Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-5208-52475

**Bc. Adrián Huňa**

# Podpora odpovedania na otázky v online komunitách študentov s využitím archívu otázok a odpovedí

Diplomová práca

Študijný program: Informačné systémy

Študijný odbor: 9.2.6 Informačné systémy

Miesto vypracovania: Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU v Bratislave

Vedúci práce: Ing. Ivan Srba, PhD.

máj 2017

STU FIIT
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

# Zadanie diplomovej práce

*Meno študenta:* **Bc. Adrián Huňa**

*Študijný program:* Informačné systémy

*Študijný odbor:* Informačné systémy

*Názov práce:* **Podpora odpovedania na otázky v online komunitách študentov s využitím archívu otázok a odpovedí**

Samostatnou výskumnou a vývojovou činnosťou v rámci predmetov Diplomový projekt I, II, III vypracujte diplomovú prácu na tému, vyjadrenú vyššie uvedeným názvom tak, aby ste dosiahli tieto ciele:

*Všeobecný cieľ:*

Vypracovaním diplomovej práce preukážte, ako ste si osvojili metódy a postupy riešenia relatívne rozsiahlych projektov, schopnosť samostatne a tvorivo riešiť zložité úlohy aj výskumného charakteru v súlade so súčasnými metódami a postupmi študovaného odboru využívanými v príslušnej oblasti a schopnosť samostatne, tvorivo a kriticky pristupovať k analýze možných riešení a k tvorbe modelov.

*Špecifický cieľ:*

Vytvorte riešenie zodpovedajúce návrhu textu zadania, ktorý je prílohou tohto zadania. Návrh bližšie opisuje tému vyjadrenú názvom. Tento opis je záväzný, má však rámcový charakter, aby vznikol dostatočný priestor pre Vašu tvorivosť.

Riaďte sa pokynmi Vášho vedúceho.

Pokiaľ v priebehu riešenia, opierajúc sa o hlbšie poznanie súčasného stavu v príslušnej oblasti, alebo o priebežné výsledky Vášho riešenia, alebo o iné závažné skutočnosti, dospejete spoločne s Vaším vedúcim k presvedčeniu, že niečo v texte zadania a/alebo v názve by sa malo zmeniť, navrhnite zmenu. Zmena je spravidla možná len pri dosiahnutí kontrolného bodu.

*Miesto vypracovania:* Ústav informatiky a softvérového inžinierstva, FIIT STU Bratislava

*Vedúci práce:* **Ing. Ivan Srba**

*Termíny odovzdania:*

Podľa harmonogramu štúdia platného pre semester, v ktorom máte príslušný predmet (Diplomový projekt I, II, III) absolvovať podľa Vášho študijného plánu

*Predmety odovzdania:*

V každom predmete dokument podľa pokynov na www.fiit.stuba.sk v časti: home > Informácie o > štúdiu > organizácia štúdia > diplomový projekt.

V Bratislave dňa 15. 2. 2016


prof. Ing. Pavol Návrat, PhD.
riaditeľ Ústavu informatiky a softvérového inžinierstva

STU FIIT
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

# Návrh zadania diplomovej práce

*Finálna verzia do diplomovej práce* [1]

## Študent:

**Meno, priezvisko, tituly:** Adrián Huňa, Bc.
**Študijný program:** Informačné systémy
**Kontakt:** huna.adrian@gmail.com

## Výskumník:

**Meno, priezvisko, tituly:** Ivan Srba, Ing.

## Projekt:

**Názov:** Podpora odpovedania na otázky v online komunitách študentov s využitím archívu otázok a odpovedí

**Názov v angličtine:** Supporting Question Answering in Online Student Communities by Utilizing Archives of Solved Questions

**Miesto vypracovania:** Ústav informatiky a softvérového inžinierstva, FIIT STU

**Oblasť problematiky:** Adaptívne a odporúčacie systémy

### Text návrhu zadania[2]

Archívy systémov pre odpovedanie na otázky v komunitách (angl. Community Question Answering - CQA) obsahujú veľké množstvo informácií a vedomostí. Základnou úlohou pre opätovné použitie obsahu v archíve je vyhľadanie podobných otázok pre rôzne dopyty používateľov systému. Napriek dostupnosti nástrojov pre takéto vyhľadávanie v archívoch vyriešených otázok, vzniká v CQA systémoch veľké množstvo nových duplicitných otázok. Tento problém je ešte výraznejší v online komunitách študentov, ktoré sú špecifické pravidelným opakovaním sa otázok. Je to spôsobené striedaním sa študentov na jednotlivých predmetoch prednášaných na univerzitách alebo v tzv. MOOC (angl. Masive Open Online Courses) kurzoch.

Analyzujte rôzne prístupy na opätovné využitie znalostí a vedomostí, ktoré sú uložené v archívoch CQA systémov. Navrhnite a zrealizujte metódu, ktorá podporí proces odpovedania na otázky študentov s využitím vedomostí, ktoré sa už v systéme nachádzajú. Využite pri tom špecifické vlastnosti vzdelávacieho procesu a informácie dostupné v online komunitách študentov. Navrhnuté riešenie experimentálne overte nad dátami z existujúceho CQA systému určeného pre online komunitu študentov.

---

[1] Vytlačiť obojstranne na jeden list papiera

[2] 150-200 slov (1200-1700 znakov), ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

**Literatúra**[3]


- CAO, Xin, Gao CONG, Bin CUI a Christian S. JENSEN, 2010. A Generalized Framework of Exploring Category Information for Question Retrieval in Community Question Answer Archives. In Proceedings of the 19th International Conference on World Wide Web - WWW '10. New York, New York, USA: ACM Press, s. 201–210.
- ZHANG, Kai, Wei WU, Haocheng WU, Zhoujun LI a Ming ZHOU, 2014. Question Retrieval with High Quality Answers in Community Question Answering. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14. New York, New York, USA: ACM Press, s. 371–380.


Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval(a) Bc. Adrián Huňa, konzultoval(a) a osvojil(a) si ho Ing. Ivan Srba a súhlasí, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V Bratislave dňa 24.1.2016

Podpis študenta

Podpis výskumníka

**Vyjadrenie garanta predmetov Diplomový projekt I, II, III**

Návrh zadania schválený: áno / ~~nie~~[4]

Dňa: 15.2.2016

Podpis garanta predmetov

[3] 2 vedecké zdroje, každý v samostatnej rubrike a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uveďte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

[4] Nehodiace sa prečiarknite

Čestne vyhlasujem, že som túto prácu vypracoval samostatne, na základe konzultácií a s použitím uvedenej literatúry.

V Bratislave, 10.5.2017

Bc. Adrián Huňa

# Poďakovanie

Ďakujem vedúcemu mojej diplomovej práce, Ing. Ivanovi Srbovi, PhD. za množstvo cenných rád, vynaložený čas a neoceniteľnú odbornú pomoc, ktorá mi pomohla počas riešenia diplomovej práce. Zároveň sa chcem poďakovať členom výskumnej skupiny PeWe, ktorí mi pomohli pri anotovaní dátovej sady.

Bc. Adrián Huňa

# Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informačné systémy

Autor: Bc. Adrián Huňa
Diplomová práca: Podpora odpovedania na otázky v online komunitách študentov s využitím archívu otázok a odpovedí
Vedúci diplomovej práce: Ing. Ivan Srba, PhD.
máj 2017

Dostupnosť vzdelávacieho materiálu na webe viedla k vzniku mnohých online komunít študentov. Tisícky študentov sa zapisujú do hromadných otvorených online kurzov (angl. Massive Open Online Courses - MOOCs), kde majú prístup k vzdelaniu bez demografických a geografických obmedzení. Pri interakcii so študijným materiálom vzniká prirodzene množstvo otázok, ktoré môžu študenti komunikovať na fórach, ktoré sú súčasťou platforiem ponúkajúcich MOOC kurzy. Proces pýtania sa a získavania odpovedí od komunity sa na otvorenom webe odohráva predovšetkým v systémoch pre odpovedanie na otázky v komunitách (angl. Community Question Answering - CQA). V tejto práci sa venujeme použitiu CQA systémov v doméne MOOC kurzov.

S cieľom riešiť problém vytvárania duplicitných otázok v CQA systémoch a MOOC kurzoch, navrhujeme metódu automatickej podpory používateľov. Otázky v MOOC kurzoch sa pravidelne opakujú predovšetkým medzi iteráciami kurzu, nakoľko používatelia väčšinou nemajú možnosť vyhľadávať v obsahu predchádzajúcej inštancie kurzu. Nami navrhnutá metóda identifikuje podobné otázky a v prípade zhody novej otázky s niektorou otázkou z archívu, automaticky zodpovie novú otázku použitím pôvodnej odpovede.

Navrhnutú metódu sme overili pomocou syntetického experimentu. Dosiahli sme 72%-nú úspešnosť pri odporúčaní jednej odpovede pre otázku, pričom odpoveď bola odporúčaná približne jednej z piatich otázok, ktoré mali v archíve podobnú otázku.

# Annotation

Slovak University of Technology Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Information systems

Author: Bc. Adrián Huňa
Master's Thesis: Supporting Question Answering in Online Student Communities by Utilizing Archives of Solved Questions
Supervisor: Ing. Ivan Srba, PhD.
2017, May

The advent of educational material available online resulted in many online student communities. Massive Open Online Courses (MOOCs) face thousands of students enrolled thanks to no demographic or geographic restrictions. Many questions naturally emerge as a result of students' interactions with course's material. Discussion forums, that are part of MOOCs, serve as a place for such questions. On the open web asking a question and obtaining answers from a community is a process faciliated by Community Question Answering (CQA) systems. In this work, we address usage of CQA systems in MOOCs.

We propose a question answering support method that solves the problem of duplicate questions. Questions in MOOCs tend to repeat especially as a course repeats in time. We observed that MOOCs usually repeat on the yearly or half-yearly basis, just like courses at physical universities. Students in MOOCs have no way to search questions in previous instances of a course and thus cannot utilize the knowledge already present in archives of questions and answers. Our method utilizes such archives and automatically answers new questions if a match in the archive is found.

The proposed method was evaluated by means of an offline experiment. We achieved 72% precision for recommendation of one answer to a new question. An answer was recommended to almost one in five questions, that had a similar question in the dataset.

# Obsah

# Kapitola 1

# Úvod

Rozvoj elektronického vzdelávania v posledných rokoch umožnil vznik mnohých typov online komunít študentov. Publikácia vzdelávacieho obsahu voľne dostupného na webe znamená neobmedzený prístup ku vzdelávaniu pre všetkých záujemcov bez ohľadu na demografické a geografické obmedzenia. V posledných rokoch sa do popredia dostali predovšetkým hromadné otvorené online kurzy (angl. Massive Open Online Courses - MOOCs), ktoré predstavujú štruktúrovaný prístup k vzdelávaniu snažiaci sa napodobniť proces výučby na univerzitách. V MOOC kurzoch zvyknú byť naraz zapísané tisícky študentov, ktorí tvoria komunitu v online priestore.

Študenti denného štúdia na univerzitách si tiež zvyknú vytvárať online komunity. Takéto komunity vznikajú predovšetkým na sociálnych sieťach, kde si študenti vytvárajú skupiny. Skupiny slúžia predovšetkým na zdieľanie informácií a vedomostí, podobne ako v prípade študentov online kurzov.

Dôležitou časťou interakcie študentov v online komunitách je pýtanie sa a odpovedanie na otázky. Používatelia sa zvyknú pýtať otázky ostatných členov komunity v prípadoch, keď vyhľadávanie informácie prostredníctvom webového vyhľadávača zlyhá. Typickými prípadmi zlyhávania webových vyhľadávačov je hľadanie odpovedí na príliš špecifické alebo komplexné otázky.

Proces položenia otázky na zodpovedanie komunitou je hlavnou náplňou systémov pre odpovedanie na otázky v komunitách (angl. Community Question Answering systems - CQA). Používatelia v týchto systémoch môžu položiť otázku opísanú ľubovoľne dlhým textom a dostať na ňu odpoveď od zvyšku komunity. Takéto odpovede bývajú špecializované na detaily otázky, vďaka čomu si používateľ nemusí hľadanú informáciu odvodzovať, ako je to napríklad potrebné pri hľadaní odpovedí cez vyhľadávače na webe.

Medzi hlavné metódy podpory používateľa na webe patrí personalizácia a odporúčanie. Dôležitým prvkom odporúčania v CQA systémoch je odporúčanie zaujímavých vyriešených otázok z archívu otázok a odpovedí. Odpovede v archíve predstavujú cenné úložisko vedomostí, ktoré je možné využiť pri procese podpory používateľa.

V tejto práci sa venujeme podpore používateľov v online komunitách študentov s využitím archívov otázok a odpovedí. Jedným z identifikovaných problémov online komunít študentov je pravidelné opakovanie otázok spôsobené iteráciami kurzu v čase. Diskusné fóra používané v MOOC kurzoch neumožňujú vyhľadávať v obsahu predchádzajúcich iterácií kurzu a preto používateľ nemá možnosť sa k daným informáciám dostať.

Našim cieľom je navrhnúť, implementovať a overiť metódu podpory používateľov realizovanú automatickým odpovedaním na otázky s využitím odpovedí v archíve odpovedí. Pri hľadaní rovnakých otázok sa zameriavame na sémantickú podobnosť otázok a ďalšie črty, medzi ktoré paria aj črty, ktoré môžeme nájsť len v doméne online komunít študentov. Špecifiká online komunít študentov berieme do úvahy taktiež pri zoraďovaní nájdených vhodných odpovedí v procese hľadania najvhodnejšej odpovede.

Zvyšok práce je organizovaný nasledovne. V kapitole 2 sa venujeme analýze CQA systémov a možnostiam podpory používateľov v nich. V kapitole 3 sa venujeme analýze online komunít študentov a bližšie sa zaoberáme komunitami v MOOC kurzoch. V kapitole 4 analyzujeme CQA systémy určené pre doménu vzdelávania a venujeme sa rozdielom medzi CQA systémami a diskusiami v MOOC kurzoch. Konceptuálny návrh našej metódy predstavujeme v kapitole 5. V kapitole 6 opisujeme realizáciu navrhnutej metódy pre MOOC kurz týkajúci sa výučby informatiky. Kapitola 7 je venovaná overeniu navrhnutej metódy. Zhodnotenie výsledkov našej práce a možnosti ďalšieho smerovania uvádzame v kapitole 8.

# Kapitola 2

# Systémy pre odpovedanie na otázky v komunitách

Web vždy slúžil ako miesto pre získanie nových informácií a nadobudnutie znalostí. Aby sme boli schopní využiť tieto informácie, je potrebné mať webové stránky indexované, o čo sa starajú webové vyhľadávače. Ak má používateľ nejakú otázku alebo problém, jeho prvý krok zvyčajne býva zadanie kľúčových slov do webového vyhľadávača a prezretie relevantných nájdených výsledkov. Vyhľadávače však niekedy nie sú schopné poskytnúť uspokojivé výsledky vyhľadávania a používateľ je odkázaný na alternatívne možnosti získania odpovede.

Systémy pre odpovedanie na otázky v komunitách (angl. Community Question Answering - CQA) umožňujú používateľom získať riešenia na príliš špecifické alebo komplexné problémy a otázky. Tieto systémy zaznamenali v posledných rokoch nárast popularity a s tým súvisiace zaujímavé problémy pre výskumníkov. Úspech CQA systémov je založený predovšetkým na komunite používateľov systému, ktorí zdieľajú svoje vedomosti pri poskytovaní personalizovaných odpovedí. Jadrom CQA systémov je interakcia medzi používateľmi, predovšetkým vo forme pýtania sa a odpovedania na otázky. Medzi ďalšie aktivity používateľov patrí:

- komentovanie otázok a odpovedí;
- hlasovanie o kvalite otázok, odpovedí a komentárov;
- označenie najlepšej odpovede;
- označovanie zaujímavého obsahu; a
- nasledovanie aktivity v systéme.

Používatelia môžu nasledovať aktivitu ostatných používateľov napríklad pre jednotlivé otázky, kategórie, či používateľov. Doméne CQA systémov sme sa venovali aj v našej predchádzajúcej práci (Huna et al., 2016), v ktorej sme navrhli inovatívnu metódu určenia reputácie používateľov.

Podpora používateľov v CQA systémoch je dôležitá oblasť pre udržanie živej komunity. V doméne CQA systémov sú najdôležitejšími príkladmi podpory používateľov:

- odporúčanie nových otázok na zodpovedanie (kapitola 2.2.1); a
- získavanie vyriešených otázok z archívu otázok a odpovedí (kapitola 2.2.2).

Obe tieto metódy majú za cieľ ovplyvniť spôsob spolupráce používateľov.

## 2.1 Populárne CQA systémy

Medzi najpopulárnejšie CQA systémy patria Yahoo! Answers[1], Stack Overflow[2] a Quora[3]. Podľa zamerania CQA systémov rozlišujeme dva typy. Prvým typom sú systémy všeobecné (Yahoo! Answers a Quora), v ktorých je možné sa pýtať otázky zo širokého spektra kategórií od biológie, cez medziľudské vzťahy, až po biznis stratégie. Druhým typom sú špecificky zamerané systémy, napríklad Stack Overflow zameraný na oblasť vývoja softvéru.

Obsah v CQA systémoch zvykne byť organizovaný buď do kategórií (Yahoo! Answers, Quora) alebo podľa značiek (Stack Overflow). Kategórie v systéme Yahoo! Answers majú hierarchickú štruktúru a používateľ musí pridať novú otázku do jednej z listových kategórií. Organizácia prostredníctvom značiek ponúka väčšiu flexibilitu pri kategorizovaní otázky, nakoľko otázka môže patriť do viacerých tém.

Najväčšie CQA systémy majú niekoľko miliónov používateľov (napr. Yahoo! Answers s vyše 10 miliónmi aktívnych používateľov mesačne a Stack Overflow s vyše 5 miliónmi registrovaných používateľov) a obsahujú milióny otázok a odpovedí (napr. Stack Overflow - vyše 13 750 000 otázok a 21 700 000 odpovedí (platné pre máj 2017). Kvalita vytvoreného obsahu takýmto množstvom používateľov je prirodzene rôzna. V systéme Yahoo! Answers môžeme nájsť množstvo nekvalitných odpovedí, no naopak systém Quora je známy práve odpoveďami s vysokou kvalitou, ktoré pridávajú odborníci z rôznych oblastí.

Na udržanie živej komunity obsahujú CQA systémy viaceré gamifikačné mechanizmy. Najčastejšie je to reputácia, ktorá umožňuje na prvý pohľad identifikovať expertného používateľa, avšak často slúži aj na sprístupnenie určitých funkcií. Používatelia v systéme Stack Overflow môžu napríklad komentovať príspevky až po získaní 50 bodov reputácie. V systéme Yahoo! Answers majú používatelia obmedzený počet otázok a odpovedí, ktoré môžu pridať za jeden deň podľa toho, v ktorom leveli podľa výšky reputácie sa nachádzajú. CQA systém Quora viditeľný mechanizmus reputácie neobsahuje. V našej predchádzajúcej práci (Huna et al., 2016) sme sa venovali určeniu reputácie používateľov s využitím náročnosti otázok a užitočnosti otázok a odpovedí.

Ďalším gamifikačným prvkom sú odznaky, ktoré môžeme nájsť v systéme Stack Overflow, respektíve vo všetkých systémoch Stack Exchange platformy[4], ktorej je Stack Overflow súčasťou. Tieto odznaky slúžia len na zvýraznenie dosiahnutých míľnikov.

[1] https://answers.yahoo.com/
[2] https://stackoverflow.com/
[3] https://www.quora.com/
[4] https://stackexchange.com/

## 2.2 Podpora spolupráce používateľov v CQA systémoch

Podpora používateľov na webe je populárna téma so širokými možnosťami realizácie. Medzi najčastejšie spôsoby podpory používateľa patrí personalizácia a odporúčanie obsahu, napríklad vo forme:

- odporúčania tovaru v internetových obchodoch;
- odporúčania televíznych programov, filmov a hudby;
- personalizovanej reklamy;
- personalizovaného zobrazovania obsahu na sociálnych sieťach.

V CQA systémoch sú hlavnými typmi obsahu, ktorý môžeme používateľom odporúčať otázky a odpovede. Stack Overflow, Yahoo! Answers a Quora obsahujú zoznam otázok, ktoré sú zobrazené na úvodnej stránke po príchode do systému. Yahoo! Answers a Quora využívajú tento zoznam len na zobrazenie vyriešených zaujímavých otázok, no Stack Overflow v ňom zobrazuje najnovšie vyriešené aj nevyriešené zaujímavé otázky. Používateľské rozhranie na smerovanie nevyriešených otázok majú Yahoo! Answers a Quora v samostatnej časti systému.

Medzi ďalšie konkrétne príklady podpory používateľov v CQA systémoch patrí vyhľadávanie otázok na základe krátkeho dopytu, podobne ako fungujú webové vyhľadávače pri vyhľadávaní webových stránok. Podobné odporúčanie otázok prebieha, keď používateľ začne písať novú otázku a systém priebežne vyhľadáva a zobrazuje podobné otázky, o ktoré by používateľ mohol mať záujem.

Za metódu podpory aktivity používateľa môžeme považovať gamifikačné machanizmy opísané v predchádzajúcej podkapitole. Tieto nástroje prispievajú k udržiavaniu aktívnej komunity používateľov, zvyšujú motiváciu používateľov a prispievajú tým k množstvu a kvalite odpovedí v CQA systémoch.

Podpora používateľov v CQA systémoch je dôležitá téma. Wang et al. (2011) pozorovali, že iba 17,6% otázok v systéme Yahoo! Answers získalo uspokojivú odpoveď do 48 hodín od pridania otázky a takmer 20% všetkých otázok zostalo bez odpovede. Podobné percento nezodpovedaných otázok pozorovali aj Shtok et al. (2012) (približne 25%). Veľký počet nezodpovedaných otázok naznačuje, že schopnosť CQA systémov priamo vyriešiť otázky používateľov je slabá a je potrebné navrhnúť a overiť nové metódy podpory používateľov.

Podporu spolupráce používateľov v CQA systémoch najčastejšie realizuje ako:

- odporúčanie nových otázok na zodpovedanie (kapitola 2.2.1); a
- získavanie vyriešených otázok z archívu otázok a odpovedí (kapitola 2.2.2).

### 2.2.1 Smerovanie nových otázok

Smerovanie nových otázok (angl. question routing) má za cieľ maximalizovať využitie znalostí jednotlivých používateľov vzhľadom na ich skúsenosti a vedomosti. Je neefektívne, ak sú expertom zobrazované jednoduché otázky, ktoré by vedeli zodpovedať aj menej expertní používatelia. V takýchto prípadoch dochádza k informačnému preťaženiu experta a najnáročnejšie otázky ostávajú nezodpovedané. Príkladom práce, ktorá do svojho modelu zahŕňa zmienené obmedzenia smerovania otázok je (Yang et al., 2014), kde sa autori venovali smerovaniu otázok v prostredí online kurzov.

Princíp smerovania nových otázok je nasledovný: pre novú otázku $q$ vytvor zoradený zoznam top $k$ používateľov $u_1$, $u_2$, ..., $u_k$, ktorí sú najvhodnejší na zodpovedanie otázky $q$ (Srba a Bielikova, 2016a). Určenie daného zoznamu používateľov je založené predovšetkým na ich expertíze, záujmoch a motivácii odpovedať na otázky.

Príkladom nesprávneho smerovania otázok je odporúčanie otázky ohľadom bezpečnosti webových aplikácií grafickému dizajnérovi alebo odporúčanie otázky ohľadom zložitých chemických vzorcov študentovi strednej školy.

### 2.2.2 Získavanie vyriešených otázok z archívu

Zatiaľ čo smerovanie nových otázok sa spolieha na vedomosti používateľov, získavanie otázok z archívu sa zameriava na využitie znalostí, ktoré sú už v systéme dostupné. Vyhľadávanie v CQA systémoch umožňuje pre zlepšenie presnosti využiť dodatočné informácie o histórii a správaní sa používateľov, ktoré webové vyhľadávače nemajú k dispozícii. Rozlišujeme tri typy využitia archívu otázok a odpovedí:

- vyhľadávanie otázok,
- odporúčanie otázok,
- odpovedanie na otázky.

Všetky tieto typy sú charakterizované prítomnosťou (Srba a Bielikova, 2016a):

- profilu, ktorý reprezentuje informačnú potrebu zachytenú v používateľovom dopyte;
- profilov, ktoré opisujú existujúce otázky a ich odpovede;
- porovnávacím modelom, ktorý počíta podobnosť medzi týmito profilmi.

#### Vyhľadávanie otázok

Cieľom vyhľadávania otázok (angl. question search) je na základe dopytu používateľa nájsť v archíve také otázky, ktoré sú sémanticky čo najviac zhodné s daným dopytom. Príklady takýchto otázok uvádzame v tabuľke 2.1. Ako sme uviedli v úvode kapitoly 2.2, typickým príkladom vyhľadávania otázok je vrátenie výsledkov pri hľadaní otázok pomocou kľúčových slov a odporúčanie existujúcich otázok pri vytváraní novej otázky, čo slúži ako prevencia vytvárania duplicitných otázok.

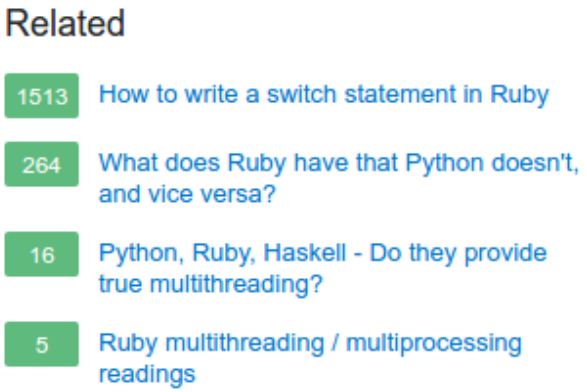
**Tabuľka 2.1:** Príklady výsledkov vyhľadávania na základe dopytu používateľa. Prevzaté a preložené z (Duan et al., 2008).

| **Dopyt:** |
|---|
| $O_1$: Aké sú dobré kluby v Berlíne alebo Hamburgu? |
| **Očakávané:**<br>$O_2$: Ktoré sú najlepšie/najzábavnejšie kluby v Berlíne? |
| **Neočakávané:**<br>$O_3$: Aké sú dobré hotely v Berlíne?<br>$O_4$: Lacné divadlá v Hamburgu? |

Najväčšou výzvou vyhľadávania otázok je, že používatelia zvyknú opisovať rovnakú informáciu rôznymi slovami. Tento jav sa označuje ako *lexikálna medzera* a bolo navrhnutých niekoľko spôsobov, ako sa s ním vysporiadať. Jednotlivým spôsobom sa budeme bližšie venovať v ďalšej časti práce (kapitola 2.3).

### Odporúčanie otázok

Odporúčanie otázok (angl. question recommendation) je veľmi podobné vyhľadávaniu otázok, avšak namiesto hľadania sémanticky rovnakých otázok je cieľom odporúčať sémantický príbuzné otázky, ktoré dopĺňajú spracovávanú otázku. Takýto typ odporúčania môžeme v CQA systémoch nájsť pri pohľade na jednu konkrétnu otázku. Príklad nájdenia podobných otázok v systéme Stack Overflow je na obr. 2.1.




**Obr. 2.1:** Zobrazenie podobných otázok v systéme Stack Overflow ku otázke "Does ruby have real multithreading?".

### Odpovedanie na otázky

Odpovedanie na otázky (angl. question answering) je podobná technika ako vyhľadávanie otázok, avšak obsahuje ešte jeden krok navyše. Po nájdení sémanticky najpodobnejších otázok na novú otázku sa najlepšia odpoveď použije ako odpoveď pre novú otázku. Cieľom tejto techniky je preto znovupoužiť existujúce odpovede v archíve odpovedí na zodpovedanie nevyriešených otázok. Automatické odpovedanie možno použiť napr. pre vyriešenie dlhodobo nezodpovedaných otázok alebo na rýchle poskytnutie odpovedí novým otázkam.

## 2.3 Identifikovanie sémanticky podobných otázok

Po identifikovaní dvoch hlavných možností podpory spolupráce používateľov - *smerovanie nových otázok*, a *získavanie vyriešených otázok z archívu*, sa v tejto kapitole bližšie venujeme druhej možnosti. Archív otázok a odpovedí predstavuje vysoký potenciál vo forme uložených vedomostí, ktoré je možné znovupoužiť. Kým smerovanie otázok potrebuje k úspešnému fungovaniu zásah ľudského používateľa, na druhej strane využitie archívu formou automatického odpovedania na otázky je plne autonómny proces. Iné využitia archívu, napríklad vyhľadávanie v archíve je funkcionalita, ktorá sa prirodzene očakáva od CQA systémov. Samotné využitie archívu otázok a odpovedí má aj ďalšie výhody, ako je napríklad možnosť poskytnutia okamžitej odpovede na novú otázku a odľahčenie expertov.

Všetky tri typy využitia archívu otázok a odpovedí sú fundamentálne založené na identifikovaní sémanticky podobných otázok. Jednotlivé typy sa líšia v tom ako využívajú výsledok nájdenia podobných otázok:

- *vyhľadávanie otázok* - hľadá exaktnú sémantickú zhodu;
- *odporúčanie otázok* - hľadá otázky, ktoré sa týkajú rovnakej témy;
- *odpovedanie na otázky* - hľadá okrem exaktne zhodných otázok aj otázky, ktoré sú veľmi súvisiace.

Prehľad relevantnej literatúry zaoberajúcej sa identifikovaniu sémanticky podobných otázok uvádzame v tabuľke 2.2. Existujúce riešenia možno charakterizovať:

- modelom pre výpočet podobnosti textov (kapitola 2.3.1),
- obsahom, ktorý je použitý pri výpočte podobnosti (kapitola 2.3.2).

Mnohé z predchádzajúcich prác sa snažili základné modely určitým spôsobom rozšíriť a zvýšiť tak ich robustnosť. Identifikovali sme dva hlavné smery, ktorým sa práce venovali:

- rozšírenia zachytávajúce kontext (kapitola 2.3.3),
- rozšírenia s využitím metadát (kapitola 2.3.4).

### 2.3.1 Modely pre výpočet podobnosti textov

V kapitole 2.2.2 sme identifikovali, že najväčšou výzvou vyhľadávania sémanticky podobných otázok je problém *lexikálnej medzery*. Lexikálnu medzeru spôsobuje použitie rôznych slov na vyjadrenie rovnakej informácie. Z tohto dôvodu výpočet podobnosti na jednoduchých vektorových reprezentáciách textov nie je postačujúci (napr. vektory reprezentujúce počet výskytu slov v texte). V predchádzajúcich prácach boli navrhnuté viaceré spôsoby, ako sa s týmto problém vysporiadať.

Pre nájdenie sémanticky podobných otázok v CQA systémoch rozlišujeme modely uvedené v tabuľke 2.3 (pozn.: stĺpec *Označenie* korešponduje s označením v tabuľke 2.2).

**Tabuľka 2.2:** Prehľad prác zaoberajúcich sa hľadaním rovnakých otázok a odpovedaním na otázky. Písmeno v stĺpci zameranie označuje riešený podproblém (S - vyhľadávanie otázok, R - odporúčanie otázok, A - odpovedanie na otázky). Skratka *naj. odpoveď* v stĺpci použitý zdroj textu znamená *najlepšia odpoveď*.

| Práca | Zameranie | Použité modely | Text pre nájdenie podobných otázok | Dátová sada | Zlatý štandard | Metriky |
|---|---|---|---|---|---|---|
| (Cao et al., 2010) | (S) Kategórie otázok | LMIR, TB, TBLM, VSM, BM25 | Otázka | Y!A | Manuálne | MAP, MRR, P@5, R-Presnosť |
| (Duan et al., 2008) | (S) Kontext otázok | LMIR, TB | Nadpis otázky | Y!A | Manuálne | MAP, MRR, R-Presnosť |
| (Chen et al., 2013) | (S) Typ otázok | LMIR, TBLM | Informácia nie je dostupná | Y!A, WikiAnswers | Manuálne | MAP, P@10 |
| (Ji et al., 2012) | (S) Latentné témy | LMIR, TBLM, Topic | Otázka a naj. odpoveď | Y!A | Manuálne | MAP, P@10 |
| (Li a Manandhar, 2011) | (R) Informačná potreba | Topic + TB | Otázka | Y!A | Manuálne | MRR, P@{5,10} |
| (Pera a Ng, 2011) | (A) Automatické odpovedanie | Word-Corr | Otázka | Y!A, TREC | Manuálne | Presnosť a úplnosť |
| (Shtok et al., 2012) | (A) Automatické odpovedanie | Topic | Otázka a naj. odpoveď | Y!A | Manuálne | Presnosť a úplnosť |
| (Suryanto et al., 2009) | (A) Zoradenie odpovedí | LMIR, VSM, BM25 | - | Y!A | Manuálne | Správnosť, MAP, MRR, P@{1,5,10} |
| (Wang et al., 2009) | (S) Syntaktický strom | SyntaxTree | Otázka a naj. odpoveď | Y!A | Manuálne | MAP, P@1 |
| (Wang et al., 2011) | (R) Popularita otázok | - | Otázka a odpovede | Y!A | Manuálne | MAP, MRR, P@10 |
| (Wu et al., 2014) | (S) Obsah zdrojov tretích strán | LMIR, TB, TBLM | Otázka, zdroje 3. strán | Y!A, Quora | Manuálne | nDCG@{1,3,5} |
| (Xue et al., 2008) | (S) Mapovanie textov otázok na text odpovedí | LMIR, TB, TBLM, BM25 | Otázka a naj. odpoveď | Wondir | Manuálne | MAP, P@10 |
| (Zhang et al., 2014) | (S) Kvalita odpovedí | LMIR, TBLM, Topic | Otázka a odpovede | Y!A, Baidu Knows | Manuálne | MAP, MRR, R-Presnosť, P@1 |
| (Zhou et al., 2011) | (S) Mapovanie fráz | LMIR, TB, TBLM | Otázka a naj. odpoveď | Y!A | Manuálne | MAP |

**Tabuľka 2.3:** Prehľad modelov používaných pri výpočte podobnosti textov otázok.

| Názov | Označenie |
|---|---|
| Jazykový model (pravdepodobnosti dopytu) označovaný aj ako jazykový model pre získavanie informácií (angl. language model for information retrieval) | LMIR |
| Model mapovania slov (angl. translation-based model) | TB |
| Jazykový model rozšírený o model mapovania slov (angl. translation-based language model) | TBLM |
| Vektorový model (angl. vector space model) | VSM |
| Okapi BM25 model | BM25 |
| Model založený na syntaktických štruktúrach | SyntaxTree |
| Model využívajúci latentné témy (angl. latent topic model) | Topic |
| Model založený na koreláciách výskytu slov | Word-Corr |
| word2vec model | word2vec |

### Jazykový model (LMIR)

Jazykový model (pravdepodobnosti dopytu) (angl. (query-likelyhood) language model) sa často označuje aj ako *jazykový model* alebo *jazykový model pre získavanie informácií* (angl. language model for information retrieval). Základnou myšlienkou jazykového modelu je určiť jazykový model pre každú otázku a následne zoradiť otázky podľa podobnosti s jazykovým modelom dopytu. Model pracuje s pravdepodobnosťami výskytu slova $t$ v dokumente $d$ a v kolekcii $coll$. Tento model zlyháva ak sémanticky podobné otázky majú málo spoločných slov. Tento model bol použitý v prácach (Cao et al., 2010; Duan et al., 2008; Chen et al., 2013; Ji et al., 2012; Wu et al., 2014; Xue et al., 2008; Zhang et al., 2014; Zhou et al., 2011).

### Model mapovania slov (TB)

Princípom modelu mapovania slov (angl. translation-based model) je využitie pravdepodobností mapovania jedného slova na iné slovo, čím sa vyrieši problém rozličných slov medzi textami. Pravdepodobnosť, že sa nejaké slovo namapuje samo na seba je 1. Pravdepodobnosti mapovania slov sa väčšinou získava cez IBM translation model 1 a IMB-4, a v predchádzajúcich prácach na ich určenie bol použitý nástroj GIZA++[5]. Pri hľadaní mapovaní je dôležité, aby texty mali podobnú dĺžku, inak ostane veľa slov z dlhšieho textu nenamapovaných. Tento model bol použitý v prácach (Cao et al., 2010; Duan et al., 2008; Li a Manandhar, 2011; Wu et al., 2014; Xue et al., 2008; Zhou et al., 2011).

[5] http://www.statmt.org/moses/giza/GIZA++.html

Zhou et al. (2011) navrhujú použiť mapovanie fráz namiesto jednoduchého mapovania slov. Slová, ktoré sú pri sebe lepšie zachytávajú kontext otázky a umožňujú vhodnejšie naučenie sa mapovaní. Ako príklad autori uvádzajú frázu *upchatý nos*, ktorá sa namapuje na slovo *nádcha* s väčšou pravdepodobnosťou, akoby sa namapovali jednotlivé slová samostatne. Aby sa vyhli problému odlišnej dĺžky textov otázok a odpovedí, tak v prvom kroku extrahujú kľúčové slová a predpokladajú, že dopytovaná otázka je mapovaná len z kľúčových slov. Kľúčové slová identifikujú metódou *zarovnanie slov* (angl. word alignment).

Po predspracovaní dát najskôr nájdu mapovania jednotlivých slov, aplikujú metódu *zarovnania slov* a extrahujú kľúčové frázy. Výkon svojho modelu porovnávali s modelmi založenými na mapovaní slov a pozorovali, že ich model prekonal všetky porovnávané modely. Autori ďalej porovnávali efekt maximálnej dĺžky fráz na výkon modelu a zistili, že od dĺžky 4 slová už nedochádza k štatisticky významnému zlepšeniu.

### Jazykový model rozšírený o model mapovania slov (TBLM)

Jazykový model rozšírený o model mapovania slov (angl. translation-based language model), navrhnutý autormi Xue et al. (2008), je kombináciou jazykového modelu a modelu mapovania slov. Bolo preukázané, že dosahuje lepší výkon ako oba modely, ktorých je kombináciou. Navrhnutý bol na odstránenie problému jazykového modelu, ktorý nedokáže dobre identifikovať podobnosť otázok ak majú rozličné slová. Úspešnosť modelu prekonala aj model mapovania slov a Okapi BM25 model. Tento model bol použitý v prácach (Cao et al., 2010; Chen et al., 2013; Ji et al., 2012; Wu et al., 2014; Xue et al., 2008; Zhang et al., 2014; Zhou et al., 2011).

Autori Xue et al. (2008) pri navrhovaní tohto modelu využívajú fakt, že texty otázok a odpovedí môžu byť považované za *paralelný korpus* a je možné sa naučiť pravdepodobnosti mapovania slov medzi textom otázok a odpovedí. Nakoľko sú otázky a odpovede v CQA systémoch písané v rovnakom jazyku, je možné použiť model mapovania slov pre identifikovanie pravdepodobnosti mapovania slov v otázke na slová v odpovedi a naopak. Autori sa rozhodli oba smery mapovaní skombinovať a v závere pozorovali, že mapovanie z textu odpovedí na text otázok bolo o niečo úspešnejšie.

### Vektorový model (VSM)

Vektorový model (angl. Vector Space Model) zvýhodňuje krátke otázky. Model pracuje s počtom otázok v kolekcii, počtom otázok ktoré obsahujú určitý pojem (angl. inverse document frequency - IDF) a frekvenciami pojmu (angl. term frequency - TF) v dokumente $d$. Tento model bol použitý v práci (Cao et al., 2010).

### Okapi BM25 model (BM25)

BM25 model berie v úvahu dĺžku otázok, čím eliminuje zvýhodňovanie krátkych otázok, ako je tomu v prípade vektorového modelu. Model pracuje s počtom otázok, ktoré obsahujú pojem $t$, frekvenciou pojmu v dokumente $d$, dĺžkou otázky a priemernou dĺžkou otázok v kolekcii. Model bol použitý v prácach (Cao et al., 2010; Xue et al., 2008).

### Model založený na syntaktických štruktúrach otázok (SyntaxTree)

Cieľom tohto modelu je reprezentovať syntaktickú štruktúru vety pomocou stromu. Autori Wang et al. (2009) sa venovali reprezentácii otázok pomocou stromu, ktorý dokáže zachytiť lexikálne aj syntaktické a sémantické vlastnosti a využiť ich v modeli hľadania podobných otázok. Navrhnutý model sa dokázal vysporiadať s gramatickými a pravopisnými chybami, no bol vhodnejší na krátke otázky a nie dlhšie texty.

Nájdenie podobných otázok autori realizovali v dvoch krokoch. V prvom kroku nájdu podobné otázky na základe textu otázok, následne vezmú 100 najpodobnejších otázok a vyhodnotia podobnosť odpovedí. Úspešnosť bola vyhodnotená pre rôzne kombinácie navrhnutého modelu a najkomplikovanejší model, ktorý pracoval so sémantikou textu a textom odpovedí dosiahol najlepší výkon.

### Model využívajúci latentné témy (Topic)

Viacerí autori sa snažili o identifikovanie latentných tém v obsahoch textov otázok. Použitie takéhoto prístupu je robustnejšie a všeobecnejšie, než spoliehanie sa na témy určené kategóriami, ako to spravili autori Cao et al. (2010). Autori Shtok et al. (2012) použili LDA model (z angl. Latent Dirichlent Allocation) ako vlastnosť textov pre naučenie klasifikátora, ktorý rozhodoval o vhodnosti použitia odpovede pre otázku. Pre 3 top úrovňové kategórie v systéme Yahoo! Answers naučili 200 LDA tém. Zhang et al. (2014) identifikovali LDA témy z textov otázok a odpovedí, pričom kvalita odpovedí určovala silu s akou daná odpoveď ovplyvní naučenie sa latentného priestoru.

Ji et al. (2012) taktiež využívali text otázok aj odpovedí pre naučenie latentných tém. Navrhnutý model vylepšujú obmedzením množstva slov priradených do tém v otázkach a odpovediach. Text odpovedí zvykne byť dlhší než text otázok, čo môže spôsobiť, že text otázky bude mať väčší vplyv na určenie tém daného páru otázka-odpoveď. Vďaka príslušnému obmedzeniu je tento problém odstránený.

Navrhnutý model kombinujú s jazykovým modelom a jazykovým modelom rozšíreným o model mapovania slov. Vyhodnotenie všetkých kombinácií modelov preukázalo, že nimi navrhnutý model v kombinácií s jazykovým modelom rozšíreným o model mapovania slov dosiahol najlepšie výsledky. Autori ďalej pozorovali, že naučenie sa tém z obsahu otázok a odpovedí bolo efektívnejšie, než naučenie sa tém len z obsahu otázok alebo len z obsahu odpovedí (avšak rozdiel v úspešnosti sa pohyboval v rozmedzí jedného percenta).

### Model založený na koreláciách výskytu slov (Word-Corr)

Autori Pera a Ng (2011) sa rozhodli využiť faktory korelácií slov, ktoré vygenerovali na 880 000 článkoch z webovej encyklopédie Wikipedia[6]. Každý korelačný faktor indikuje podobnosť medzi dvoma slovami na základe ich frekvencie spoločného výskytu vedľa seba v texte a vzájomnej vzdialenosti v textoch. Otázky z CQA systému Yahoo! Answers reprezentujú

[6] https://sk.wikipedia.org/

pomocou extrahovaných kľúčových slov a ováhovaním jednotlivých slov. Samotné identifikovanie sémanticky podobných otázok realizovali pre otázky, ktoré mali spoločné kľúčové slová (resp. kľúčové slová museli byť veľmi podobné). Podobnosť otázok je založená na súčte korelačných faktorov pre všetky slová otázok. Pri experimentálnom vyhodnotení dosiahla nimi navrhnutá metóda o 36% vyššiu úspešnosť (70% vs 44%) pre metriku správnosť voči nájdeným otázkam, ktoré získali z použitia vyhľadávača v systéme Yahoo! Answers.

**word2vec model (word2vec)**

Model *word2vec* (Mikolov et al., 2013) je inovatívny model založený na neurónových sieťach, ktorý reprezentuje slová vo vektorovom priestore. Hoci *word2vec* ešte nebol priamo použitý pre účely hľadania podobných otázok v CQA systémoch, myslíme si, že je jeho použitie vhodné aj na takúto úlohu. Alternatívou k *word2vec* modelu je GloVe (Pennington et al., 2014), ktorý tiež umožňuje získať vektory slov na základe ich spoločných výskytov a vzdialeností v texte.

Nakoľko model *word2vec* reprezentuje slová a nie texty, nie je možné tento model použiť priamo. Viaceré prístupy riešenia tohto problému boli navrhnuté - napríklad model *Doc2Vec*[7], ktorý poskytuje vektory pre texty ľubovoľnej dĺžky. Jednoduchším riešením je použitie váženého priemeru vektorov všetkých slov v dokumente.

### 2.3.2 Obsah použitý na vyhľadávanie podobných otázok

Modely hľadania podobných otázok môžu pracovať s rôznym obsahom v CQA systémoch. Otázky v CQA systémoch sa skladajú z nadpisu a samotného textu otázky. K otázkam sú pridávané odpovede, ktorých kvalita je ohodnotená spätnou väzbou komunity. Ďalším znakom kvality odpovedí je akceptovanie odpovede za riešenie pýtajúcim sa používateľom.

Text otázok a odpovedí v CQA systémoch je pomerne krátky a preto niektorí autori obohacovali text otázok o informácie z dodatočných zdrojov. V predchádzajúcich vedeckých prácach sme identifikovali vyhľadávanie sémanticky podobných otázok na základe:

- *nadpisu otázky* (Duan et al., 2008),
- *nadpisu a textu otázky* (Cao et al., 2010; Li a Manandhar, 2011; Pera a Ng, 2011),
- *nadpisu a textu otázky a textu najlepšej odpovede* (Ji et al., 2012; Shtok et al., 2012; Wang et al., 2009; Xue et al., 2008; Zhou et al., 2011),
- *nadpisu a textu otázky a textu všetkých odpovedí* (Zhang et al., 2014),
- *obohateniu textu o informácie zo zdrojov tretích strán* (Wu et al., 2014).

Prvé štyri prístupy využívajú dáta priamo dostupné v CQA systémoch. Posledný prístup využíva dáta, ktoré môžu byť z rôznych zdrojov a preto sa mu venujeme podrobnejšie.

[7] https://radimrehurek.com/gensim/models/doc2vec.html

**Obohateniu textu o informácie zo zdrojov tretích strán**

Wu et al. (2014) vo svojej práci použili logy vyhľadávacích dopytov používateľov a výsledky webových vyhľadávaní na zlepšenie relevancie nájdených podobných otázok v systéme Yahoo! Answers. Vyhľadávacie dopyty používateľov v Yahoo! Answers mali priemerne 4 slová a často sa nejednalo o kompletné vety. Relevantné otázky sa snažia detegovať pomocou identifikácie používateľovho zámeru, ktorý získavajú z troch zdrojov:

1. *textu odpovedí*, ktorý hovorí o zámere otázok;
2. *logov z webového vyhľadávača*, ktoré hovoria o preferenciách bežných webových používateľov; a
3. *z top výsledkov vyhľadávača*, ktoré hovoria o populárnych témach týkajúcich sa používateľovho dopytu.

Výkon svojho modelu porovnávali s tradičnými modelmi (jazykový model, jazykový model rozšírený o model mapovania slov) aj s modelom zameraným rovnako na zámer používateľa (Li a Manandhar, 2011), ktorý je založený na LDA modely. Z výsledkov vyplýva, že LDA model dosiahol najhorší výkon a nimi navrhnutá metóda najlepší.

### 2.3.3 Rozšírenie modelov o kontext otázky

Viaceré predchádzajúce práce sa snažili identifikovať kontext, v akom je otázka položená. Ak sme schopní získať kontext (napr. geografická poloha ku ktorej sa otázka viaže alebo identifikovanie aktuálneho významu homoným v texte), vieme lepšie identifikovať pravú podstatu zamerania otázok a lepšie tak nájsť podobné otázky. Nasledujúce práce sa snažili kontext identifikovať z textového obsahu otázok a odpovedí.

**Využitie stromovej štruktúry pri práci s kontextom otázok**

Autori Duan et al. (2008) vo svojej práci využívali tému a zameranie otázok. Téma otázky väčšinou reprezentuje kontext/obmedzenia otázky (napr. geografická poloha - Berlín, Hamburg). Zameranie otázky zase reprezentuje určité zameranie otázky (napr. zábavné atrakcie, lacné hotely). Cieľom autorov je nájsť podobné otázky, ktoré sú relevantné aj rovnako v rámci kontextu/obmedzenia otázky, ako aj jej zamerania.

Autori navrhli model rezania stromu (angl. tree cut model) pre automatické identifikovanie témy a zamerania otázky. Otázky v archíve reprezentujú stromovou štruktúrou príbuzných otázok, a na tento strom následne aplikujú rez stromom pre identifikovanie podobných otázok. Navrhnutú metódu skombinovali s jazykovým modelom a modelom mapovania slov. Experimentálne výsledky preukázali, že ich model prekonal vektorový model a jazykový model, ktoré použili ako metódy pre porovnanie výkonu. Variant s modelom mapovania slov bol najúspešnejší (rozdiel približne 5% s najlepšou úspešnosťou 34% pre MRR).

**Využitie informačnej potreby používateľa**

Li a Manandhar (2011) sa venovali zachyteniu kontextu otázky vo forme informačnej potreby používateľa. Nadpis otázky považujú za používateľov dopyt a text otázky za vyjadrenie jeho informačnej potreby. Autori pozorovali, že v CQA systémoch text otázky často nie je vyplnený a preto sa časť ich práce venuje hľadaniu mapovaní medzi nadpisom otázky a jej textom. Cieľom práce je zlepšenie úspešnosti pri riešení problému *odporúčanie otázok*. Na meranie podobnosti textov otázok použili niekoľko metód:

- založené na TF-IDF,
  - *TF* znamená frekvencia pojmu (angl. term frequency) a *IDF* inverzná frekvencia v dokumente (angl. inverse document frequency). Metóda hovorí o tom ako dôležité sú jednotlivé pojmy v dokumente vzhľadom na korpus dokumentov.
- založené na vedomostiach,
  - Metódy založené na metódach vyhodnotenia podobnosti s využitím korpusu WordNet.
- založené na LDA.

Výsledky preukázali, že najlepší výkon dosiahli metódy založené na LDA (úspešnosť podľa metriky MRR dosiahol LDA model 95,8% a model podľa TFIDF 86,2%). Po rozšírení modelu o predpovedanie slov reprezentujúcich potreby používateľa dosiahli ďalšie zlepšenie výkonu.

### 2.3.4 Rozšírenia modelov o metadáta otázky a odpovedí

Okrem práce s textom je vhodné pri identifikovaní sémanticky podobných otázok používať aj dodatočné dostupné metadáta (napr. informácia o kategórii otázky), prípadne si metadáta odvodiť (napr. LDA témy).

**Použitie informácie o kategórii otázky**

Cao et al. (2010) využili informáciu o kategórii v systéme Yahoo! Answers na vylepšenie úspešnosti jednotlivých modelov na získavanie podobných otázok z archívu. Každá otázka v systéme Yahoo! Answers má priradenú kategóriu a môžeme preto predpokladať, že otázky v jednej kategórii majú rovnakú tému. Otázky *Aké sú dobré reštaurácie v Bostone* a *Aké sú dobré reštaurácie v Tokiu* sú síce lexikálne veľmi podobné, no navzájom nerelevantné. Pri použití informácii o kategórii otázok (napr. *Cestovanie.USA.Boston* a *Cestovanie.Ázia.Tokio*) je vidieť, že otázky z rovnakej kategórie by mali byť vyhodnotené ako podobnejšie.

Autormi navrhnutý model je založený na intuícii, že čím je dopyt $q$ relevantnejší ku kategórii $kat$, tým je pravdepodobnejšie, že daná kategória obsahuje otázky relevantné ku $q$. Výsledné skóre pre zoradenie otázok podľa relevancie je určené interpoláciou dvoch relevantností:

1. Globálne skóre relevantnosti medzi dopytom $q$ a kategóriou $kat(d)$, ktorá obsahuje dokument $d$.
2. Lokálne skóre relevantnosti dopytu $q$ a dokumentu $d$, ktorý patrí do kategórie $kat$.

V tomto prípade sú za dokumenty považované otázky. Na určenie príslušných hodnôt skóre je možné použiť rôzne modely z kapitoly 2.3.1. Jednotlivé hodnoty relevantnosti je potrebné normalizovať. Autori použili $min-max$ normalizáciu do intervalu $<0,1>$.

Dôležitým prvkom navrhnutého modelu je počítanie lokálnej relevantnosti slov vzhľadom na otázky z kategórie. Vďaka lokálnej relevantnosti slová, ktoré sa vyskytujú vo veľa otázkach jednej kategórie strácajú dôležitosť. Experimentálne výsledky preukázali, že všetky modely dosahovali lepší výkon, keď boli skombinované s modelom využívajúcim informáciu o kategórii (zlepšenie v rozmedzí 2-5% pre metriky MAP a MRR).

**Využitie typu otázky**

Chen et al. (2013) sa venovali odlíšeniu objektívnych (faktoidných) otázok od subjektívnych názorov a sociálnych interakcií. Navrhnutý model určí pravdepodobnosť príslušnosti otázky do jednotlivých kategórií. Klasifikáciu vykonávajú s využitím algoritmu strojového učenia *support vector machines* (SVM). Okrem tradičných vlastností textu (napr. počet slov) používajú aj niekoľko ďalších metadát. Medzi najdôležitejšie metadáta patrí *kategória otázky*, *čas vytvorenia otázky* a *skúsenosti používateľa* (počet otázok, ktoré sa používateľ spýtal).

Naučenie modelu realizovali použitím učenia s čiastočným učiteľom - metódou *co-training*. Pri použití metódy využívajú fakt, že textové vlastnosti a metadáta sú na sebe nezávislé a dopĺňajú sa, čo umožňuje aplikovať metódu strojového učenia aj na veľkom množstve neoznačených dát. Výsledným návrhom ich modelu je lineárna kombinácia jazykového modelu rozšíreného o model mapovania slov a ich modelu, ktorý zachytáva zámer používateľa.

Experiment preukázal, že kombinácia textových vlastností a metadát dosahovala najlepšie výsledky (textové vlastnosti úspešnosť 69,3%, metadáta 60,9% a text + metadáta 73,1% pre objektívne otázky). Vyhodnotenie úspešnosti hľadania podobných otázok ukázalo, že ich model prekonal jazykový model a aj jazykový model rozšírený o model mapovania slov (0,26% vs 0,35% pre P@1).

**Dvojkrokový prístup hľadania sémanticky podobných otázok**

Autori Shtok et al. (2012) pracujú s hypotézou, že podobné otázky by mali mať podobné odpovede. Ich cieľom je ponúknuť používateľovi automatickú odpoveď (*odpovedanie na otázky*), a preto pracujú len s otázkami, ktoré majú označenú najlepšiu odpoveď. V prvej fáze vyhodnocujú podobnosť len na úrovni textov otázok. Na výpočet podobnosti používajú kosínusovú podobnosť. Zoradenie podobných otázok vykonávajú v dvoch krokoch. V prvom kroku vypočítajú podobnosť nadpisov otázok a ďalej pracujú len s otázkami, ktoré dosiahli určitú podobnosť $\alpha$. V druhom kroku vypočítajú podobnosť otázky na základe nadpisov aj textov otázok. Texty reprezentujú unigram modelom vo vektorovom priestore s váhami určenými pomocou TF-IDF.

### Model založený na využití kvality odpovedí

Zhang et al. (2014) upozorňujú, že v praxi text otázok a odpovedí nemožno považovať za paralelný (tak ako to robili v (Xue et al., 2008)) kvôli nízko kvalitným odpovediam, ktoré vnášajú do dát šum. Autori predpokladajú, že texty otázok a odpovedí zdieľajú spoločnú latentnú tému a na naučenie sa témy by mali mať väčší vplyv kvalitné odpovede.

Tento spôsob zároveň modeluje správanie používateľov v CQA systémoch. Učenie sa latentného priestoru je riadené kvalitou odpovedí. Všetky odpovede otázky majú vplyv na naučenie sa latentných tém, avšak čím je odpoveď kvalitnejšia, tým väčší vplyv na naučenie sa latentného priestoru bude mať. Latentné témy sú identifikované pre otázky a odpovede zvlášť a následne skombinované na základe váh jednotlivých odpovedí.

Kvalitu odpovedí definovali na základe štyroch vlastností:

1. dĺžka odpovede;
2. pomer najlepších odpovedí používateľa;
3. unikátny počet slov v odpovedi; a
4. počet prekrývajúcich sa slov v textoch otázky a odpovede.

Nad týmito vlastnosťami autori natrénovali model logistickej regresie. Za zlatý štandard pre toto učenie použili odpovede označené ako najlepšie.

Efektivitu svojho modelu porovnávajú s tradičnými modelmi (jazykový model, jazykový model rozšírený o model mapovania slov), modelmi založeným na latentných témach (Ji et al., 2012) a porovnali aj použitie čistého LDA pre naučenie tém. Autormi navrhnutý model prekonal všetky porovnávané modely o maximálne 5% (úspešnosť modelov sa pohybovala na úrovni 86-89% pre MRR metriku vyhodnotenú na Yahoo! Answers dátach). Modely založené na témach zároveň prekonali všetky modely, ktoré latentné témy nevyužívajú. Užitočným pozorovaním bolo, že použitie hlasovania používateľov na určenie kvality odpovedí viedlo k nízkej úspešnosti modelu.

### Model založený na popularite otázok

Wang et al. (2011) rozšírili jazykový model o predikciu popularity otázok čím dosiahli lepšie odporúčania. Vo svojej práci neporovnávajú úspešnosť jednotlivých jazykových modelov, avšak sústreďujú sa hlavne na zakomponovanie popularity otázok v návrhu ich metódy. Popularitu otázok určujú na základe pravdepodobnosti, že otázku zopakujú ďalší ľudia v kombinácií s počtom odpovedí, ktoré otázka dostala. Pravdepodobnosť, že otázka bude zopakovaná získavajú na základe centrality otázky v grafe otázok. Hrany medzi otázkami existujú ak otázky majú aspoň určitú hraničnú podobnosť (v práci bola použitá hodnota 0,5). Pri počítaní podobnosti používajú kosínusovú podobnosť a používajú texty otázok aj odpovedí. Úspešnosť navrhnutej metódy prekonala základné metódy o 16% pre metriku MRR (64% vs 80%).

### 2.3.5 Prístupy k určeniu vhodnosti odpovedí

Základným cieľom podpory používateľov prostredníctvom automatického odpovedania na otázky je poskytnúť na novú otázku najvhodnejšiu odpoveď. Autori Suryanto et al. (2009) sa zamerali na identifikáciu najvhodnejších odpovedí bez toho, aby sa zaoberali identifikovaním podobných otázok. Vo svojej práci používajú podobné otázky získané vyhľadávačom v CQA systéme Yahoo! Answers. Pre nájdenie najvhodnejšej odpovede využívajú črty otázok a vypočítanú expertízu autorov otázok. Na výpočet relevantnosti odpovedí k otázkam využívajú modely Vector Space Model, Okapi BM25 a jazykový model. Autori pozorovali, že všetky varianty metód, ktoré využívali kvalitu odpovedí, mali vyššiu úspešnosť než metódy, ktoré kvalitu odpovedí ignorovali.

Autori Pera a Ng (2011) využívajú rovnako ako pri identifikovaní sémanticky podobných otázok, tak aj pri zoraďovaní odpovedí korelačné faktory slov. Odpovede zoraďujú pre top 10 najpodobnejších identifikovaných otázok. Odpovede zoraďujú na základe podobnosti odpovede s predchádzajúcou otázkou, podobnosti s novou otázkou a dĺžky odpovede.

Automatickému odpovedaniu sa venovali aj autori Shtok et al. (2012), ktorí pracujú len s najlepšou odpoveďou otázky, ktorá bola označená ako najpodobnejšia prvým krokom ich metódy. Vyhodnocovanie vhodnosti použitia viacerých odpovedí nechávajú ako prácu do budúcna. V tejto fáze natrénovaným klasifikátorom vyhodnocujú, či je vhodné použiť odpoveď aj pre novú otázku. Klasifikátor využíva 95 vlastností, medzi ktoré napríklad patria počet obrázkov, dĺžka textu a LDA témy. Na základe práce s malým množstvom dát porovnávali klasifikátory náhodný les, logistickú regresiu, SVM a naivný bayesov klasifikátor. Náhodný les dosahoval najlepší výkon. Klasifikátor dosiahol presnosť 75-80%.

Ďalším typom odpovedania na otázky, ktorý autori vyhodnotili, bolo odpovedanie na dlhodobo nezodpovedané otázky. V tomto prípade použili 10 000 otázok a pozorovali, že dokázali zodpovedať približne rovnaký pomer otázok ako v prípade nových otázok. Tento záver indikuje, že veľa nezodpovedaných otázok nie je ťažkých, avšak nezodpovedané zostávajú predovšetkým preto, že si ich všimlo málo ľudí alebo boli odignorované.

Autori Belinkov et al. (2015) použili vektorovú reprezentáciu slov pre úlohu určenia vhodnosti odpovedí pre otázky v CQA systémoch. Vo svojej práci využívali dátovú sadu zo súťaže SemEval, ktorá obsahovala páry otázka-odpoveď. Autori určovali vhodnosť odpovedí kategorizovanú do troch kategórií: *dobré* odpovede, *potenciálne dobré* a *zlé*. Texty reprezentovali rôznymi vlastnosťami: textovými (napr. prekryv slov, kosínusová podobnosť frekvencie termov), vektorovými (vypočítané z *word2vec* vektorov), založenými na metadátach (či je autor odpovede rovnaký ako autor odpovede) a založenými na poradí odpovedí. Vektorové vlastnosti využívajú modely *word2vec* a *Doc2Vec*. Autori pozorovali, že kombinácia vektorových a textových vlastností mala nižší výkon ako kombinácia, kde boli použité len vektorové vlastnosti.

### 2.3.6 Predspracovanie textu

Pri práci s textami v prirodzenom jazyku je vhodné texty predspracovať. Predpokladáme, že určitý spôsob predspracovania textu použili všetci autori v predchádzajúcich opísaných prácach. Niektorí autori na predspracovanie textu špeciálne upozornili. Pri predspracovaní textu sú dôležité tieto metódy:

- tokenizácia,
- odstránenie stop slov,
  - Odstránenie sémanticky nevýznamných slov - napríklad preložky, spojky.
  - Použité v (Cao et al., 2010; Ji et al., 2012; Li a Manandhar, 2011; Zhang et al., 2014).
- určenie slovného základu slov (angl. stemming),
  - Počas tohto procesu dôjde k redukcii slov na ich základný tvar (koreň slova).
  - *Príklad:* rybárovej -> ryb.
  - Použité v (Wang et al., 2009; Zhang et al., 2014).
- lematizácia,
  - Lematizácia je pokročilejší prístup k určeniu spoločného reprezentanta slov. V tomto prípade je cieľom priradenie slovníkového tvaru každému slovu.
  - *Príklad:* rybárovej -> rybár.
  - Použité v (Li a Manandhar, 2011).
- oprava pravopisu a gramatiky.
  - Preklepy v texte prinášajú do dátovej sady šum a preto je vhodné ich odstrániť. Eliminovanie takýchto chýb však nie je triviálne a autori Li a Manandhar (2011) na to použili nástroj s otvoreným zdrojovým kódom[8].
  - Použité v (Li a Manandhar, 2011).

### 2.3.7 Spôsoby vyhodnotenia úspešnosti modelov

V predchádzajúcich vedeckých prácach pozorujeme opakujúci sa trend pri vyhodnocovaní úspešnosti modelov hľadania sémanticky podobných otázok a zoradenia odpovedí. Takmer všetky doteraz opísané práce pracovali s dátovou sadou zo systému Yahoo! Answers a pri overovaní pomocou offline experimentu boli zapojení ľudskí experti. Úlohou expertov bolo manuálne vyhodnocovať relevantnosť nájdených podobných otázok a poradie odpovedí.

Cao et al. (2010) použili dátovú sadu obsahujúcu vyše 3 000 000 otázok zo systému Yahoo! Answers. Na natrénovanie mapovaní medzi slovami použili inú dátovú sadu obsahujú 1 000 000 otázok a prislúchajúcich odpovedí. Otázky z týchto dátových sád patrili do

[8] http://www.afterthedeadline.com/

26 kategórií na prvej úrovni a 1263 kategórií na najnižšej úrovni. Autori následne použili zoznam dopytov z predchádzajúcej práce (Cao et al., 2009). Pre každý porovnávaný model zaznamenali top 20 vrátených otázok. Tieto vrátené otázky boli manuálne vyhodnotené ako relevantné alebo nerelevantné k dopytu a použité ako zlatý štandard pri vyhodnocovaní úspešnosti metód.

Duan et al. (2008) získali zo systému Yahoo! Answers vyše 500 000 otázok, ktoré mali označenú najlepšiu odpoveď. Pre získanie zlatého štandardu použili vektorový model na získanie top 20 výsledkov pre 400 náhodne vybratých otázok z archívu. Relevantnosť výsledkov k dopytu bola vyhodnotená manuálne do binárnej klasifikácie. Ako metódy pre porovnávanie úspešnosti použili vektorový model a jazykový model.

Chen et al. (2013) vyhodnotili svoj model na dvoch dátových sadách. Prvá dátová sada pochádzala zo systému Yahoo! Answers a obsahovala vyše 4 400 000 otázok. Druhá dátová sada obsahovala vyše 800 000 otázok zo systému WikiAnswers[9]. Úspešnosť metód vyhodnotili podobne ako predchádzajúce práce. V prvom kroku náhodne vybrali 50 otázok a top výsledky vyhľadávania pre každú otázku boli manuálne vyhodnotené ako relevantné a nerelevantné, čím získali zlatý štandard pre vyhodnotenie navrhnutej metódy.

Ji et al. (2012) použili vyše 500 000 otázok zo systému Yahoo! Answers. Ako dopytovacie otázky zvoli 200 náhodných otázok a zlatý štandard vytvorili spojením top 20 výsledkov pre všetky varianty modelov pre každý dopyt, ktoré následne manuálne oanotovali ako relevantné a nerelevantné. Na takto vytvorenom zlatom štandarde následne vyhodnotili úspešnosť navrhnutej metódy.

Li a Manandhar (2011) vytvorili dátovú sadu obsahujúcu 2 000 000 otázok z dvoch kategórií v systéme Yahoo! Answers. Otázky rozdelili na vyriešené a nevyriešené na základe existencie najlepšej odpovede. Autori použili 400 vyriešených otázok na vytvorenie testovacej dátovej sady. Z každej z týchto 400 otázok použili názov otázky ako dopyt používateľa a cez rozhranie systému Yahoo! Answers získali nájdené otázky na dané dopyty. Následne boli tieto výsledky manuálne vyhodnotené ako relevantné alebo nerelevantné a použité pre overenie metódy.

Pera a Ng (2011) použili dátovú sadu zo systému Yahoo! Answers, ktorá pozostávala z takmer 4 500 000 otázok. Ďalších 300 otázok použili ako otázky, ku ktorým sa snažili nájsť podobné odpovede. Týchto 300 otázok pochádzalo zo zadaní úloh z rôznych sprievodných akcií konferencií. Otázky nájdené ich metódou boli manuálne vyhodnotené ako relevantné alebo nie relevantné. Nájdenie vhodnej odpovede bolo tiež vyhodnotené manuálne, pričom zoradené odpovede boli opäť vyhodnotené ako relevantné alebo nie relevantné.

Shtok et al. (2012) vyhodnotili svoj prístup offline aj online experimentom v systéme Yahoo! Answers. V offline experimente použili vyriešené otázky z troch kategórií v systéme Yahoo! Answers obsahujúcich spolu takmer 1 000 000 otázok. V prvom kroku vybrali 1 200 otázok, ktoré nepatrili do dátovej sady (nemali označenú najlepšiu odpoveď) a pomocou

[9] http://answers.wikia.com/wiki/Wikianswers

ich modelu našli najlepšie odpovede. Takéto páry otázok a nájdených odpovedí následne oanotovali ako relevantné alebo nerelevantné. Pre každú z kategórií následne natrénovali klasifikátor a vyhodnotili úspešnosť na manuálne vytvorenom zlatom štandarde.

Pre online experiment vytvorili troch robotov, ktorých odpovede vytvorené použitím odpovedí z vyriešených otázok dosiahli lepší pomer najlepších odpovedí ako priemerní používatelia. Roboti zároveň boli vždy schopní pridať odpoveď ako prví. Okrem porovnania štatistík s priemerným používateľom autori náhodne vybrali 200 otázok z každej kategórie na ktoré ich roboti odpovedali a ich odpovede boli manuálne vyhodnotené.

Suryanto et al. (2009) vybrali 50 otázok z domény *počítače a internet* v systéme Yahoo! Answers. Otázky vybrali tak, aby pre ich text vyhľadávač v Yahoo! Answers vrátil aspoň 10 relevantných výsledkov v top 20 výsledkoch. Najlepšie odpovede z top 20 otázok vrátených vyhľadávačom v Yahoo! Answers boli manuálne vyhodnotené. V ich práci sa zameriavajú na rozlíšenie medzi relevantnosťou a kvalitou odpovedí a preto manuálne vyhodnotili odpovede ako relevantné/nerelevantné a dobré/zlé.

Wang et al. (2009) použili približne 500 000 párov vyriešených otázok a odpovedí zo systému Yahoo! Answers. Predpokladajú, že v jednom vlákne môže byť položených niekoľko otázok a preto vyextrahovali z textu otázky jednotlivé otázky vo forme jednej vety. Pre vytvorenie zlatého štandardu najskôr použili k-means zhlukovanie, aby našli podobné odpovede. Tento krok autori zdôvodňujú hypotézou, že podobné odpovede majú podobné otázky. Následne vyberajú 20 otázok z každej kategórie a manuálne označujú podobné otázky. Takto vytvorený zlatý štandard používajú pri vyhodnotení úspešnosti metódy.

Wu et al. (2014) vytvorili dve dátové sady zo systémov Yahoo! Answers (vyše 100 000 000 otázok) a Quora (takmer 650 000 otázok). Autori pracovali so záznamom dopytov používateľov z webového vyhľadávača z ktorého náhodne zvolili 1 782 dopytov. Priemerná dĺžka dopytu bola 1,94 slova. Otázky v dátovej sade zindexovali systémom Lucene.Net[10] a tento systém použili aj na nájdenie otázok na vyhľadávané dopyty. Relevantnosť otázok k dopytu bola následne vyhodnotená manuálne expertmi, ktorí otázky rozdelili do štyroch skupín. Oanotovaná dátová sada je dostupná na adrese `http://home.ustc.edu.cn/~ustcwhc/`.

Xue et al. (2008) použili dátovú sadu obsahujúcu takmer 1 000 000 otázok zo služby Wondir[11]. Na dopytovanie použili 50 otázok z webových vyhľadávačov a relevantnosť nájdených podobných otázok manuálne vyhodnotili, čím získali zlatý štandard pre následne vyhodnotenie úspešnosti metód.

Zhang et al. (2014) vytvorili dve dátové sady zo systémov Yahoo! Answers (takmer 1 200 000 otázok) a Baidu Knows[12] (vyše 770 000 otázok). Otázky a odpovede zindexovali nástrojom Lucene.Net a nástrojom tiež našli v priemere 15 podobných otázok. Relevantnosť nájdených podobných otázok vyhodnotili manuálne. Rovnaký spôsob použili aj autori Wu et al. (2014).

[10] `http://lucenenet.apache.org/`
[11] Služba Wondir už v súčasnosti nie je dostupná.
[12] `http://zhidao.baidu.com/`

Dátová sada, ktorú použili je dostupná na adrese `http://research.microsoft.com/en-us/people/wuwei/wuwei.aspx`.

Zhou et al. (2011) použili vyše 500 000 vyriešených otázky zo systému Yahoo! Answers. Na naučenie pravdepodobností mapovaní slov využili 1 000 000 párov otázka-odpoveď z inej dátovej sady. Testovaciu množinu vytvorili na základe náhodného výberu 300 otázok a pomocou vektorového modelu identifikovali top 20 najpodobnejších otázok, ktoré následne manuálne vyhodnotili ako relevantné alebo nerelevantné. Na takto získanom zlatom štandarde následne vyhodnotili úspešnosť metód.

### 2.3.8 Metriky vyhodnotenia úspešnosti metód

Autori v predchádzajúcich prácach používali na vyhodnotenie úspešnosti modelov tradičné metriky z domény získavania informácií. Rozlišujeme dva typy použitých metrík: metriky pracujúce so zoradeným zoznamom a klasifikačné metriky. Klasifikačné metriky sú *správnosť*, *presnosť*, *úplnosť*; a metriky pracujúce so zoradeným zoznamom sú *presnosť na N*, *MAP*, *MRR*, *R-presnosť*, *nDCG*. Prehľad použitia v konkrétnych prácach je uvedené v tabuľke 2.2 na strane 9. V nasledujúcom prehľade jednotlivé používané metriky stručne opisujeme:

- Správnosť (angl. Accuracy),
  - Správnosť predstavuje pomer korektne klasifikovaných záznamov ku celkovému počtu záznamov.
- Presnosť (angl. Precision),
  - Presnosť je zlomok vrátených relevantných dokumentov ku všetkým vráteným dokumentom.
- Úplnosť (angl. Recall),
  - Úplnosť je zlomok vrátených relevantných výsledkov ku všetkým relevantným dokumentom v dátovej sade.
- Presnosť na N (angl. Precision at N - P@N),
  - Označuje sa aj ako P@N a hovorí o pomere relevantných výsledkov na prvých N pozíciách. V predchádzajúcich prácach sa najčastejšie používalo P@1,5,10.
- MAP (z angl. Mean Average Precision),
  - MAP zvýhodňuje prístupy, ktoré vrátia relevantné výsledky medzi prvými a ktoré vracajú výsledky správne zoradené.
- MRR (z angl. Mean Reciprocal Rank),
  - MRR reflektuje ako hlboko sa relevantný výsledok nachádza v zozname.
  - Vzorec pre MRR je v tvare:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{2.1}$$

kde $|Q|$ predstavuje počet dopytov, $rank_i$ predstavuje pozíciu prvého relevantného výsledku pre $i$-ty dopyt.

- R-Presnosť (angl. R-Precision),
  - R presnosť je presnosť vrátenia relevantných dokumentov po tom, čo bolo vrátených $R$ dokumentov, kde $R$ je celkový počet relevantných dokumentov.
- nDCG (z angl. Normalized Discounted Cumulative Gain),
  - Táto metrika prikladá väčšiu dôležitosť správnemu zotriedeniu dokumentov na začiatku zoznamu, než na jeho konci.
  - Vzorec pre nDCG je nasledovný:

$$nDCG = \frac{DCG_p}{IDCG_p} \tag{2.2}$$

Vzorec pre $DCG_p$ a $IDCG_p$ je podobný:

$$\{DCG_p, IDCG_p\} = \sum_{i=1}^{\{p,|REL|\}} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{2.3}$$

Rozdiel je v tom, že raz počítame hodnoty pre zoradenie získané z vyhodnocovanej metódy a v druhom prípade pre ideálne zoradenie. Hodnota $p$ určuje počet pozícií, na ktorých vyhodnocujeme hodnotu $nDCG$, a $rel_i$ je relevancia záznamu na danej pozícii. $|REL|$ predstavuje ideálne zoradený zoznam dokumentov po pozíciu $p$.

## 2.4 Diskusia

V tejto kapitole sme identifikovali potrebu podpory používateľov v CQA systémoch. Medzi najčastejšie spôsoby podpory používateľov v CQA systémoch patrí smerovanie nezodpovedaných otázok a získavanie vyriešených otázok z archívu. V oblasti využitia archívu otázok a odpovedí sme sa bližšie venovali podoblasti vyhľadávania sémanticky podobných otázok.

Veľkým problémom vyhľadávania sémanticky podobných otázok je tzv. *lexikálna medzera* - jav, kedy používatelia vyjadria rovnakú otázku rôznymi slovami. Predchádzajúci výskum priniesol niekoľko modelov, ktoré sa s týmto problémom snažia vysporiadať. Pozorovali sme štyri hlavné skupiny modelov pre vyhľadávanie sémanticky podobných otázok: 1) jazykový model, 2) model založený na mapovaní slov, 3) model založený na syntaktických štruktúrach a 4) model založený na latentných témach. Jednotlivé modely môžu pracovať buď len s textami otázok, alebo aj textami odpovedí. Pozorovali sme, že využitie kvalitných odpovedí prinieslo zlepšenie výkonu navrhovaných modelov.

Najúspešnejším modelom pre identifikovanie sémanticky podobných otázok sa javí model založený na latentných témach navrhnutý v práci (Zhang et al., 2014). Zároveň sme identifikovali nový spôsob určenia podobnosti textov založený na vektorovej reprezentácii slov *word2vec*.

Pozorovali sme problém so zlatým štandardom pre vyhodnotenie úspešnosti metód. Vo všetkých predchádzajúcich prácach autori využili manuálne vyhodnotenie relevantnosti vrátených podobných otázok. Čiastočnú výnimku tvorí práca autorov Shtok et al. (2012), ktorí pre vyhodnotenie úspešnosti online experimentu použili spätnú väzbu od používateľov systému Yahoo! Answers.

# Kapitola 3

# Online komunity študentov

Rozvíjanie elektronického vzdelávania (e-vzdelávanie) po roku 2000 prinieslo online dostupnosť edukačného materiálu a otvorené možnosti na štúdium širokej mase používateľov internetu. Online obsah umožňuje používateľom samovzdelávať sa kdekoľvek a kedykoľvek, čo je výhodné hlavne pre pracujúcich ľudí, ktorí cítia potrebu celoživotného vzdelávania, alebo pre študentov bez finančných možností zúčastniť sa klasického vzdelávacieho procesu. Existujú viaceré typy edukačného materiálu dostupného na webe. Sú to predovšetkým:

- publikované materiály z prednášok a cvičení populárnych univerzít,
- články o vede,
- video kurzy,
- kurzy snažiace sa napodobniť spôsob výučby na univerzitách, a
- akademické vedecké diskusné fóra.

V posledných rokoch nárast popularity zaznamenali predovšetkým online kurzy poskytované prestížnymi svetovými univerzitami. Do takýchto kurzov býva naraz zapísaných niekoľko tisíc študentov a označujú sa ako hromadné otvorené online kurzy (angl. Massive Open Online Course - MOOC). Zapísaní študenti vytvárajú prirodzenú komunitu formovanú okolo kurzu do ktorého sa zapísali. Iným typom populárnych e-vzdelávacích platforiem sú napríklad Khan Academy[1] a Udemy[2], ktoré fungujú mimo univerzitného systému a kladú dôraz na umožnenie používateľovi prejsť si materiál kurzu vlastným tempom.

Špecifikom online kurzov je, že študenti sa takmer nikdy reálne nepoznajú, pretože kurzov sa zúčastňujú študenti z celého sveta. Na druhej strane, študenti zapísaní do denného študijného programu na svojej univerzite majú tiež tendenciu formovať online komunitu. Takéto komunity slúžia predovšetkým na šírenie informácií, riešenie problémov s preberaným učivom a zdieľanie tipov, čo by v podobnom rozsahu nebolo bez online prostriedkov možné.

V minulosti boli centrom online komunít univerzitných študentov fóra, avšak s rastúcou popularitou sociálnych sietí komunikácia na týchto fórach začala stagnovať. V súčasnosti

[1] https://www.khanacademy.org/

[2] https://www.udemy.com/

môžeme pozorovať, že napríklad študenti informatiky na Fakulte informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave medzi sebou komunikujú predovšetkým na sociálnej sieti Facebook[3], kde si vytvárajú ročníkové skupiny.

Inováciu v oblasti online komunít študentov predstavuje CQA systém Askalot[4], ktorý je zameraný na doménu vzdelávania a obsahuje na to určené podporné nástroje (Srba a Bielikova, 2015). Askalot je vyvíjaný na Fakulte informatiky a informačných technológií a do produkčného prostredia bol nasadený v letnom semestri akademického roka 2013/2014. Od svojho spustenia sa doň prihlásilo 1676 používateľov, ktorí položili 790 otázok a odpovedali na ne prostredníctvom 1057 odpovedí (údaje z mája 2017).

## 3.1 Hromadné otvorené online kurzy

Hromadné otvorené online kurzy (MOOC) sú založené na myšlienkach otvoreného vzdelávania - princípe, že vedomosti by mali byť šírené zadarmo, a túžba po vzdelaní by mala byť adresovaná bez demografických, ekonomických a geografických obmedzení (Zheng et al., 2015). Hromadné otvorené online kurzy sú často poskytované platformami, ktoré zlučujú viaceré kurzy na jedno miesto. Môže ísť napríklad o kurzy poskytované jednou univerzitou (napr. Lagunita[5] od Stanford University) alebo o kurzy od viacerých univerzít.

Medzi najväčších a najznámejších poskytovateľov MOOC kurzov patria Coursera[6] a edX[7]. Systém edX bol založený v roku 2012 Massachusettskou technickou univerzitou (MIT) a Harvardovou univerzitou. V súčasností sa na edX nachádzajú kurzy od vyše 90 univerzít a inštitúcií. Coursera bola založená profesormi zo Stanfordovej univerzity tiež v roku 2012 a na konci roka 2015 mala vyše 15 miliónov používateľov.

S vývojom MOOC kurzov sa začali rozlišovať dva typy MOOC kurzov - cMOOC a xMOOC (Alario-Hoyos et al., 2014; Cui a Wise, 2015; Zheng et al., 2015). Kým kurzy typu cMOOC kladú dôraz na interakcie medzi študentami (napr. vzájomné riešenie problémov a práca na projektoch), v xMOOC verzii je prúd vedomostí jednosmerný od inštruktora ku študentom a vzájomné interakcie študentov sú väčšinou obmedzené na diskusiu vo fórach. Kurzy typu xMOOC majú taktiež tradičnejšiu štruktúru a to predovšetkým vo forme:

- vopred určeného sylabu,
- predpripravených vyučovacích videí,
- pripravených testov.

Hlavnou výhodou MOOC kurzov je ich otvorenosť - sú dostupné cez web a poväčšinou je možné sa ich zúčastniť zadarmo s možnosťou zaplatiť si overený certifikát o úspešnom

[3] https://www.facebook.com
[4] https://askalot.fiit.stuba.sk/
[5] https://lagunita.stanford.edu/courses
[6] https://www.coursera.org/
[7] https://www.edx.org/

dokončení kurzu. Okrem tradičného obsahu vzdelávacích kurzov (videá, materiály na čítanie a kvízy) väčšina MOOC platforiem obsahuje diskusné fóra na podporu interakcie medzi študentami a s inštruktormi kurzov. Z dôvodu veľkého počtu zapísaných študentov však dochádza k nemožnosti individuálneho prístupu inštruktorov ku jednotlivým študentom (Shatnawi et al., 2014). Samotní študenti preto musia navzájom spolupracovať pri hľadaní riešení na svoje otázky a problémy.

MOOC kurzy trpia predovšetkým problémom nízkeho percenta študentov, ktorí kurz úspešne dokončia. Nakoľko sú kurzy dostupné zadarmo, niektorí používateľa sa po zaregistrovaní do kurzu ani raz neprihlásia. Tento fakt je spôsobení aj tým, že používatelia musia niekedy čakať až niekoľko týždňov, kým sa kurz začne. Ďalších niekoľko desiatok percent študentov nepokračuje v kurze po prvom týždni a úspešne dokončiť kurz sa podarí len približne 5-10% študentom (Alario-Hoyos et al., 2014; Anderson et al., 2014; Coetzee et al., 2014).

Medzi hlavné príčiny tohto problému patrí nedostatok času študentov na štúdium, slabá sebadisciplína v dodržiavaní termínov kurzu a chýbajúca motivácia pre dokončenie kurzu. Zheng et al. (2015) vo svojej práci na základe rozhovorov s účastníkmi MOOC kurzov identifikovali 8 hlavných príčin na nedokončenie kurzu:

- väčšie časové nároky, než bolo deklarované v sylaboch;
- zložitý náročný obsah;
- nedostatok času;
- nedostatok motivácie;
- chýbajúci pocit komunity;
- vplyv odporúčania kurzu od známych;
    - Keď študent vedel, že ani jeho známy nedokončil daný kurz, mal tendencie hľadať výhovorky, aby kurz tiež nedokončil.
- neskorý začiatok kurzu;
- cieľ študovať len určitú časť kurzu.
    - Akonáhle používateľ naplnil svoje očakávania/potreby, tak kurz opustil.

Paradoxne, niektoré online kurzy dosahujú veľké percento študentov, ktorí úspešne dokončia kurz. Feng et al. (2015) sa venovali analýze fóra v sociologickom kurze s 18 959 používateľmi, ktorý bol poskytovaný Šanghajskou univerzitou a kredity získané za jeho dokončenie boli uznávané lokálnymi univerzitami. Tento kurz väčšina študentov úspešne dokončila, čo možno pravdepodobne pripísať práve uznávaniu kreditov za úspešné dokončenie kurzu lokálnymi univerzitami.

## 3.2 Komunikácia v MOOC kurzoch

Hlavným komunikačným kanálom v MOOC kurzoch sú diskusné fóra, ktoré sú súčasťou platforiem poskytujúcich hromadné otvorené online kurzy. Okrem diskusných fór zvyknú inštruktori kurzov využívať emailovú komunikáciu na oznámenie dôležitých termínov alebo upozornení na zaujímavé diskusie. V súčasnosti niektorí inštruktori využívajú aj sociálne siete, predovšetkým vo forme skupín na sociálnej sieti Facebook, prípadne používajú účet na sociálnej sieti Twitter[8] na zverejňovanie informácií o novom publikovanom obsahu.

Kurz CS50[9] poskytovaný Harvardovou univerzitou zameraný na výučbu základov programovania v systéme edX je extrémnym príkladom kurzu, ktorý obsahuje množstvo komunikačných kanálov. Inštruktori tohoto kurzu sa rozhodli natívne diskusné fórum vo svojom kurze vôbec nepoužívať a namiesto toho vyžívajú primárne inštanciu Stack Exchange platformy (`http://cs50.stackexchange.com/`) a niekoľko ďalších nástrojov a služieb:

1. Facebook (`https://facebook.com/groups/cs50`),
2. Gitter (`gitter.im/cs50/x`),
3. Imzy (`https://www.imzy.com/cs50/`),
4. LinkedIn (`https://www.linkedin.com/groups/7437240/profile`),
5. Reddit (`https://www.reddit.com/r/cs50`),
6. Slack (`https://cs50x.slack.com/`), a
7. Twitter (#cs50, @cs50).

Celkovo je tak komunikácia rozdelená na ôsmich rôznych miestach - troch sociálnych sieťach (Facebook, Twitter, LinkedIn), troch komunitných platformách (Stack Exchange, Imzy a Reddit) a dvoch nástrojoch na komunikáciu v reálnom čase.

Analýzou využívania 5 komunikačných kanálov v rámci jedného MOOC kurzu sa venovali autori v prácach (Alario-Hoyos et al., 2013; Alario-Hoyos et al., 2014). Ich práca sa venuje 9 týždňovému kurzu, ktorý bol ponúkaný cez španielsku platformu MiríadaX[10]. Inštruktori v tomto kurze využili 2 interné (diskusné fórum a CQA systém) a 3 externé komunikačné nástroje (Facebook, Twitter, MentorMob[11]). Autori pozorovali, že najaktívnejší používateľ pre každú službu bol iný. Fórum zaznamenalo najväčšiu aktivitu zo strany študentov a druhá najväčšia aktivita bola pozorovaná na sociálnej sieti Facebook.

I keď sa na prvý pohľad zdá, že veľa rôznych prostriedkov umožňuje používateľovi vybrať si ten, ktorý mu najviac vyhovuje, v závere článku (Alario-Hoyos et al., 2014) autori upozorňujú na problém informačného preťaženia. Študenti (a rovnako inštruktori) museli preskúmať niekoľko portálov, aby našli nový obsah, čo vyústilo až do niekoľkých sťažností

[8] `https://twitter.com/`

[9] `https://courses.edx.org/courses/HarvardX/CS50x3/2015/`

[10] `https://miriadax.net/`

[11] `https://www.mentormob.com/`

od študentov. Inštruktori sa sťažovali predovšetkým na problém detekcie urgentných problémov a opakujúce sa diskusie v rôznych systémoch (Alario-Hoyos et al., 2014).

Napriek tomu, že MOOC kurzy natívne neobsahujú chatové rozhranie, autori Coetzee et al. (2014) upozorňujú, že v doméne online vzdelávania sa kombinácia asynchrónnej (diskusné fóra) a synchrónnej (chat, súkromné správy) komunikácie osvedčila. Autori vo svojej práci skúmajú použitie chatu v jednom z kurzov, ktorý bol poskytovaný na platforme edX. Prvým skúmaným problémom bolo zobrazenie chatu - prvá skupina používateľov videla chat ako záložku v kurze a druhej bol chat zobrazený ako súčasť rozhrania študijného materiálu. Oba prístupy zobrazovali tú istú diskusnú miestnosť. Kontrolnej skupine študentov nebol chat zobrazovaný vôbec. Aktivita v chate bola nízka (8,2 správy za hodinu) a 24,9% správ bolo odoslaných inštruktormi kurzu. Analýza aktivity a dosiahnutých výsledkov študentov z jednotlivých skupín preukázala, že použitie chatu nemalo žiadny výrazný efekt na výslednú známku študenta, dokončenie kurzu, či aktivitu v diskusnom fóre.

## 3.3 Použitie fór v MOOC kurzoch

Diskusné fóra predstavujú dôležitý prvok hromadných otvorených online kurzov. Aktivita používateľov na fórach je však nízka. Počet študentov, ktorí sa zapoja do fóra predstavuje približne 5-15% (Alario-Hoyos et al., 2014), avšak až 50% týchto študentov úspešne dokončilo kurz (Alario-Hoyos et al., 2014). Podobne autori Feng et al. (2015) pozorovali, že čím viac sa študenti zapájali do fóra, tým dosiahli lepšie výsledky.

Fóra na webe sa používajú vo viacerých kontextoch a je otázne, či v MOOC fórach prebieha diskusia alebo majú skôr typ systémov pre otázky a odpovede. Analýza autorov Anderson et al. (2014) ukázala, že diskusia na fórach je väčšinou priamočiara, obsah rastie s príchodom nových používateľov a komunikácia medzi používateľmi, ktorí prví do určitého vlákna prispeli je slabá.

Jednou z výrazných charakteristík diskusných fór v hromadných otvorených online kurzoch je rôznorodý charakter obsahu jednotlivých tém. Fóra sa používajú predovšetkým na pýtanie sa otázok, diskusie k obsahu kurzu a predstavovanie sa. Alario-Hoyos et al. (2013) zistili, že niektorí používatelia sa do kurzu zaregistrovali len preto, aby pridávali na fórum sťažnosti, ktoré často nesúviseli s témou kurzu, ale napríklad politikou. Iní používatelia zase robili reklamu svojim produktom snažiac sa využiť množstvo potenciálnym zákazníkov. Negatívnym správaním, ktoré autori pozorovali bolo aj zverejňovanie odpovedí na testy.

### 3.3.1 Identifikácia druhov tém v diskusných fórach

Diskusné fóra v MOOC kurzoch majú väčšinou definované kategórie, ktoré zoskupujú príspevky s určitou témou. V praxi sa však často stáva, že študenti toto rozdelenie nedodržiavajú a preto je potrebné byť schopný automaticky identifikovať skutočnú tému príspevku.

Hecking et al. (2015) vo svojej práci používali rozdelenia na: otázky týkajúce sa kurzu, všeobecné diskusné príspevky a iné. Na identifikovanie typov príspevku vytvorili klasifikátor naučený na štrukturálnych (napr. pozícia príspevku v diskusnom vlákne) a sémantických vlastnostiach príspevkov (napr. výskyt otáznikov). Autori pozorovali, že iba tretina používateľov fóra sa zapojila do príspevkov týkajúcich sa obsahu kurzu.

Identifikovaniu tém, ktoré súvisia s obsahom kurzu sa venovali aj Cui a Wise (2015). Autori vytvorili klasifikačný model, ktorý na základe lingvistických charakteristík textu dokázal úspešne rozlíšiť relevantné a nerelevantné témy. V tomto kurze bolo fórum štruktúrované do kategórií:

- všeobecná diskusia,
- otázky a odpovede, úlohy,
- technické problémy,
- otázky k študijnému materiálu, a
- predstavenie sa študentov.

Dátovú analýzu robili len nad kategóriami *všeobecná diskusia* a *otázky a odpovede*.

Prvý príspevok v každej diskusnej téme bol manuálne vyhodnotený ako relevantný alebo nerelevantný ku kurzu. Následne naučili binárny SVM klasifikačný model na vlastnostiach extrahovaných z textu. Dodatočne použité vlastnosti boli informácie o počte videní, hlasovanie používateľov a počet odpovedí. Autori pozorovali, že len 28% tém v kategórii *všeobecná diskusia* sa týkalo kurzu. Inštruktori kurzu odpovedali na 14% všetkých tém, pričom adresovali 18% tém, ktoré sa týkali obsahu kurzu, z čoho vyplýva, že študenti nie vždy dostali odpoveď od inštruktorov, keď mali relevantný problém s kurzom. Úspešnosť klasifikačného modelu sa pridaním dodatočných informácií nezlepšila.

### 3.3.2 Sémantická analýza obsahu diskusných fór

Sémantickú analýzu diskusného fóra v čínskom kurze s vysokou úspešnosťou vykonali Feng et al. (2015). Po predspracovaní dát priradili slovám relevantným ku kurzu pozitívnu váhu a nerelevantným negatívnu. Autori pozorovali, že približne tretina príspevkov na fóre nebola relevantná k obsahu kurzu. Zaujímavým zistením bolo, že najväčší pomer relevantných a nerelevantných príspevkov mali študenti, ktorý dosiahli priemerný výsledok v kurze. Ďalším zaujímavým pozorovaním bolo, že inštruktori kurzu, respektíve ich asistenti odpovedali na existujúce témy minimálne, avšak vytvárali nové témy v ktorých predovšetkým informovali o nových veciach týkajúcich sa kurzu.

Yang et al. (2015) sa venovali identifikovaniu zmätenosti študentov z obsahu kurzu. Využili pri tom dáta zachytávajúce interakciu študentov so študijným materiálom prostredníctvom klikania myši a lingvistické vlastnosti textu príspevkov na diskusných fórach. Po naučení klasifikátora na manuálne oanotovaných dátach analyzovali, aký má vplyv zmätenosť študentov na nedokončenie kurzu a pozorovali závislosť medzi týmito dvoma atribútmi. V prípade, že

je zmätenosť študentov dostatočne rýchlo adresovaná a študent dostane na svoju otázku uspokojivú odpoveď je vplyv zmätenosti na nedokončenie kurzu čiastočne eliminovaný.

### 3.3.3 Automatická podpora vzdelávania v diskusných fórach

Hlavnou výzvou MOOC kurzov z pohľadu vzdelávania je veľké množstvo študentov, ktorí pripadajú na jedného učiteľa. Tento pomer často dosahuje až 5000:1. Autori Shatnawi et al. (2014) sa problém rozhodli adresovať identifikovaním príspevkov relevantných ku kurzu a poskytnutím automatickej spätnej väzby študentom (automatické odpovedanie na otázky). Ich riešenie spočíva vo vytvorení nástroja na ontológie.

Každý kurz má definovanú vlastnú ontológiu, ktorú vytvoria z rôznych zdrojov podľa témy kurzu. V prvom kroku identifikujú najčastejšie slová v jednotlivých zdrojoch a následne manuálne vytvoria ontológiu aj s atribútmi a spätnou väzbou pre jednotlivé slová. Pri spracovaní jedného príspevku následne hľadajú zhodu v ontológii a v prípade zhody vrátia používateľovi preddefinovanú spätnú väzbu. Napriek tomu, že ich systém dosahoval dobré výsledky, problém takéhoto prístupu je potreba manuálneho vytvorenia ontológie a definovanie spätnej väzby. Jedným z možných riešení je použitie CQA systému, kde znalosti vytvorí komunita v priebehu času.

## 3.4 Typy používateľov v MOOC kurzoch

Hromadné otvorené online kurzy často dosahujú zápis niekoľkých tisícov študentov, ktorí pochádzajú z rôznych demografických a geografických oblastí. Ako sme už uviedli v kapitole 3.1, len približne 10% používateľov úspešne dokončí kurz. Identifikovanie rôznych typov používateľov je užitočné z viacerých dôvodov. Najaktívnejší používatelia väčšinou preberú rolu mentorov a dopĺňajú inštruktorov v poskytovaní personalizovanej podpory ostatným študentom. Ak poznáme typy používateľov, môžeme ich vhodne prepájať (napr. smerovanie otázok - kapitola 2.2.1) a vytvárať tak užšie spolupracujúce komunity, alebo vhodne priraďovať partnerov na vzájomné vyhodnotenie úloh.

Alario-Hoyos et al. (2014) rozdelili používateľov *na základe ich výkonu v kurze* a analyzovali ich množstvo aktivity na rôznych komunikačných nástrojoch. Komunikácia študentov nemala vplyv na ich zaradenie do jednotlivých profilov. Autori definovali tri hlavné kategórie používateľov:

1. Používatelia, ktorí sa zaregistrujú a pozrú si maximálne niekoľko videí.
2. Používatelia, ktorí nedokončia kurz, avšak dokončia jeho časť.
3. Používatelia, ktorí dokončia kurz.

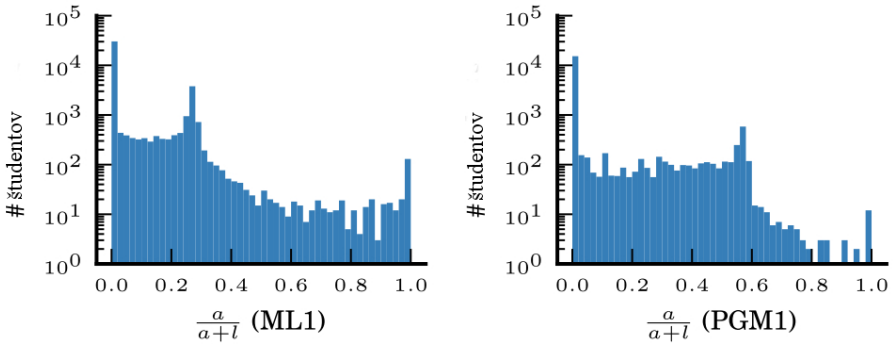
Tieto kategórie ešte delia do siedmych podkategórií, ktoré sú zobrazené v tabuľke 3.1. V poslednom stĺpci tabuľky je uvedené percento výskytu profilu v populácii študentov analyzova-

ného kurzu. Najväčší počet používateľov, ktorí úspešne dokončili kurz bol v podkategóriách 7 a 4. Používatelia, ktorí sa nezapájali do náročných aktivít, mali problémy s úspešným dokončením kurzu - len 31,8% takýchto používateľov úspešne ukončilo kurz.

**Tabuľka 3.1:** Podkategórie typov používateľov v MOOC kurzoch z práce (Alario-Hoyos et al., 2014).

| Kategória | Podkategória | % |
|---|---|---|
| Zaregistrujú sa a pozrú si maximálne niekoľko videí | (1) Nepozrú si ani jedno video, ani nevyriešia úlohu. | 26,56 |
| | (2) Pozrú si videá, avšak neriešia testy a úlohy. | 42,04 |
| Nedokončia kurz, avšak dokončia jeho časť | (3) Venujú sa aktivitám kurzu, avšak kurz nedokončia. | 13,21 |
| | (4) Venujú sa aktivitám kurzu, avšak kurz dokončia neúspešne a navyše sa do kurzu zaregistrovali až po jeho spustení. | 5,38 |
| | (5) Venujú sa aktivitám kurzu, kurz nedokončia a zaregistrovali sa do kurzu po jeho spustení. | 5,22 |
| Dokončia kurz | (6) Dokončia kurz, avšak nezapájajú sa do aktivít, ktoré vyžadujú veľkú námahu. | 1,57 |
| | (7) Dokončia kurz so všetkými jeho aktivitami. | 6,02 |

Na základe *interakcie študentov s materiálom kurzov* identifikovali profily správania aj autori Anderson et al. (2014). Základné aktivity podľa ktorých určujú profily sú sledovanie videa a odoslanie úlohy alebo testu. Dôležitým faktorom je pomer jednotlivých aktivít. Tento pomer vypočítali pre šesť kurzov na Coursera platforme. V histogramoch rozloženia všetkých kurzov pozorovali tri body s maximálnymi hodnotami - body na krajoch histogramu a jeden v strede. Dva histogramy z práce (Anderson et al., 2014) uvádzame na obr. 3.1. Označenia ML1 a PGM1 sú označením kurzov, ku ktorým sa daný histogram vzťahuje.




**Obr. 3.1:** Histogram pomeru počtu odovzdaných úloh ($a$) voči počtu odovzdaných úloh a pozretých výučbových lekcií ($a + l$). Prevzaté a preložené z (Anderson et al., 2014).

Na základe tohto pozorovania je možné identifikovať tri prirodzené typy správania sa používateľov:

1. *Diváci*, ktorí predovšetkým len sledujú študijný materiál.
2. *Riešitelia*, ktorí odovzdávajú úlohy a testy bez predchádzajúceho štúdia materiálu.
3. *Všestranní*, ktorí majú vyváženú aktivitu medzi oboma aktivitami.

Následne odvodzujú ešte ďalšie dve ďalšie kategórie:

4. *Zberateľia*, ktorí videá sťahujú, avšak nemusia si ich aj pozrieť.
5. *Nezúčastnení*, ktorí sa do kurzu zaregistrovali, avšak ich aktivita je minimálna.

Pri analýze času registrácie používateľov zistili, že 60% používateľov sa registrovalo pred začiatkom kurzu a 18% používateľov po jeho skončení. Takýchto používateľov označujú ako *archeológovia*. Autori zisťovali, aký typ používateľov navštevuje diskusné fóra. Zastúpenie jednotlivých profilov v populácii študentov, ktorí si prečítali aspoň jednu tému na fóre bolo výrazne odlišné, než zastúpenie v populácii kurzu. Aktívni používatelia mali vyššiu tendenciu zúčastňovať sa aktivít na fóre. Kým *nezúčastnení* tvorili 50% registrovaných študentov kurzu, na fóre predstavovali len 10% a až 90% *všestranných* používateľov čítalo fórum. Autori tiež zistili, že používateľ, ktorý založil novú tému, dosahoval nižšie záverečné skóre, než používatelia ktorí mu v téme odpovedali.

### Najaktívnejší používatelia diskusných fór

Podobne ako je to v mnohých online komunitách, aj v diskusných fórach MOOC kurzov je možné identifikovať malú skupinu používateľov, ktorí sú zodpovední za veľké množstvo príspevkov. Vlastnosti takýchto používateľov skúmali autori Huang et al. (2014). Najaktívnejších používateľov fór označujú ako *superprispievatelia* a rozlišujú ich podľa:

1. počtu príspevkov, ktoré pridali,
2. hodnotenia príspevkov hlasovaním ostatných používateľov,
3. reputácie na základe hodnotenia príspevkov.

Autori pracovali s dátami zo 44 kurzov, ktoré boli ponúkané na platforme Coursera. Coursera určuje reputáciu používateľov ako odmocninu z priemerného množstva hlasov na všetkých príspevkoch používateľa. Za *superprispievateľov* považujú top 5% používateľov podľa vyššie uvedených troch metrík. Ich prvým zistením je, že *superprispievatelia* viac odpovedali na existujúce témy, ako vytvárali nové diskusné témy. Z tohto pozorovania možno usudzovať, že *superprispievatelia* viac pomáhali ostatným študentom, ako sami hľadali pomoc. Zároveň pozorovali, že ich príspevky sú dlhšie, než príspevky priemerných používateľov. Dôležitým zistením bolo, že správanie sa *superprispievateľov* bolo konzistentné naprieč viacerými kurzami a teda sa väčšinou nejedná len o špecifický záujem používateľa o konkrétnu oblasť, na ktorú je kurz zameraný.

Najzaujímavejšou sledovanou vlastnosťou *superprispievateľov* je porovnanie ich úspešnosti v kurze s ostatnými študentami. Všetky kategórie *superprispievateľov* dosahovali lepšie výsledky než priemerní študenti a *superprispievatelia* na základe reputácie dosahovali úplne najlepšie výsledky. Ďalším zistením bolo, že *superprispievatelia* zvyknú prispievať do témy skôr ako priemerní používatelia avšak ich príspevky dosahujú podobné hodnotenie od ostatných používateľov ako priemerní používatelia.

## 3.5 Diskusia

Online komunity študentov sa stali neoddeliteľnou súčasťou vzdelávacieho procesu. Študenti na univerzitách majú prirodzenú tendenciu združovať sa aj v online priestore a vymieňať si užitočné informácie a poznatky. Druhým typom online komunít študentov sú komunity vytvorené v doméne e-vzdelávania.

Inštruktori MOOC kurzov sa snažia vyjsť študentom v ústrety a umožňujú im komunikovať prostredníctvom viacerých služieb. Autori Alario-Hoyos et al. (2014) však pozorovali, že rozptýlenie komunikácie na viacero miest viedlo k často sa opakujúcim diskusiám a inštruktori aj študenti sa sťažovali na nutnosť sledovania viacerých služieb naraz.

Komunikácia vo väčšine MOOC kurzov sa odohráva na diskusných fórach, ktoré sú súčasťou platforiem poskytujúcich MOOC kurzy. Zistili sme, že iba malé percento používateľov sa zapája do fóra. Viaceré štúdie preukázali, že títo používatelia dosahujú lepšie výsledky než používatelia, ktorí sa do komunikácie na fórach nezapájajú.

Dôležitým poznatkom je, že iba časť príspevkov v diskusných fórach sa týka kurzu. Množstvo takýchto príspevkov predstavuje 30-66% v závislosti od konkrétneho kurzu. Jedným z výrazných problémov MOOC kurzov je nízke percento používateľov, ktorí dokončia kurz. Analýza diskusných fór ukázala, že problémy študentov na fórach sú len málokedy adresované inštruktormi kurzov a ich asistentmi.

Používateľov MOOC kurzov môžeme deliť podľa viacerých aspektov. Jednou z vlastností je množstvo aktivity študentov pri interakcii s obsahom kurzu. Bolo zistené, že najväčší počet používateľov, ktorí úspešne dokončia kurz sú najaktívnejší študenti. Používateľov je možné deliť aj podľa spôsobu práce s materiálom. Autori práce (Anderson et al., 2014) zistili, že aktívni používatelia, ktorí študujú materiál aj odovzdávajú úlohy a riešia kvízy, tvorili až 90% populácie používateľov, ktorí si nejakú tému ma fóre prečítali alebo sa do niektorej z nich zapojili.

MOOC kurzy majú niekoľko problémov, ktoré sú spôsobené hlavne veľkým počtom študentov na jedného vyučujúceho. Inštruktori a ich asistenti nie sú schopní sledovať všetok vznikajúci obsah na fórach MOOC kurzov a poskytovať personalizovanú podporu jednotlivým študentom. Problém s informačným preťažením majú aj študenti, ktorí sú ochotní prevziať rolu mentora a pomáhať ostatným študentom.

Z tohto dôvodu je potrebné hľadať automatizované nástroje, ktoré by boli schopné podporiť študentov v reálnom čase. Jedným z takýchto úsilí je tvorba ontológie kurzu a automatizované poskytovanie spätnej väzby študentom (Shatnawi et al., 2014). Ďalším prístupom, ktorý by mohol pomôcť zlepšiť súčasnú situáciu, sú nové typy nástrojov (napríklad CQA systémy), ktoré poskytujú prehľadnejšiu štruktúru a množstvo funkcií, ktoré na súčasných fórach v MOOC kurzoch nie sú dostupné (napr. nasledovanie aktivity používateľov, adaptívna podpora spolupráce).

# Kapitola 4

# Použitie CQA systémov v MOOC kurzoch

Existujú viaceré nástroje, ktoré riešia problém pýtania sa a odpovedanie na otázky študentov v online kurzoch. Jedným z takýchto nástrojov je systém OpenStudy[1] (Ram et al., 2011). Študenti si po registrácii môžu nájsť partnerov na štúdium, kolaboratívne pracovať na úlohách, pýtať sa otázky a odpovedať na ne. V systéme OpenStudy boli vytvorené študijné skupiny napríklad pre kurzy ponúkané pod záštitou MIT Open Courseware[2].

Podporným nástrojom sledujúcim komunikáciu študentov a poskytovaním interaktívnej spätnej väzby je systém Bazaar (Adamson et al., 2014), ktorý bol použitý na niektorých MOOC kurzoch, no je možné ho používať napríklad aj vo vzdelávacom systéme Moodle[3]. V tomto systéme si inštruktor kurzu nastaví konverzačných agentov, ktorí sledujú diskusiu študentov v IM (angl. instant messaging) nástroji a poskytujú spätnú väzbu v reálnom čase predovšetkým vo forme usmerňovania konverzácie a snahe iniciovať komunikáciu od všetkých prítomných študentov.

CQA systémy na otvorenom webe priniesli do diskusií prebiehajúcich na fórach štruktúru a lepšiu organizáciu obsahu. Podobný prínos dokážu priniesť CQA systémy aj do komunikácie v rámci MOOC kurzov. Na základe nám známych informácií je jediným poskytovateľov MOOC kurzov, ktorý natívne obsahuje CQA systém platforma MiríadaX. Ako sme uviedli v kapitole 3.2, bolo zistené, že CQA systém v jednom kurze poskytovanom na tejto platforme bol využívaný študentami a inštruktori kurzu pravidelne odpovedali na pridané otázky (Alario-Hoyos et al., 2014).

Hoci iné platformy poskytujúce MOOC kurzy neobsahujú CQA systémy, niektorí inštruktori sa rozhodli pre svoje kurzy používať externé CQA systémy. Jedným z príkladov je kurz CS50 poskytovaný Harvardovou univerzitou, ktorý sme uviedli v kapitole 3.2. Inštruktori kurzu CS50 sa okrem iného rozhodli na komunikáciu používať inštanciu Stack Exchange

[1] http://openstudy.com/

[2] http://ocw.mit.edu/courses/openstudy/

[3] https://moodle.com/

platformy. Táto inštancia je verejne prístupná na otvorenom webe[4] a na prezeranie otázok a odpovedí nie je potrebné byť účastníkom kurzu. Spôsob používania tejto inštancie však nesedí s vysokým štandardom obsahu, ktorým sa prezentujú zvyšné systémy Stack Exchange platformy. Príspevky, ktoré tam študenti pridávajú totiž majú podobnú štruktúru, ako príspevky na fórach MOOC kurzov (kapitola 3.3), a teda tu nájdeme aj všeobecné diskusie a sociálne interakcie. Nízka kvalita obsahu, vyvolala patričný nesúhlas niektorých používateľov Stack Exchange platformy. Príkladom je diskusia v meta sekcii Stack Exchange platformy: `http://meta.stackexchange.com/q/231208`.

## 4.1 CQA systémy určené pre doménu vzdelávania

Okrem otvorených CQA systémov (ktorý sa používa na kurze CS50) a CQA systémoch priamo integrovaných v platformách, ktoré poskytujú MOOC kurzy (napr. MiríadaX), existujú aj špeciálne CQA systémy určené pre doménu vzdelávania. Príkladmi takýchto systémov sú Piazza[5] a Green Dolphin (Aritajati a Narayanan, 2013). Tieto systémy boli priamo vytvorené pre doménu vzdelávania a poskytujú uzavretý priestor pre účastníkov kurzu. Piazza je systém umožňujúci inštruktorom vytvorenie vlastného CQA systému. Študenti v tomto systéme majú možnosť kolaboratívne upravovať odpovede na otázky a viesť diskusie k jednotlivým otázkam. Systém Piazza v súčasnosti používa viac ako 1 000 vzdelávacích inštitúcií vrátane najlepších svetových univerzít.

Green Dolphin je experimentálny CQA systém určený pre študentov programovania. Aj v tomto systéme študenti kolaboratívne upravujú jednu odpoveď na otázku. Medzi ďalšie významné vlastnosti tohto systému patrí integrovanie programátorských úloh do systému, pričom študentom je po odoslaní úlohy poskytnutá spätná väzba o kvalite kódu. Ďalšou vlastnosťou CQA systému Green Dolphin je identifikovanie expertných používateľov. Interakcia používateľov v systéme je založená na ekonomike bodov, ktoré je možné získať napríklad za pýtanie sa a odpovedanie na otázky a následne ich minúť na smerovanie otázok k jednotlivým používateľom, alebo získanie rýchlejšej odpovede na otázku.

Medzi CQA systémy určené pre doménu vzdelávania patrí aj systém Askalot (Srba a Bielikova, 2015) vyvíjaný na Fakulte informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave. Tento systém podporuje učiteľov napríklad zvýrazňovaním ich odpovedí a možnosťou vyhodnocovania kvality obsahu, ktorý pridajú študenti. V akademickom roku 2015/2016 bol autor tejto práce členom tímu, ktorý pracoval na integrácii Askalotu do systému edX (Srba a Bielikova, 2016b).

[4] `http://cs50.stackexchange.com/`

[5] `https://piazza.com/`

## 4.2 Rozdiely medzi fórami v MOOC kurzoch a obsahom štandardných otvorených CQA systémov

V kapitole 3.3 sme identifikovali, že aj na fórach MOOC kurzov prebieha proces komunitného zdieľania vedomostí prostredníctvom pýtania sa otázok a odpovedania na ne. Hlavné rozdiely medzi otázkami a odpoveďami v MOOC kurzoch a otázkami a odpoveďami v CQA systémoch sú:

- *prirodzené autority*,
  - Procesu výmeny vedomostí v diskusných fórach sa zúčastňujú jednak študenti, ako aj inštruktori kurzov a ich asistenti. Inštruktori a asistenti predstavujú typ používateľa, ktorého môžeme automaticky považovať za expertného na tému kurzu. Vo všeobecných online komunitách študentov sa do komunity môžu rovnako zapájať učitelia a teda tiež majú prirodzenú autoritu v komunite. V štandardných otvorených CQA systémoch niečo takéto neexistuje a je potrebné pre každého používateľa odhadnúť expertízu.
- *prirodzene sa viažuci obsah*,
  - V bežných CQA systémoch sa používateľ pýta v explicitne nedefinovanom kontexte a ak chceme zistiť viac o problematike jeho otázky, musíme nejakým spôsobom spracovať text otázky a napríklad podľa kľúčových slov dohľadať informácie napr. z webovej encyklopédia Wikipedia[6]. Výhodou MOOC kurzov je, že študenti sa môžu pýtať otázky priamo pri obsahu kurzov, vďaka čomu máme k dispozícii dodatočný obsah vo forme študijných textov alebo textového prepisu video a audio materiálov.
- *periodické opakovanie sa otázok*.
  - Otázky sa zvyknú opakovať v oboch typoch systémov. Špecifikom diskusií v online komunitách študentov je však výrazné periodické opakovanie otázok spôsobené opakovaním jednotlivých kurzov v čase (ročná alebo semestrálna báza v prípade klasických univerzít a niekoľkomesačná až ročná v prípade MOOC kurzov). Kým používatelia v CQA systémoch majú k dispozícii vyhľadávanie v archíve otázok a odpovedí, MOOC kurzy začínajú s prázdnym obsahom fóra a tak študenti nemajú možnosť dostať sa k vedomostiam a vyriešeným problémom z predchádzajúcich iterácií kurzu.

[6] https://www.wikipedia.org/

## 4.3 Diskusia

Identifikovali sme viaceré online systémy umožňujúce študentom pýtať sa otázky a odpovedať na ne. Jediným poskytovateľom MOOC kurzov, ktorý natívne podporuje CQA systém je platforma MiríadaX. Inštruktori kurzov však prejavujú záujem používať externé CQA systémy vo svojich kurzoch.

Medzi procesom pýtania sa a odpovedania na otázky v CQA systémoch a v online komunitách študentov sú určité výrazné rozdiely. V online komunitách študentov identifikujeme prirodzené autority (učitelia), ktorých je však vzhľadom na počet študentov veľmi málo. Problém duplicitných otázok v MOOC kurzoch je ešte výraznejší ako v CQA systémoch, nakoľko sa otázky periodicky opakujú, tak ako sa opakujú iterácie jedného kurzu. Tento problém je spôsobený aj nedostupnosťou archívneho obsahu v nových kurzoch.

Výhodou otázok v MOOC kurzoch je povinné vyplnenie textovej časti, čím odpadá problém hľadania podobných otázok len na základe krátkeho niekoľkoslovného dopytu. Autori Wu et al. (2014) napríklad pozorovali, že len 81% otázok v systéme Yahoo! Answers a 58% otázok v systéme Quora malo text otázky vyplnený. Ďalšou výhodou je prirodzene sa viažuci obsah k otázke vo forme výučbového materiálu kurzu, ktorý je možné využiť pre lepšie pochopenie otázky študenta.

V MOOC kurzoch pozorujeme potrebu automatickej podpory študentov z dôvodu veľkého počtu študentov na jedného učiteľa. Výrazným problémom je opakovanie sa otázok v každej iterácii kurzu. Použitie automatického odpovedania na otázky s využitím odpovedí z archívu otázok a odpovedí sa javí ako užitočné riešenie podpory študentov. Pri procese automatického odpovedania môžeme navyše využiť dodatočné informácie a špecifiká, ktorými sa MOOC kurzy odlišujú od klasických CQA systémov.

# Kapitola 5

# Konceptuálny návrh metódy pre automatické odpovedanie na otázky

Cieľom navrhnutej metódy je podporiť študentov v online komunitách využitím archívu otázok a odpovedí. Po analýze možností podpory používateľov v kapitole 2 sme sa rozhodli pre automatické odpovedanie na otázky. Problémom duplicitných otázok trpia CQA systémy aj online komunity študentov, kde sa opakované otázky navyše prirodzene objavujú periodicky, tak ako sa opakuje výučba v čase. Kvôli duplicitným otázkam dochádza k preťaženiu inštruktorov kurzov, ktorí nemajú čas sa venovať všetkým otázkam. Nami navrhnutá metóda je zameraná na použitie predovšetkým v doméne MOOC kurzov, ale môže byť použitá aj v inom vzdelávacom prostredí, kde sa študenti pýtajú otázky týkajúce sa učiva.

Cieľom našej práce je využiť metadáta dostupné v doméne vzdelávania pre zvýšenie úspešnosti nájdenia podobných otázok. Po analýze prístupov z predchádzajúcich vedeckých prác (kapitola 2.3) sme sa rozhodli použiť textovú podobnosť a špecifiká online komunít študentov (kapitola 4.2) pre identifikovanie sémanticky podobných otázok a následné zoradenie odpovedí. Našu hypotézu definujeme ako:

> Ak v procese nájdenia podobných otázok a vhodných odpovedí využijeme metadáta dostupné v online komunitách študentov, budeme schopní lepšie odporúčať podobné otázky a odpovede než model, ktorý nezohľadňuje tieto špecifické metadáta. Zároveň budeme schopní odpovedať na signifikantný počet otázok s vysokou mierou presnosti.

V prípade použitia nami navrhnutej metódy v reálnom systéme musíme poskytovať odpovede na nové otázky s vysokou presnosťou. Pri snahe o vysokú presnosť nesmieme zabúdať aj na pomer otázok, ktoré sa nám podarí zodpovedať. Ak by sme boli schopní zodpovedať len 1% otázok, samotná réžia s nastavovaním metódy v reálnom systéme by bola pravdepodobne väčšia než samotný prínos metódy. Empiricky odhadujeme, že už pri 5-10% zodpovedaných otázkach by bolo použitie našej metódy prínosné.

Proces automatického odpovedania na otázky delíme na dve hlavné nezávislé časti:

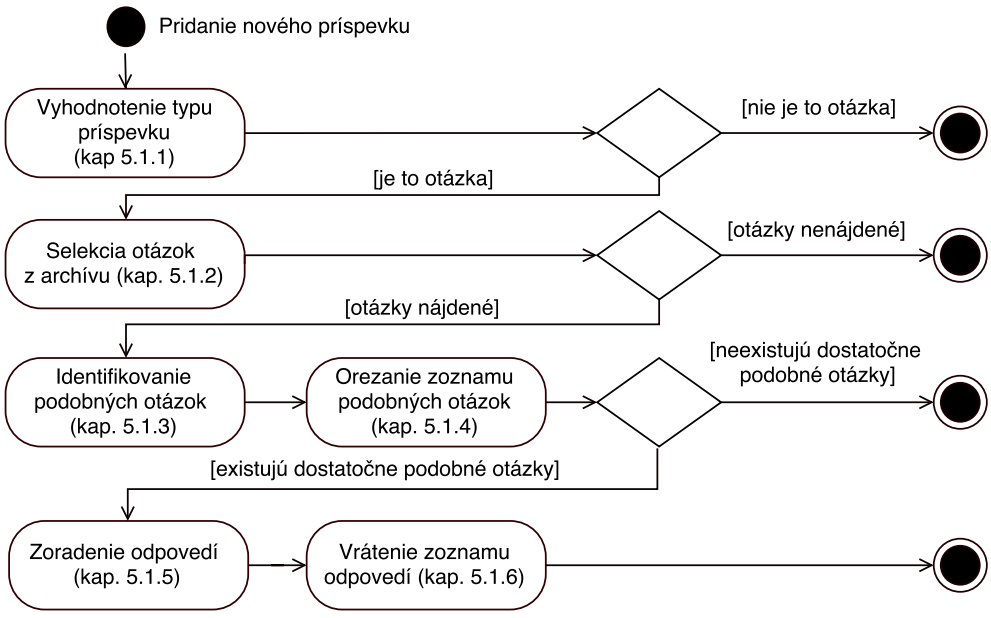
1. identifikovanie podobných otázok len na základe otázok, a
2. zoradenie odpovedí z podobných otázok.

Pre dva nezávislé kroky sme sa rozhodli na základe analýzy predchádzajúcich prác v kapitole 2.3. Autori v práci (Zhang et al., 2014) poukazujú na fakt, že text odpovedí vnáša do procesu nájdenia otázok šum. Vo viacerých predchádzajúcich prácach bol tiež využívaný len text otázok (Cao et al., 2010; Duan et al., 2008; Li a Manandhar, 2011). Ďalším dôvodom pre toto rozhodnutie je, že dodržiavanie štruktúry otázka-odpoveď v online komunitách študentov, predovšetkým na fórach MOOC kurzov, býva slabé. Vďaka odfiltrovaniu málo podobných otázok zároveň predpokladáme, že v procese zoraďovania odpovedí budeme zoraďovať odpovede už len z relevantných otázok.

Celkový návrh metódy sa skladá z viacerých krokov:

1. vyhodnotenie typu príspevku (kapitola 5.1.1),
2. selekcia otázok z archívu (kapitola 5.1.2),
3. identifikovanie podobných otázok (kapitola 5.1.3),
4. orezanie zoznamu podobných otázok (kapitola 5.1.4),
5. zoradenie odpovedí (kapitola 5.1.5),
6. vrátenie zoznamu odpovedí (kapitola 5.1.6).

Proces, v ktorom jednotlivé kroky vystupujú je zobrazený na diagrame aktivít na obr. 5.1




**Obr. 5.1:** Pohľad na proces zodpovedania novej otázky prostredníctvom našej metódy

## 5.1 Opis jednotlivých krokov metódy

### 5.1.1 Vyhodnotenie typu príspevku

Pri analýze komunikácie na MOOC fórach v kapitole 3.3 sme zistili, že pomerne malé množstvo príspevkov v diskusných fórach MOOC kurzov sa týka vyučovaného materiálu. V prvom rade nás preto zaujíma, či je príspevok faktoidná otázka alebo diskusia. Ak sa jedná napríklad o oznámenie nových informácii alebo otázku typu "Na akej škole študujete?", tak nemá zmysel hľadať podobné príspevky a vykonávanie našej metódy končí. V našej práci sa plánujeme spoliehať na informáciu poskytnutú používateľom, kde používateľ pri vytváraní príspevku zvolí, či sa jedná o faktoidnú otázku alebo nie. Alternatívne riešenie je využiť metódy strojového učenia (napr. binárna klasifikácia) pre predikciu typu príspevku.

### 5.1.2 Selekcia otázok z archívu

Cieľom metódy je automaticky odpovedať na otázku a preto má zmysel pracovať s otázkami z archívu, ktoré už majú odpoveď. V závislosti od typu systému/kurzu, v ktorom by bola metóda použitá, môžeme v tomto kroku vybrať aj iba otázky, ktoré majú označenú najlepšiu odpoveď a teda ich môžeme považovať za vyriešené. Ak žiadne vyhovujúce otázky neexistujú (napr. keď bol kurz nedávno spustený), tak vykonávanie našej metódy končí.

### 5.1.3 Identifikovanie podobných otázok

Úlohu nájsť podobné otázky sme sa rozhodli riešiť binárnym klasifikátorom, ktorý určí, či sú dve otázky podobné alebo nie. Dve otázky považujeme za podobné, ak sa na základe ich nadpisov a textov zdá, že odpoveď z jednej otázky by mohla byť vhodná ako odpoveď aj pre druhú otázku. Cieľom preto nie je nájsť len priamo duplicitné otázky, ale aj dostatočne podobné otázky. Takýmto spôsobom hľadali podobné otázky pre automatické odpovedanie aj v prácach (Shtok et al., 2012; Pera a Ng, 2011). Príklad podobných otázok uvádzame v tabuľke 5.1. Na základe analýzy prác v kapitole 2.3 sme sa rozhodli použiť nasledujúce typy čŕt (zvýraznené sú špecifické pre vzdelávaciu doménu):

1. textová podobnosť;
2. pomer dĺžok textov otázok;
3. pomer počtu stop slov;
4. pomer počtu otáznikov;
5. počet zhodných slov;
6. počet zhodných slov na základe typu (podstatné mená, slovesá, opytovacie zámená);
7. **podobnosť na základe kategórie**;
8. **podobnosť na základe času**.

**Tabuľka 5.1:** Príklad podobných otázok, ktoré nie sú úplne duplicitné, avšak v rámci našej metódy ich považujeme za podobné. Otázky boli vybraté a preložené z kurzu *MITx: 6.00.2x*.

| **Otázka** |
|---|
| *Nadpis*: Aký kurz by som mal spraviť po kurze 6002x?<br><br>Ahoj! Veľmi som si užil tento kurz, napriek kvízu a extrémne náročnej sady problémov 5 ;) Mohol by niekto odporučiť aké ďalšie kurzy si mám spraviť, tu na edX-e, MIT OCW, alebo inej platforme? Alebo mohol by niekto odporučiť kurz, kde sa môžem dozvedieť viac o teórii grafov? Mám rád dátovú vedu a rád by som si zlepšil svoje znalosti z informačných technológií. Ďakujem. |
| **Podobná otázka** |
| *Nadpis*: Čo ďalej po 6.00.1x a 6.00.2x<br><br>Dôkladne som si užil oba kurzy a popri tom sa naučil veľa vecí. Dúfal som, že na konci nájdem pár videí o tom, kam ďalej pokračovať. Aké kurzy, knihy, materiály môžeme použiť pre ďalšie budovanie vedomostí v oblasti dátovej vedy a strojového učenia. |
| **Odpoveď z podobnej otázky** |
| Zdieľam, čo sa mi podarilo nájsť. Mohol by si sa hlbšie venovať algoritmom: s týmto MIT kurzom (odkaz), alebo touto knihou (odkaz), takisto vieš nájsť nejaké MOOC kurzy, avšak väčšina nepoužíva Python. Takisto by si mohol študovať dátovú vedu - tu je kurz The analytics edge na edX-e, ktorý práve začal a mal by výborný, avšak používa R. Ak sa chceš držať jazyka Python, ja som práve začal čítať túto knihu (odkaz) a zatiaľ je skvelá. CS 109 od Harvardu má taktiež množstvo dobrého materiálu online. Ak niekto nájde ďalšie materiály, prosím pridajte ich sem. Veľa šťastia tým, ktorí pokračujú na tejto ceste :) |

Črta *podobnosť na základe kategórie* (7) bola použitá v práci Cao et al. (2010), no v doméne vzdelávania predpokladáme, že je jej dôležitosť ešte vyššia. Výučba v predmetoch býva často organizovaná po týždňoch, pričom každý týždeň je zameraný na určitú tému. V informačnom systéme zvykne byť výučbe v jednom týždni kurzu (alebo semestra) priradená jedna kategória, ktorá má svoje podkategórie reprezentujúce príslušné podtémy. Ak existuje takáto hierarchia kategórií, tak môžeme predpokladať, že podobné otázky týkajúce sa určitej témy budú v kategóriách, ktoré sú pri sebe v hierarchii kategórii bližšie.

Črta *podobnosť na základe času* (8) je špecifická výlučne pre doménu vzdelávania. Ak sa obsah kurzu/predmetu príliš nezmení, môžeme prirodzene očakávať podobné otázky vytvorené v rovnakom čase relatívne ku začiatku výučby. Ak by sa obsah kurzu/predmetu zmenil v poradí obsahu jednotlivých týždňov, môžeme podobnosť na základe času identifikovať podľa rovnakého názvu kategórií, do ktorých otázky patria.

Prvá uvedená črta sa venuje podobnosti textov otázok. V kapitole 2 sme identifikovali viacero spôsobov ako vypočítať podobnosť textov. Najjednoduchší spôsob je transformovať texty pomocou algoritmu TF-IDF. Medzi zložitejšie reprezentácie patrí model LDA a word2vec. Všetky tri modely je možné použiť v našej metóde. LDA model sa v predchádzajúcich prácach ukázal ako najúspešnejší (Li a Manandhar, 2011) a bol používaný vo viacerých ďalších prácach (Zhang et al., 2014; Shtok et al., 2012; Ji et al., 2012).

### TF-IDF model

TF-IDF je model reprezentácie textu na základe *TF* - frekvencie pojmu (angl. term frequency) a *IDF* - inverznej frekvencie pojmu v dokumente (angl. inverse document frequency). Medzi výslednou vektorovou reprezentáciou textov, ktoré použitím tejto metodiky vzniknú je následne možné vypočítať podobnosť napríklad pomocou Jaccardovej, kosínusovej podobnosti alebo Euklidovskej vzdialenosti. Vzorec pre výpočet IDF pojmu je nasledovný:

$$\mathrm{IDF}_t = \log \frac{N}{\mathrm{df}_t} \tag{5.1}$$

kde $|N|$ je počet dokumentov a $\mathrm{df}_t$ je počet dokumentov, ktoré obsahujú pojem $t$.

### LDA model

LDA model (angl. Latent Dirichlet allocation) je model založený na identifikovaní latentných tém a pravdepodobností patrenia dokumentov do jednotlivých tém. Postup vypočítania podobnosti otázok pomocou latentných tém je nasledovný:

1. Naučenie sa LDA tém z textov existujúcich otázkach
    - Problém s naučením LDA tém máme v prípade, že je otázok nedostatok a teda nemáme LDA témy na čom vytvoriť (problém studeného štartu (angl. cold-start problem)). V takýchto prípadoch môžeme témy naučiť na otázkach z predchádzajúcej iterácie kurzu alebo dostupnom výučbovom materiále kurzu (učebné texty, prepis videí).
2. Určenie pravdepodobností príslušnosti otázky do jednotlivých LDA tém.
3. Nájdenie najpodobnejších otázok podľa pravdepodobností patrenia do LDA tém.

Jedným z problémov tohto prístupu je, že LDA témy je potrebné pravidelne prepočítať, aby odrážali všetky novopribudnuté otázky.

### word2vec model

V prípade použitia word2vec modelu je proces jednoduchší:

1. Vypočítanie a uloženie vektora otázky spriemerovaním vektorov všetkých slov.
2. Nájdenie najpodobnejších otázok na základe vektorov otázok pomocou Jaccardovej, kosínusovej podobnosti alebo Euklidovskej vzdialenosti.

Alternatívnou možnosťou je priame použitia modelu, ktorý vie transformovať celý text a nie len slová (napr. GenSim[1]). Výhodou tohto prístupu je, že funguje už od pridania prvej otázky a nie je potrebné pravidelné prepočítavanie hodnôt.

[1] `https://radimrehurek.com/gensim/`

### 5.1.4 Orezanie zoznamu podobných otázok

Výstupom predchádzajúceho kroku našej metódy je zoznam otázok, ktoré klasifikátor označil ako podobné. Väčšina implementácií populárnych klasifikátorov umožňuje získať pravdepodobnosť patrenia záznamu do jednotlivých klasifikačných tried. Práve túto informáciu v tomto kroku použijeme na vyfiltrovanie najpodobnejších otázok.

V zozname otázok ponecháme len tie otázky, ktoré spĺňajú určitú podobnosť $\gamma$. Ak máme dostatok dát, hodnotu $\gamma$ vieme určiť pomocou jej systematického menenia a vyhodnotenia počtu a podobnosti otázok, ktoré ostanú v zozname podobných otázok. Ak nemáme dostatok dát (angl. cold-start problem), hodnotu $\gamma$ určíme empiricky. Ak v zozname podobných otázok neostanú žiadne otázky, tak vykonávanie našej metódy končí.

### 5.1.5 Zoradenie odpovedí

Vstupom do tohto kroku je zoznam otázok, ktoré spĺňajú minimálnu podobnosť $\gamma$ a cieľom tohto kroku je zoradiť ich odpovede podľa vhodnosti pre novú otázku. V tomto kroku môžeme pracovať s rôznymi podmnožinami odpovedí z otázok. Môžeme použiť len odpovede, ktoré majú najväčšiu podobnosť s textom otázok alebo s odpoveďami, ktoré majú najviac hlasov a podobne. Autori v prácach (Pera a Ng, 2011; Suryanto et al., 2009) zoraďovali všetky odpovede z nájdených podobných otázok. Autori Suryanto et al. (2009) však overili svoju metódu len pre zoraďovanie odpovedí označených ako najlepšie.

Pre zoradenie odpovedí sme sa rozhodli použiť techniku strojového učenia *učenie sa zoraďovať* (Liu, 2009) (angl. Learning to Rank, L2R). Táto technika bola použitá v práci (Belinkov et al., 2015) pre výber najvhodnejšej odpovede na otázku. Aplikovaním L2R vytvoríme model, ktorý dokáže zoraďovať položky podľa ich vhodnosti. Existujú tri typy L2R prístupov:

1. bodový prístup (angl. pointwise),
    - V prípade bodového prístupu je cieľom modelu odhadnúť určitú absolútnu hodnotu, ktorá reprezentuje kvalitu položky (napr. počet pozitívnych hlasov od používateľov.)
2. párový prístup (angl. pairwise),
    - Cieľom párového prístupu je určiť, ktorá z dvoch položiek je vhodnejšia.
3. prístup založený na zozname (angl. listwise).
    - Tento prístup sa snaží optimalizovať jednu z vyhodnocovacích metrík (napr. MRR, MAP) pre všetky položky v dátovej sade.

Z dôvodu nízkeho počtu hlasov používateľov v doméne online kurzov, nie sme schopní získať zoradený zoznam všetkých odpovedí ani poznať absolútnu hodnotu kvality odpovede. Jedinou možnosťou použitia L2R je preto párový prístup. Zoradené páry odpovedí

potrebné pre trénovanie modelu vieme získať manuálnym porovnaním vhodnosti dvoch odpovedí.

Pre naučenie modelu potrebujeme získať jednotlivé črty odpovedí. Navrhujeme použitie čŕt získaných po analýze predchádzajúcich prác v kapitole 2.3, a doplnenie o črty špecifické pre doménu vzdelávania (črty 14 a 15). Zoznam typov čŕt je nasledovný:

1. textová podobnosť novej otázky a odpovede;
2. textová podobnosť pôvodnej otázky a odpovede;
3. počet zhodných slov;
4. počet zhodných slov na základe typu (podstatné mená, slovesá);
5. skóre odpovede vo forme hlasovania komunity;
6. informácia, či bola odpoveď označená za najlepšiu;
7. dĺžka odpovede;
8. počet otáznikov v odpovedi;
9. počet odkazov na webové stránky v odpovedi;
10. počet obrázkov v odpovedi;
11. počet stop slov v odpovedi;
12. pomer najlepších odpovedí používateľa;
13. informácia, či je autor odpovede ten istý používateľ, ktorý položil otázku;
14. **informácia, či je autor odpovede učiteľ**;
15. **informácia, či je odpoveď z tej istej iterácie kurzu**.

### 5.1.6 Vrátenie zoznamu odpovedí

Vstupom tohto kroku je zoradený zoznam odpovedí. Výstupom môže byť jedna najpodobnejšia odpoveď, alebo viacero odpovedí v závislosti pre aký typ použitia metódy sa rozhodneme. Prvou z možností použitia našej metódy je odpovedanie na novú otázku prostredníctvom vytvorenia novej odpovede v systéme. V tomto prípade potrebujeme iba jednu odpoveď a výstupom tohto kroku metódy je preto len jedna odpoveď.

Druhým spôsobom použitia je odporúčanie viacerých odpovedí pýtajúcemu sa používateľovi po vytvorení novej otázky. Vytváranie novej odpovede v systéme, ktorá by obsahovala texty viacerých odpovedí nie je vhodné z dôvodu dlhého výsledného textu. Vhodnejším riešením je preto zobrazovať určitý počet odpovedí len pýtajúcemu sa používateľovi a umožniť mu označiť správnu odpoveď, ktorá sa následne vytvorí reálne v systéme.

Rozšírením druhého spôsobu je odporúčať odpovede už počas toho, ako používateľ vytvára otázku. V prípade odporučenia existujúcej odpovede ešte pred vytvorením otázky nemusí k samotnému vytvoreniu novej otázke v systéme vôbec dôjsť, čím dosiahneme ušetrenie času, ktorý by bol strávený čítaním textu novej otázky ostatnými používateľmi.

## 5.2 Návrh realizácie procesu strojového učenia

Skôr než použijeme strojové učenie v krokoch *identifikovanie podobných otázok* a *zoradenie odpovedí*, potrebujeme dáta predspracovať. Na základe analýzy v kapitole 2.3.6 navrhujeme texty otázok a odpovedí predspracovať pomocou *lematizácie* a *odstránenia stop slov*. Lematizáciu pred určením slovného základu slova (angl. stemming) uprednostňujeme, pretože použitie tohto procesu nám vytvorí slová, ktoré presnejšie odrážajú pôvodné slovo, vďaka čomu vieme lepšie identifikovať zhodné slová v dvoch textoch.

V kapitolách 5.1.3 a 5.1.5 sme navrhli niekoľko typov čŕt, ktoré budú použité pre identifikovanie podobných otázok a zoradenie odpovedí. Samotná reprezentácia týchto čŕt môže byť rôzna a preto v tejto časti navrhujeme konkrétne riešenie. Pri použití metód strojového učenia je potrebné mať hodnoty čŕt normalizované. Za týmto účelom navrhujeme použitie škálovania a centrovania hodnôt jednotlivých čŕt.

**Črty pre identifikovanie podobných otázok**

Prvou navrhnutou črtou je *textová podobnosť*. Otázky v systéme sa skladajú z dvoch častí - nadpisu a textu otázky. Tieto dve časti môžeme spojiť, čím nám vznikne ďalšia reprezentácia otázky (Shtok et al., 2012). Štvrtou reprezentáciou otázok, ktorú navrhujeme použiť je extrakcia opytovacích viet (vety končiace otáznikom), nakoľko predpokladáme, že v takýchto vetách je zachytená podstata otázky. Textovú podobnosť navrhujeme počítať pre tieto štyri reprezentácie.

Aj ďalšie navrhnuté textové črty - *pomer dĺžok textov otázok*, *pomer počtu stop slov*, *pomer počtu otáznikov*, *počet zhodných slov* a *počet zhodných slov na základe typu* - vieme vyhodnotiť pre štyri rôzne reprezentácie otázok. Zároveň môžeme okrem vzájomného pomeru počtov napr. zhodných slov vypočítať aj pomer počtu zhodných slov ku celkovému počtu slov v otázke.

Posledné dva typy čŕt sú špecifické pre doménu vzdelávania. Prvý typ *podobnosť na základe kategórie* umožňuje využiť hierarchickú organizáciu výučbového materiálu. Za týmto účelom navrhujeme štyri konkrétne črty ako reprezentantov podobnosti na základe kategórie:

1. podobnosť ako umiestnenie kategórie v hierarchii kategórií;
2. binárna informácia, či sú kategórie otázok rovnaké;
3. binárna informácia, či sú kategórie otázok rovnaké naprieč dvoma iteráciami kurzu/predmetu;
4. binárna informácia, či sú otázky z rovnakej iterácie kurzu/predmetu.

Druhým typom črty špecifickej pre doménu vzdelávania je *podobnosť na základe času*. Tu navrhujeme dve konkrétne reprezentácie tejto črty:

1. rozdiel v čase pridania otázok relatívne ku prvej otázke v kurze/predmete; a
2. rozdiel v čase pridania prvej otázky v kategóriách, do ktorých otázky patria.

Reprezentácia otázok 4 spôsobmi (text, nadpis, text + nadpis, časť s opytovacími vetami) spôsobuje, že dostávame veľký počet čŕt. Z tohto dôvodu je potrebné identikovať najdôležitejšie črty (angl. feature selection) a len tie použiť na vstupe do klasifikátora. Ak by sme redukciu počtu čŕt nevykonali, narazili by sme na problém strojového učenia, ktorý sa označuje ako *prekliatie dimenzionality* (angl. curse of dimensionality). Tento problém značí stav, kedy máme príliš málo záznamov na to, aby sme nimi dostatočne pokryli viacrozmerný priestor a vo výsledku by sa klasifikátor nemal na čom naučiť rozdiely medzi záznamami.

**Črty pre zoradenie odpovedí**

Pri vyhodnocovaní podobnosti otázok a odpovedí môžeme tiež využiť štyri identifikované reprezentácie textov otázok. Aj v tomto prípade môžeme okrem absolútneho počtu napr. podstatných mien vypočítať aj pomer počtu podstatných mien ku všetkým slovám v odpovedi. Všetky zvyšné typy čŕt sú samoopisujúce a ich realizácia je priamočiara. Podobne ako aj v predchádzajúcom kroku, aj tu navrhujeme identifikovanie najdôležitejších čŕt a ich následné použitie za účelom zredukovania dimenzionality.

## 5.3 Sumarizácia navrhnutej metódy

Navrhnutá metóda sa skladá z dvoch hlavných komponentov - 1) klasifikátora na identifikovanie podobných otázok a 2) techniky strojového učenia - *učenie sa zoraďovať* - pre zoradenie odpovedí. Tieto dva komponenty sú na sebe nezávislé. Ako vstup do komponentu pre zoraďovanie odpovedí vchádzajú otázky, ktoré spĺňajú určitú minimálnu podobnosť získanú z klasifikátora na identifikovanie podobných otázok.

Vykonávanie navrhnutej metódy sa deje len pre príspevky, ktoré sú otázkami. Pri vyhodnotení typu príspevku sa spoliehame na informáciu poskytnutú používateľom pri vytváraní príspevku. Výstupom našej metódy je zoradený zoznam odpovedí podľa ich vhodnosti pre použitie ako odpoveď na novú otázku. Počet prvkov v zozname závisí od spôsobu aplikovania našej metódy v produkčnom systéme.

Navrhnutá metóda je univerzálna pre použitie na texty v akomkoľvek jazyku, pre ktorý existuje natrénovaný *word2vec* model. Výnimku tvorí predspracovanie textu, ktoré je jazykovo závislé.

# Kapitola 6

# Realizácia metódy automatického odpovedania na otázky

Konceptuálny návrh metódy opísaný v predchádzajúcej kapitole nie je viazaný na žiadny programovací jazyk, algoritmus strojového učenia, a ani nie je obmedzený na konkrétnu online komunitu študentov. V tejto kapitole sa venujeme našej realizácii navrhnutej metódy, ktorej jednotlivé body bližšie špecifikujeme.

V rámci realizácie navrhnutej metódy sme sa rozhodli pracovať s obsahom fóra v systéme poskytujúcom MOOC kurzy edX[1]. Medzi hlavné dôvody prečo sme sa pre edX rozhodli patrí fakt, že edX je druhým najväčším poskytovateľom MOOC kurzov a na diskusnom fóre podporuje rozdelenie typu príspevkov na otázky a diskusie. Benefitom, prečo sme si vybrali edX bol aj fakt, že v akademickom roku 2015/2016 bol autor tejto publikácie súčasťou tímu, ktorý pracoval na integrácii systému Askalot (Srba a Bielikova, 2015) do systému edX a teda mal so systémom edX určité skúsenosti.

Napriek tomu, že kurzy v systéme edX nedisponujú CQA systémom, predpokladáme, že rozdelenie príspevkov na fóre na otázky a diskusie nám umožňuje zamerať sa len na faktoidné otázky, ktoré by boli súčasťou CQA systému. Počas realizácie navrhnutej metódy sme pracovali s kurzami, ktorých obsah bol v anglickom jazyku. Metódu sme realizovali v programovacích jazykoch R a Ruby.

## 6.1 Realizácia vyhodnotenia typu príspevku

V konceptuálnom návrhu metódy v kapitole 5.1.1 sme navrhli pri vyhodnotení typu príspevku (otázka alebo diskusia) sa spoliehať na informáciu od autora príspevku. Vďaka tomu, že príspevky na diskusnom fóre kurzov v systéme edX takýmto delením disponujú, spoliehame sa v súlade s návrhom na informáciu pochádzajúcu zo vstupných dát.

[1] `https://www.edx.org/`

## 6.2 Realizácia selekcie otázok z archívu

V kapitole 5.1.2 navrhujeme pracovať buď s otázkami, ktoré majú nejakú odpoveď alebo s otázkami, ktoré majú označenú najlepšiu odpoveď a teda je pravdepodobnejšie, že sú naozaj vyriešené. Pri analýze dát zo systému edX sme pozorovali, že len malé množstvo otázok obsahuje odpoveď, ktorá je označená za najlepšiu. Z tohto dôvodu sme sa rozhodli pracovať so všetkými otázkami, ktoré majú aspoň jednu odpoveď.

## 6.3 Realizácia identifikovania podobných otázok

Pre nájdenie a zoradenie otázok navrhujeme použiť binárny klasifikátor (kapitola 5.1.3). Okrem rozhodnutia, či sú otázky podobné, určí klasifikátor aj pravdepodobnosť s ktorou sú si otázky podobné. Túto pravdepodobnosť potrebujeme pre ďalší krok našej metódy.

### 6.3.1 Predspracovanie textu

Predspracovanie textu navrhnuté v kapitole 5.2 zahŕňa lematizáciu slov a odstránenie stop slov. Lematizáciu slov sme realizovali prostredníctvom knižnice *stanford-core-nlp*[2], ktorá nám umožňuje z jazyka Ruby pracovať s nástrojmi pre prácu s textom. Tieto nástroje vedia pracovať okrem angličtiny (ktorú potrebujeme pre našu realizáciu metódy) aj s francúzštinou a nemčinou. Pomocou tejto knižnice sme texty tokenizovali a získali lémy jednotlivých tokenov, ktoré sme následne transformovali na tvar skladajúci sa len z malých písmen. Takto spracovaný text sme si uložili pre ďalšie použitie.

Zoznam stop slov sme získali zo stránky ranks.nl[3]. Tento zoznam sme rozšírili aj o interpunkčné znamienka a formátovacie značky formátu markdown[4], v ktorom sa nachádzajú texty otázok.

### 6.3.2 Výpočet podobnosti textu

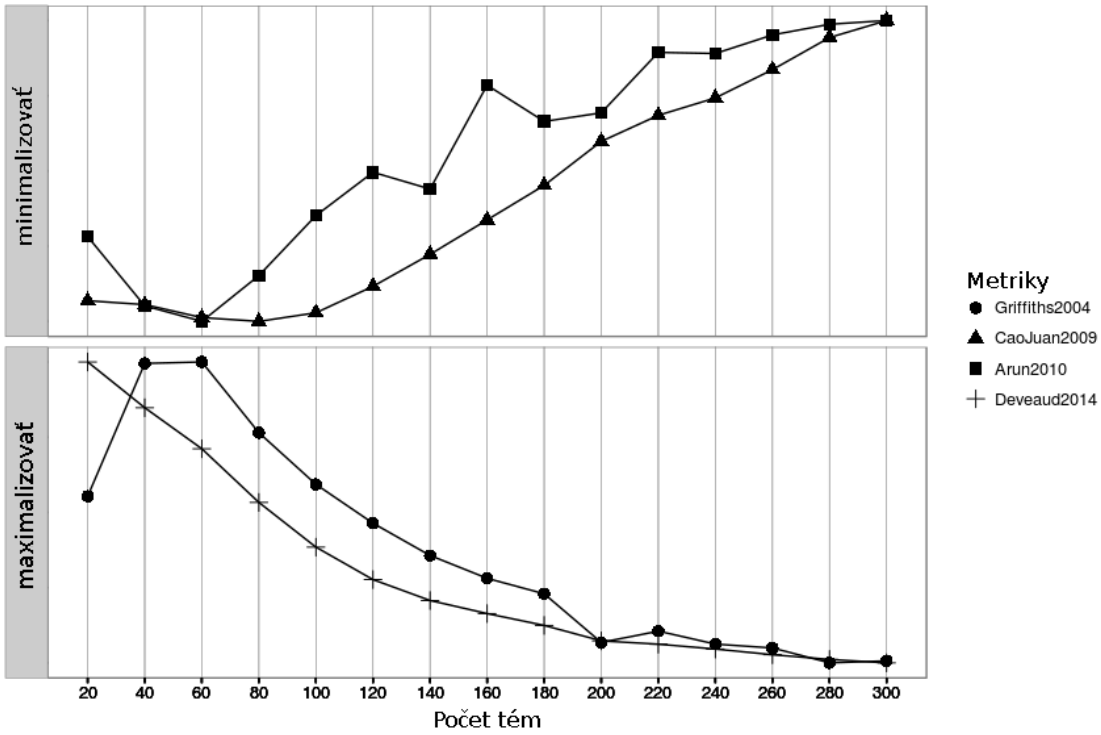
V kapitole 5.1.3 sme prezentovali tri spôsoby výpočtu podobnosti dvoch textov. V prvom kroku sme zrealizovali podobnosť na základe váh TF-IDF. Váhy TF-IDF sme získali pre unigramy (jednotlivé slová) a bigramy (dvojice slov). Podobnosť vektorovej reprezentácie textov sme vypočítali na základe kosínovej podobnosti, pričom hodnoty vektorov boli pri výpočte podobnosti normalizované L2 normalizáciou.

[2] `https://github.com/louismullie/stanford-core-nlp`

[3] `http://www.ranks.nl/stopwords`

[4] `https://tools.ietf.org/html/rfc7763`

Ako ďalšie črty do klasifikátora sme pridali aj podobnosť na základe LDA tém. Vhodný počet LDA tém sme určili pomocou knižnice *ldatuning*[5]. Ukážka výstupu z knižnice je na obr. 6.1. Pri vyhodnotení úspešnosti klasifikátora podobných otázok však došlo k zníženiu úspešnosti a preto sme črty vypočítané na základe LDA tém pri trénovaní klasifikátora nepoužili. Toto pozorovanie je v rozpore s predchádzajúcimi prácami, ktoré dosahovali lepšiu presnosť pri použití LDA tém (Li a Manandhar, 2011; Zhang et al., 2014; Shtok et al., 2012; Ji et al., 2012). Použitie LDA tém nebolo vhodné pravdepodobne preto, lebo obsah kurzov je zameraný na jednu tému, kvôli čomu latentné témy nemajú dostatočnú rozlišovaciu schopnosť.




**Obr. 6.1:** Ukážka výstupu z knižnice *ldatuning* zobrazujúca výkon metrík pre určenie vhodného počtu LDA tém.

Aj z dôvodu zníženia presnosti pri použití LDA tém sme sa rozhodli zrealizovať výpočet podobnosti textov pomocou word2vec modelu. Pri pokuse nájsť knižnicu pre jazyk R, ktorá by vedela pracovať s *word2vec* sme boli neúspešní. Z tohto dôvodu sme sa rozhodli pre alternatívne riešenie - použitie knižnice *text2vec*, ktorú používame aj pre výpočet podobnosti podľa TF-IDF váh. Knižnica umožňuje využiť algoritmus GloVe (Pennington et al., 2014), ktorý rovnako ako word2vec vie vypočítať vektorovú reprezentáciu jednotlivých slov.

Nevýhodou použitia GloVe z knižnice *text2vec* je, že knižnica nepodporuje importovanie naučených vektorov slov a preto sme museli vektory slov natrénovať. Použili sme na to texty otázok, odpovedí a komentárov z kurzu. Alternatívou by bolo využiť text výučbových materiálov. Pri použití GloVe algoritmu sme slová reprezentovali v 150 dimenzionálnom priestore a pri učení sa spoločných výskytov textov sme nastavili parameter maximálnej vzdialenosti slov na 5. Vektory slov v texte sme sčítali a získali tak vektor pre celý text.

[5] https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf

### 6.3.3 Realizácia čŕt pre klasifikátor

Pri rozpracovaní čŕt v kapitole 5.2 sme navrhli konkrétnu reprezentáciu pre jednotlivé typy čŕt, ktoré navrhujeme použiť. Výpočet navrhnutých čŕt je priamočiary s výnimkou *podobnosti ako umiestnenia kategórie v hierarchii kategórií*. Črtu prekryv slov sme rozšírili o črtu prekryv bigramov. Experimentovali sme aj s použitím trigramov (trojice slov), avšak táto črta sa neukázala ako užitočná a preto sme ďalej s trigramami nepracovali. Zoznam všetkých použitých čŕt uvádzame v prílohe D.

**Výpočet podobnosti na základe kategórie**

Výpočet *podobnosti ako umiestnenia kategórie v hierarchii kategórií* sme sa rozhodli založiť na základe hĺbky spoločného rodiča kategórií otázok. Vzorec pre výpočet podobnosti je nasledovný:

$$sim_{cat}(Q_1, Q_2) = \begin{cases} 1 & \text{ak} C_1 = C_2 \\ \dfrac{\text{depth}(C_p)}{max\_depth} & \text{ak} C_1 <> C_2 \end{cases} \tag{6.1}$$

kde $Q_1$ a $Q_2$ sú otázky, $C_1$ a $C_2$ kategórie otázok, depth($C_p$) je hĺbka spoločného rodiča a $max\_depth$ je maximálna hĺbka stromu kategórií.

### 6.3.4 Použitie klasifikátora

V realizácii našej metódy môžeme použiť akýkoľvek binárny klasifikátor. Celkovo sme použili tri rôzne klasifikátory:

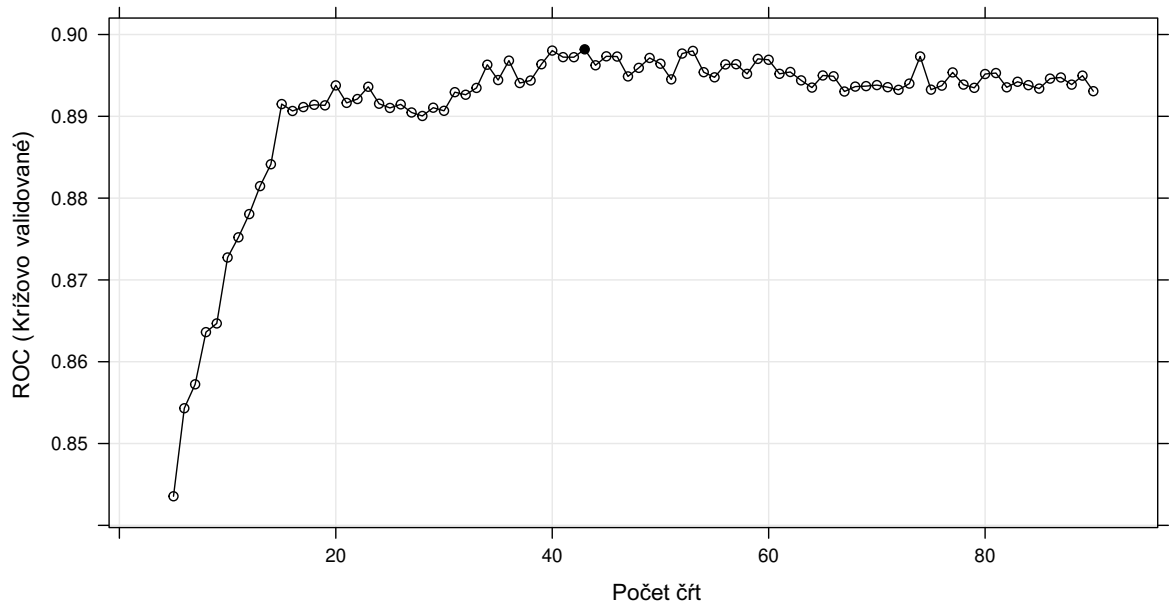
- náhodný les,
  - Klasifikátor náhodný les (angl. random forest) je založený na vytvorení viacerých rozhodovacích stromov, pričom každý zo stromov predpovedá klasifikačnú triedu záznamu. Trieda, ktorá bola predpovedaná najväčším počtom stromov, je následne použitá ako predpoveď klasifikátora.
- SVM,
  - Algoritmus SVM sa snaží nájsť čo najväčšiu medzeru medzi dátami reprezentovanými v n-dimenzionálnom priestore.
- naivný bayesov klasifikátor.
  - Naivný bayesov klasifikátor je založený na pravdepodobnostiach - na Bayesovej teoréme. Z tohto dôvodu je predpokladom, že vstupné črty sú navzájom nezávislé, a preto je potrebné spraviť analýzu korelácií čŕt a odstrániť silno korelované črty.
  - V našom riešení sme odstránili črty, kde korelácia bola väčšia/menšia ako 0,8/-0,8. Zoznam odstránených čŕt uvádzame v prílohe D.

### 6.3.5 Výber najdôležitejších čŕt pre klasifikátor

V kapitole 5.2 sme identifikovali problém veľkého počtu čŕt a navrhujeme ich redukciu výberom najdôležitejších čŕt. V rámci realizovania metódy sme použili dva spôsoby pre výpočet dôležitosti čŕt. Hodnoty čŕt sme normalizovali na základe návrhu v kapitole 5.2.

#### Rekurzívna eliminácia čŕt

Rekurzívnu elimináciu čŕt (angl. recursive feature elimination, RFE) sme realizovali pomocou knižnice *caret*[6]. Algoritmus pozostáva z výberu čŕt, ktorých počet sa rekurzívne znižuje. Externý estimátor (v našom prípade sme použili klasifikátor *náhodný les* s počtom stromov 500 a parametrom mtry = 6) validuje natrénovaný model a priraďuje jednotlivým čŕtám váhu. Trénovanie bolo validované 5-násobnou krížovú validáciou (angl. 5-fold cross-validation). Všetky použité črty aj s identifikovanou dôležitosťou uvádzame v prílohe D. Graf úspešnosti klasifikátora uvádzame na obr. 6.2. ROC krivka (na grafe na obr. 6.2 je zachytená na osi Y) odráža vzťah medzi počtom správne pozitívnych výsledkov (angl. true positive) a nesprávne pozitívnych (angl. false positive) výsledkov.




**Obr. 6.2:** Ukážka výkonu klasifikácie podobnosti otázok pri použití RFE.

#### Modely založené na gradientovom zosilnení

Pomocou knižnice *gbm*[7] sme použili modely založené na gradientovom zosilnení (angl. gradient boosted models, GBM). Okrem možnosti samotnej klasifikácie pomocou týchto modelov vieme získať aj dôležitosť jednotlivých čŕt. Zoznam čŕt s dôležitosťou získanou pomocou GBM uvádzame v prílohe D.

[6] https://topepo.github.io/caret

[7] https://cran.r-project.org/web/packages/gbm/gbm.pdf

Kým pri použití RFE dostane zoznam čŕt, ktoré máme použiť, pri použití GBM dostávame zoznam všetkých čŕt s ich dôležitosťou a počet čŕt si určíme sami. Ako vidieť na obr. 6.2, pomocou RFE prístupu získavame 43 čŕt. Ako črty podľa GBM sme sa rozhodli použiť črty s dôležitosťou minimálne 0,2. Takýchto čŕt bolo 53. Najdôležitejšom črtou podľa oboch prístupov k výberu čŕt je *podobnosť nadpis+text na základe TF-IDF*. Ďalšie dôležité črty podľa GBM sú:

- pomer počtu zhodných podstatných mien v nadpis+text ku počtu unikátnych slov v nadpis+text s väčším počtom unikátnych slov,
- podobnosť nadpisov podľa GloVe,
- pomer počtu zhodných podstatných mien v texte ku počtu unikátnych slov v nadpise s väčším počtom unikátnych slov.

Podľa RFE sú ďalšie dôležité črty:

- pomer počtu zhodných podstatných mien v nadpis+text ku počtu unikátnych slov v nadpis+text s menším počtom unikátnych slov,
- pomer počtu zhodných podstatných mien v nadpis+text ku počtu unikátnych slov v nadpis+text s väčším počtom unikátnych slov,
- počet zhodných slov v nadpis+text.

## 6.4 Realizácia orezania zoznamu podobných otázok

V kroku metódy *realizácia filtrovania podobných otázok* (kapitola 5.1.4) ponecháme v zozname podobných otázok len tie, ktoré spĺňajú určitú minimálnu podobnosť $\gamma$. Počas vyhodnocovania našej metódy pomocou offline experimentu sme hodnotu $\gamma$ stanovili na základe ladenia úspešnosti klasifikácie podobných otázok. Hodnota pre ktorú klasifikátor dosahoval najlepšiu úspešnosť bola približne 0,9 (hodnota sa mierne líšila v závislosti od klasifikátora a použitých čŕt).

## 6.5 Realizácia zoradenia odpovedí

Zoradenie odpovedí realizujeme na základe návrhu v kapitole 5.1.5 párovým prístupom techniky *učenie sa zoraďovať*. Pomocou tejto techniky dostaneme relatívne zoradenie odpovedí identifikovaných podobných otázok.

### 6.5.1 Predspracovanie textu

Predspracovanie textu odpovedí realizujeme analogicky ako predspracovávame texty otázok (kapitola 6.3.1). Konkrétne ide o lematizáciu a odstránenie stop slov.

### 6.5.2 Výpočet podobnosti textu

Výpočet podobnosti textu sme realizovali len na základe kosínovej podobnosti vektorovej reprezentácie textu pomocou TF-IDF váh. Hodnoty vektorov boli pri výpočte podobnosti normalizované L2 normalizáciou.

### 6.5.3 Realizácia čŕt pre zoradenie odpovedí

Realizovali sme všetky črty, tak ako boli navrhnuté v kapitole 5.2. Jednou z čŕt, ktoré používame je informácia o tom, či je autor odpovede učiteľom. Používatelia v systéme edX môžu patriť do viacerých rolí - učiteľ, asistent učiteľa, asistent učiteľa z komunity. V rámci našej realizácii metódy považujeme všetky tieto typy používateľov za rolu učiteľa. Zoznam realizovaný čŕt uvádzame v prílohe E.

### 6.5.4 Použitie algoritmu techniky učenie sa zoraďovať

V kapitole 5.1.5 sme identifikovali, že najvhodnejším typom techniky *učenie sa zoraďovať* je párový prístup. V literatúre bolo prezentovaných množstvo algoritmov, ktoré sa zaraďujú do tejto kategórie, avšak len malé množstvo z nich má verejne prístupnú implementáciu. Pre realizáciu našej metódy sme zároveň vyžadovali implementácie algoritmov spustiteľné z jazykov Ruby alebo R.

Počas analýzy existujúcich knižníc sme identifikovali dve možnosti, ktoré spĺňajú našu podmienku. Prvou knižnicou je súbor algoritmov *sofia-ml*[8] a jej port do jazyka R - *RSofia*[9]. Problém s touto knižnicou je, že dokumentácia ku knižnici *RSofia* už nie je dostupná.

Druhou identifikovanou možnosťou je knižnica $SVM^{rank}$ (Joachims, 2006). Táto knižnica nemá port do jazyka R ani Ruby, avšak spúšťa sa prostredníctvom príkazového riadku rovnako ako knižnica *sofia-ml*. Knižnicu $SVM^{rank}$ použili aj autori v (Belinkov et al., 2015).

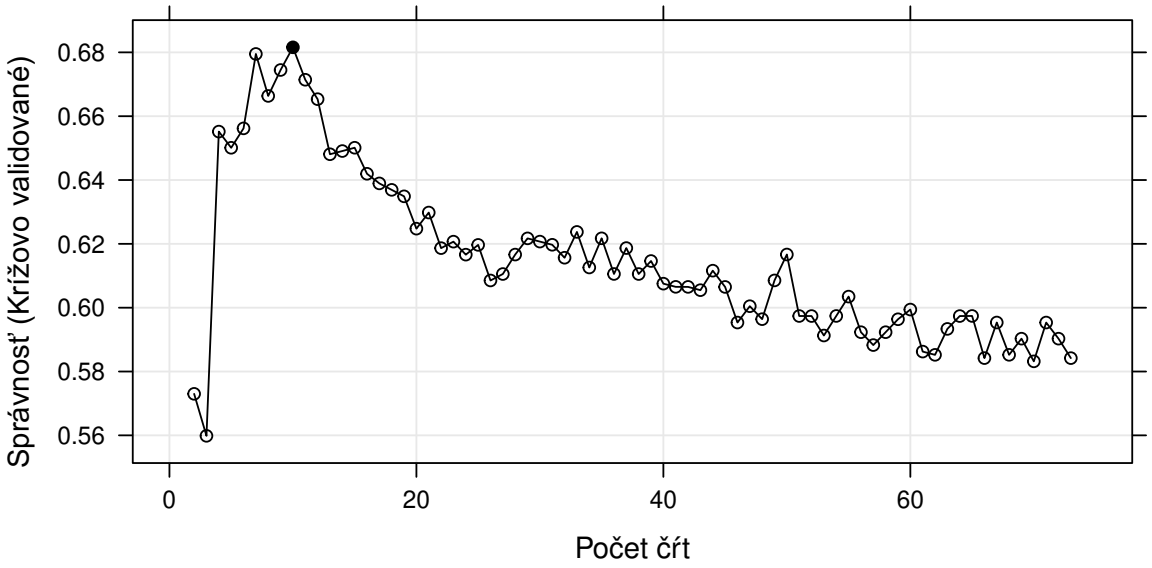
### 6.5.5 Výber najdôležitejších čŕt pre zoradenie odpovedí

Aj v prípade zoradenia odpovedí pracujeme s veľkým počtom čŕt - vytvorili sme ich 130. Zoznam čŕt uvádzame v prílohe E. Hodnoty čŕt sme normalizovali a identifikovali najdôležitejšie pomocou RFE a GBM. Identifikované črty sú troch typov 1) črty týkajúce sa odpovede, 2) podobnosť odpovede s novou otázkou, 3) podobnosť odpovede s pôvodnou otázkou. Počas vyhodnocovania sme zistili, že použitie čŕt typu podobnosť odpovede s pôvodnou otázkou spôsobovalo zníženie úspešnosti a preto sme všetky tieto črty prestali používať.

[8] `https://code.google.com/archive/p/sofia-ml/`

[9] `https://github.com/fcela/RSofia`

Najdôležitejšie črty identifikované pomocou GBM sú: *rozdiel času vytvorenia odpovede a otázky*; *dĺžka odpovede*; *informácia, či je autor odpovede učiteľ*; *pomer najlepších odpovedí ku všetkým odpovediam používateľa*; *počet unikátnych zhodných slov*; *pomer dĺžky nadpisu otázky a textu odpovede*; *pomer dĺžky nadpisu otázky bez stop slov a textu odpovede*; *skóre vo forme hlasovania*; *podobnosť textu na základe TFIDF reprezentácie*; *počet stop slov*. Môžeme pozorovať, že edukačná črta *informácia, či je autor odpovede učiteľ* je tretia najdôležitejšia. Črty identifikované pomocou RFE sú podobné a ich zoznam uvádzame v prílohe E. Graf úspešnosti klasifikátora ako výstup z RFE uvádzame na obr. 6.3.




**Obr. 6.3:** Ukážka výkonu klasifikácie vhodnosti odpovedí pri použití RFE.

## 6.6 Realizácia vrátenia zoznamu odpovedí

Krok navrhnutej metódy *vrátenie zoznamu odpovedí* sme v rámci nášho riešenia realizovali univerzálne bez snahy o konkrétne použitie, ktoré sme opísali v kapitole 5.1.6. Na výstupe našej metóde sme vrátili zoznam všetkých odpovedí zoradený podľa ich vhodnosti. Vďaka takémuto prístupu sme vedeli vyhodnotiť tento krok metódy viacerými metrikami.

# Kapitola 7

# Overenie navrhnutého riešenia

Overenie navrhnutého riešenia sme realizovali pomocou syntetického (offline) experimentu na dátovej sade z MOOC kurzov. V rámci overenia návrhu metódy prostredníctvom offline experimentu sme si stanovili tieto ciele:

- Experimentálne nájsť najvhodnejší klasifikačný model pre identifikáciu podobných otázok.
- Nájsť hodnotu parametra $\gamma$, ktorý určuje hranicu podobnosti otázok, pre ktoré zoraďujeme odpovede.
- Porovnať úspešnosti navrhnutej metódy s metódou, ktorá neuvažuje špecifiká online komunít študentov.

Overenie našej metódy sa skladá z dvoch samostatných overení pre dva hlavné kroky našej metódy 1) identifikovanie podobných otázok (kapitola 5.1.3) a 2) zoradenie odpovedí (kapitola 5.1.5). Tieto dva kroky sme overovali nezávisle jeden od druhého, aby sme videli v ktorom kroku vzniká aká veľká chyba a mohli ladiť parametre algoritmov strojového učenia v izolovanom prostredí. Na záver overujeme spojenie (angl. ensemble) algoritmov strojového učenia v celkovom overení navrhnutej metódy.

Pri identifikovaní podobných otázok sa pozeráme len na chronologicky staršie otázky - teda pre otázku na začiatku kurzu (*otázka 1*) nemôže byť ako podobná identifikovaná otázka z konca kurzu (*otázka 2*), pretože v čase, kedy by prebiehalo odporúčanie *otázky 1* by *otázka 2* ešte neexistovala. Vďaka chronologickému obmedzeniu identifikovania podobných otázok platí aj pri zoraďovaní odpovedí, že sa zoraďujú odpovede len zo starších otázok.

V prípade zoraďovania odpovedí sme však obmedzenie na chronologickosť zjemnili a zoraďujeme aj odpovede z *otázky 2*, ktoré boli vytvorené až po vytvorení *otázky 1*, keďže takýchto prípadov bol zanedbateľný počet.

## 7.1 Dátová sada

V kapitole 2.3.8 sme pozorovali, že takmer všetci autori využívali príspevky z CQA systému Yahoo! Answers, ktoré bolo možné získať cez rozhranie pre programovanie aplikácií (angl. application programming interface, API). V doméne online komunít žiadna verejne dostupná dátová sada neexistuje. Dátové sady týkajúce sa MOOC kurzov ponúkajú na základe žiadosti Stanfordova univerzita[1] a Massachusettský technologický inštitút[2] (angl. Massachussetts Institute of Technology, MIT). Dátové sady kurzov zo Stanfordovej univerzity obsahujú napríklad obsah diskusných fór, aktivitu používateľov, úspešnosť študentov v testoch a iné údaje. Počas predmetu diplomový predmet I sme získali dáta z 5 kurzov pochádzajúcich z tejto dátovej sady. Po bližšej analýze dát sme však zistili, že dáta neobsahujú kategórie otázok na fóre, otázky nemajú informáciu o svojej kategórii a chýba informácia, či ide o *otázku* alebo *diskusiu*. Z tohto dôvodu sme sa rozhodli s týmito dátami ďalej nepracovať.

Dátovú sadu, s ktorou môžeme realizovať offline experiment, sme sa rozhodli vytvoriť stiahnutím obsahu MOOC kurzov na platforme edX. V jazyku Ruby sme vytvorili skript, ktorý prechádza všetky témy na fórach zvolených kurzov a ukladá ich obsah do databázy. Jednotlivé objekty sme sa rozhodli ukladať do databázového modelu systému Askalot, vďaka čomu sme schopní ich zobrazovať v systéme Askalot a neskôr, po prispôsobení našej implementácie metódy, aj priamo v Askalote samotnú metódu automatického odpovedania na otázky používať. Okrem obsahu diskusií sme si uložili aj výučbový obsah kurzov, ktorý by mohol byť použitý v rámci vylepšovania úspešnosti nami navrhnutej metódy. Tento výučbový obsah sme však napokon nevyužili.

### Opis dátovej sady

Pri získavaní dátovej sady zo systému edX sme sa zamerali na kurzy týkajúce sa informačných technológií. Jednotlivé kurzy sme sťahovali priebežne počas predmetov diplomový projekt I a diplomový projekt II. Prehľad získaných kurzov uvádzame v tabuľke 7.1. Celkovo sme získali dáta pre 18 kurzov (pre 5 kurzov boli dostupné dve iterácie).

Väčšina kurzov v systéme edX nie je opakovaná pravidelne a preto je náročné odhadnúť, ktorý kurz sa či a kedy bude opakovať. Z tohto dôvodu sme sa snažili získať dáta pre čo najväčší počet kurzov. V tabuľke môžeme vidieť, že len 5 z 13 kurzov malo naplánovanú/realizovanú ďalšiu iteráciu (k dátumu 28.11.2016).

Jednotlivé kurzy majú rôznu dĺžku trvania a môžu byť dvoch typov. Prvý typ sú kurzy realizované ako tradičný kurz, kde nový materiál je prezentovaný každý týždeň. Druhý typ je kurz, v ktorom je všetok obsah dostupný hneď na začiatku. Pre ďalšiu prácu na offline experimente sme potrebovali zistiť počet príspevkov a používateľov v jednotlivých kurzoch. Tento štatistický prehľad uvádzame v tabuľke 7.2. Jednotlivé stĺpce tabuľky 7.2 predstavujú počet otázok v kurze, počet odpovedí na otázky, počet používateľov, ktorí sa zapojili do fóra,

[1] `https://datastage.stanford.edu/`

[2] `http://web.mit.edu/ir/mitx/index.html`

**Tabuľka 7.1:** Začiatky kurzov z vytvorenej dátovej sady a dátum začiatku ďalšej iterácie kurzu.

| Kurz | Začiatok kurzu | Ďalšia iterácia |
|---|---|---|
| BerkeleyX: CS188x_1 Artificial Intelligence | 25.08.2014 | - |
| MITx: 21W.789x Building Mobile Experiences | 25.01.2015 | - |
| Microsoft: DEV204x Programming with C# | 28.05.2015 | 11.03.2016 |
| Microsoft: DEV201x Introduction to TypeScript | 02.06.2015 | 16.05.2016 |
| TeachersCollegeX: BDE1x Big Data in Education | 01.07.2015 | - |
| DelftX: FP101x Introduction to Functional Programming | 15.10.2015 | - |
| MITx: 21W.789.2x Mobile Application Experiences Part 2: Mobile App Design | 29.02.2016 | - |
| MITx: 6.00.2x Introduction to Computational Thinking and Data Science | 02.03.2016 | 19.10.2016 |
| MITx: 21W.789.3x Mobile Application Experiences Part 3: Building Mobile Apps | 27.03.2016 | - |
| HKUSTx: COMP107x Introduction to Mobile Application Development using Android | 31.03.2016 | 14.11.2016 |
| Microsoft: DAT208x Introduction to Python for Data Science | 08.05.2016 | 13.06.2016 |
| IITBombayX: CS213.1x Foundation of Data Structures | 17.05.2016 | - |
| BerkeleyX: CS169.1x Agile Development Using Ruby on Rails - The Basics | 01.06.2016 | - |

počet kategórií s otázkami, dĺžku trvania kurzu v týždňoch a typ kurzu. Kurzy s viacerými iteráciami sú zvýraznené hrubým písmom.

### Výber kurzu pre zrealizovanie offline experimentu

V tabuľke 7.2 môžeme pozorovať veľké rozdiely medzi popularitou jednotlivých kurzov. Pre naše potreby sú navyše vhodnejšie kurzy, ktoré majú dostupných viacero iterácií. Jednou z vlastností, ktorú chceme využiť pri identifikácii podobných otázok (kapitola 5.1.3) je časová podobnosť a preto ďalším faktorom preferencie sú pre nás kurzy, ktoré majú obsah publikovaný po týždňoch. Použitie otázok z viacerých iterácií kurzu zároveň predstavuje väčšiu šancu podobných otázok a umožňuje nám overiť dôležitosť využitia otázok z archívu predchádzajúcej iterácie kurzu. Naše požiadavky na štruktúru kurzu spĺňa kurz *MITx: 6.00.2x*, ktorý má navyše najväčší počet otázok, vďaka čomu máme dostatočne veľkú dátovú sadu na zrealizovanie offline experimentu.

Počas predmetu diplomový projekt II sme zrealizovali offline experiment na prvej iterácii kurzu *MITx: 6.00.2x*. Po začatí prác s dátami z druhej iterácie kurzu sme zistili, že obsah kurzu bol aktualizovaný. Obsah jednotlivých týždňov bol preusporiadaný, niektoré časti kurzu zrušené a iné pridané. Ďalšou zmenou bol zmenený harmonogram publikovania nového obsahu. Pôvodne bol obsah zverejňovaný tradične po týždňoch a po novom začali publikovať nový obsah každé dva týždne. Napriek zmenám v štruktúre kurzu sme sa rozhodli pokračovať v realizovaní offline experimentu na tomto kurze.

**Tabuľka 7.2:** Začiatky kurzov z vytvorenej dátovej sady a dátum začiatku ďalšej iterácie kurzu. Označenie v stĺpci typ: Týž. - kurz má obsah zverejňovaný po týždňoch, Vš. - kurz má všetok obsah dostupný od spustenia kurzu.

| Kurz | Otázky | Odpovede | Používatelia | Kategórie | Dĺžka | Typ |
|---|---|---|---|---|---|---|
| BerkeleyX: CS188x_1 | 108 | 159 | 162 | 1 | 14 | Týž. |
| MITx: 21W.789x | 894 | 764 | 716 | 16 | 12 | Týž. |
| **Microsoft: DEV204x** | 237 | 144 | 250 | 13 | 12 | Vš. |
| **Microsoft: DEV204x - 2. iterácia** | 266 | 395 | 313 | 15 | 12 | Vš. |
| **Microsoft: DEV201x** | 152 | 74 | 143 | 6 | 6 | Vš. |
| **Microsoft: DEV201x - 2. iterácia** | 98 | 47 | 83 | 6 | 6 | Vš. |
| TeachersCollegeX: BDE1x | 684 | 892 | 521 | 74 | 8 | Týž. |
| DelftX: FP101x | 954 | 1474 | 774 | 41 | 8 | Týž. |
| MITx: 21W.789.2x | 95 | 128 | 81 | 5 | 5 | Týž. |
| **MITx: 6.00.2x** | 1228 | 1835 | 765 | 180 | 8 | Týž. |
| **MITx: 6.00.2x - 2. iterácia** | 1027 | 1764 | 693 | 178 | 5 (10) | Týž. |
| MITx: 21W.789.3x | 77 | 113 | 54 | 6 | 5 | Týž. |
| **HKUSTx: COMP107x** | 68 | 111 | 87 | 1 | 5 | Vš. |
| **Microsoft: DAT208x** | 235 | 299 | 342 | 49 | 6 | Vš. |
| **Microsoft: DAT208x - 2. iterácia** | 244 | 274 | 318 | 49 | 6 | Vš. |
| IITBombayX: CS213.1x | 158 | 181 | 115 | 60 | 7 | Týž. |
| BerkeleyX: CS169.1x | 288 | 517 | 221 | 1 | 6 | Týž. |

## 7.2 Tvorba zlatého štandardu

Okrem samotnej dátovej sady potrebujeme mať pre offline experiment k dispozícii aj zlatý štandard, aby sme vedeli vyhodnotiť úspešnosť metódy. Autori v predchádzajúcich prácach využívali manuálne aposteriórne vyhodnotenie relevantnosti nájdených podobných otázok, pretože neexistuje zlatý štandard, ktorý by mohol byť použitý. Rovnaký spôsob použili pri vyhodnocovaní zoradenia odpovedí.

Na rozdiel od autorov v predchádzajúcich prácach, ktorí vyhodnocovali úspešnosť navrhnutých metód manuálnou anotáciou a posteriori, my sme sa rozhodli vytvoriť si zlatý štandard a priori, teda ešte pred začiatkom samotného experimentovania. Zlatý štandard sme vytvorili manuálnym anotovaním podobnosti párov otázok a vhodnosti odpovedí. Pri analýze dátovej sady sme pozorovali, že množstvo príspevkov na fóre malo nesprávne zvolený typ príspevku - otázka alebo diskusia. Z tohto dôvodu sme manuálne opravili typ príspevkov v kurze *MITx: 6.00.2x*. Celkovo sme vykonali tri typy manuálneho anotovania:

1. oprava typu príspevku otázka/diskusia,
2. vyhodnotenie podobnosti otázok,
3. vyhodnotenie vhodnosti odpovedí.

### Manuálne anotovanie dát z prvej iterácie kurzu

Oprava typu príspevku otázka/diskusia bola vykonaná pre všetky príspevky z prvej iterácie kurzu jedným anotátorom na základe textu otázky. Celkovo sme z 1228 príspevkov identifikovali 915 otázok, z ktorých 751 otázok malo aspoň jednu odpoveď. Z dôvodu, že existuje 281 625 možných párov otázok, sme pri anotovaní podobnosti otázok postupovali nasledovne:

- Vypočítali sme podobnosť textov otázok (nadpis otázky + text otázky) pomocou kosínovej podobnosti TF-IDF reprezentácii textov otázok.
- Pre každú otázku našli 20 najpodobnejších a zo zoznamu podobných otázok sme odstránili tie otázky, kde bola podobnosť menšia ako 0,1. Zvyšné otázky sme posunuli ďalej na manuálne vyhodnotenie podobnosti.
- Pri vyhodnocovaní podobnosti sme zobrazili kategórie otázok, nadpisy a texty otázok.
- Podobnosť sme anotovali binárne: sú podobné/nie sú podobné.

Podľa definície podobných otázok v kapitole 5.1.3 hľadáme otázky, ktoré nemusia byť priamo duplicitné, stačí aby išlo o dostatočne podobné otázky, pri ktorých je predpoklad, že odpovede jednej otázky by mohli byť užitočné aj pre druhú otázku. Anotovanie párov podobných otázok prebehlo jedným anotátorom a vyhodnotených bolo 14 826 párov otázok. Počas anotovania párov otázok z prvej iterácie kurzu sme ešte neuvažovali chronologickosť a preto niektoré páry mohli byť anotované dva krát. Z tohto dôvodu sme skontrolovali ešte konzistentnosť rozhodnutí a nesúhlasiace anotácie opravili (bolo ich 366). Celkovo sme identifikovali 441 párov podobných otázok, ktoré predstavujú 175 otázok (z 751), pre ktoré sme označili podobnú otázku. Na základe toho, že sme identifikovali len 441 podobných párov z 14 826 vidíme, že ide o veľmi nevyváženú dátovú sadu.

Pri anotovaní vhodnosti odpovedí sme postupovali nasledovne:

- Anotovali sme vhodnosť odpovedí pre páry otázok, ktoré sme označili ako podobné pri anotovaní podobnosti otázok.
- Zobrazili sme kategórie otázok, a texty otázok a odpovedí.
- Vhodnosť sme anotovali do 3 úrovní: vhodná, čiastočne vhodná a nevhodná.

Aby bola odpoveď zaradená do prvej úrovne musí byť relevantná a kvalitná zároveň. Do druhej úrovne môžu patriť napríklad odpovede, ktoré obsahujú len odkaz na nejakú príbuznú diskusiu, alebo zodpovedajú len časť problému otázky. Anotovanie vhodnosti odpovedí bolo robené autorom tejto publikácie. Vyhodnotili sme 1637 párov otázka-odpoveď. Počas anotovania sme identifikovali páry otázok, ktoré boli nesprávne označené ako podobné. Nesprávne anotácie podobných otázok sme opravili.

### Manuálne anotovanie dát z druhej iterácie kurzu

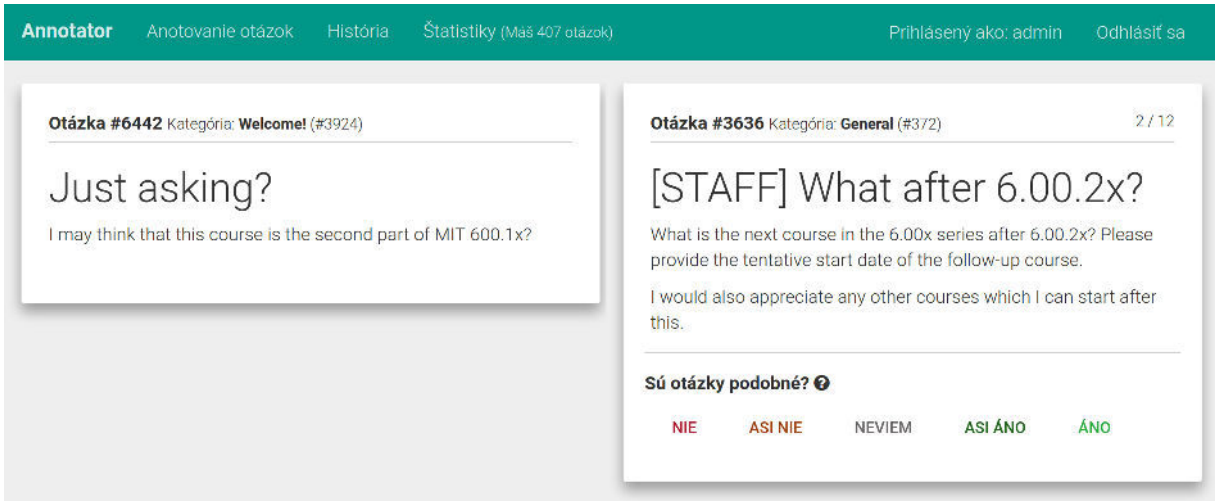
Opravu typu príspevku otázka/odpoveď sme realizovali rovnakým spôsobom ako pri anotovaní dát prvej iterácie. Z 1027 príspevkov sme 697 označili ako otázky a 657 z nich malo aspoň jednu odpoveď.

**Tabuľka 7.3:** Počet podobných otázok na jednotlivých pozícia pri anotovaní dát prvej iterácie.

| **Pozícia** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Počet podobných otázok** | 63 | 44 | 30 | 19 | 21 | 20 | 14 | 18 | 16 | 12 |
| **Pozícia** | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| **Počet podobných otázok** | 21 | 13 | 4 | 4 | 4 | 5 | 9 | 4 | 4 | 2 |

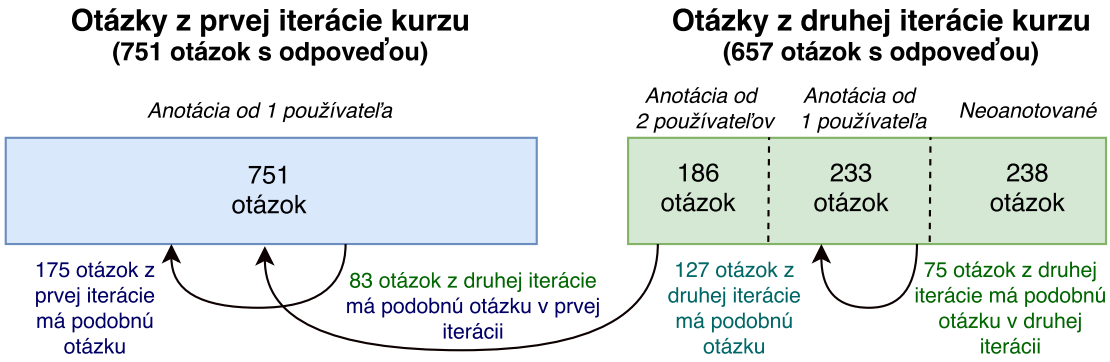
Pre anotovanie podobnosti otázok sme vytvorili nástroj na anotovanie s cieľom využiť viacerých expertov pre vytvorenie zlatého štandardu podobných otázok. Pre každú otázku sme vybrali 12 najpodobnejších otázok na základe kosínovej podobnosti TF-IDF reprezentácie textov otázok. Podobné otázky boli hľadané v rámci prvej aj druhej iterácie kurzu. Číslo 12 sme zvolili na základe analýzy počtu podobných otázok v jednotlivých pozíciách na dátach z anotovania prvej iterácie kurzu (tabuľka 7.3).

V online anotátore sme umožnili určiť podobnosť otázok do 5 úrovní: 1) sú podobné, 2) asi sú podobné, 3) neviem, 4) asi nie sú podobné, a 5) nie sú podobné. Ukážka používateľského rozhrania nástroja je na obr. 7.1. Prostredníctvom anotátora oanotovalo otázky celkovo 24 používateľov vrátene autora tejto publikácie. Pre 186 otázok sme získali anotácie od dvoch používateľov a pre ďalších 234 od jedného používateľa. Zvyšných 237 otázok ostalo neoanotovaných. Pre anotácie od dvoch používateľov sme vypočítali zhodu anotátorov pomocou Cohenovej kappy (angl. Cohen's kappa). Výsledná zhoda anotátorov bola 0,474; čo znamená priemerne dobrá zhoda.



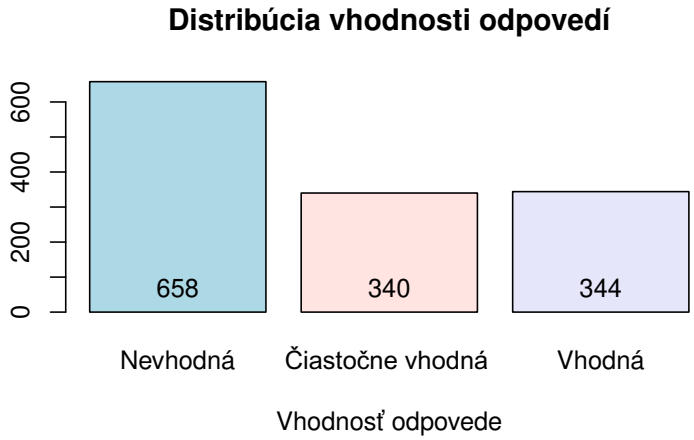

**Obr. 7.1:** Používateľské rozhranie nástroja na anotovanie podobnosti otázok.

Celkovo bolo oanotovaných 5052 párov otázok. Po skonzistentnení rozhodnutí anotátorov sme získali 273 párov podobných otázok, ktoré predstavujú 127 otázok z druhej iterácie kurzu pre ktoré bola nájdená podobná otázka. Z 127 podobných otázok bolo len 75 otázok, ktoré mali podobnú otázku v rámci druhej iterácie kurzu a 83 otázok malo podobnú otázku vzhľadom na otázky z prvej iterácie kurzu. Počet identifikovaných otázok v dátovej sade vizualizujeme na obr. 7.2.




**Obr. 7.2:** Vizualizácia dátovej sady dvoch iterácií kurzu *MITx: 6.00.2x* s počtom manuálne identifikovaných podobných otázok.

Anotovanie odpovedí bolo zrealizované jedným anotátorom rovnakým procesom ako pri dátach z prvej iterácie kurzu. Celkovo sme vyhodnotili 549 párov otázka-odpoveď. Vizualizácia počtu odpovedí s jednotlivou vhodnosťou je zobrazená na obr. 7.3. Čísla zobrazené na obr. 7.3 predstavujú súčet anotácií v prvej a druhej iterácii kurzu.




**Obr. 7.3:** Počet odpovedí s jednotlivou vhodnosťou spoločne pre dáta z prvej a druhej iterácie kurzu *MITx: 6.00.2x*.

## 7.3 Overenie identifikovania podobných otázok

V tejto kapitole opisujeme vyhodnotenie kroku *identifikovanie podobných otázok* (kapitola 5.1.3) navrhnutej metódy realizovaný pre dáta z dvoch iterácií kurzu *MITx: 6.00.2x Introduction to Computational Thinking and Data Science*. Overenie sme sa rozhodli realizovať len nad oanotovanými pármi otázok, pretože len tak vieme o nájdených pároch s istotou prehlásiť, či sú podobné alebo nie. Alternatívou by bolo priebežne vyhodnocovať podobnosť novoidentifikovaných párov. Pretože nemáme oanotovanú celú dátovú sadu, nevieme objektívne posúdiť percento otázok s existujúci podobnými otázkami, ktoré naša metóda dokázala úspešne identifikovať. Z tohto dôvodu úplnosť (angl. recall) našej metódy považujeme za sekundárnu a zameriavame sa na čo najväčšiu presnosť pri identifikovaní podobných otázok.

### 7.3.1 Metodológia overenia identifikovania podobných otázok

Identifikovanie podobných otázok overujeme pre tri klasifikátory, ktoré sme opísali v kapitole 6.3.4 - náhodný les, SVM a naivný bayesov klasifikátor. Črty špecifické pre doménu vzdelávania označujeme ako *edukačné črty*. Pri všetkých modeloch bola použitá 5-násobná krížová validácia (angl. 5-fold cross-validation). Úspešnosť vyhodnocujeme pre viacero variantov:

1. variant s črtami identifikovanými pomocou RFE alebo GBM (kapitola 6.3.5),
2. variant s RFE/GBM črtami bez korelovaných čŕt,
3. variant s RFE/GBM črtami bez edukačných čŕt,
4. variant s RFE/GBM črtami bez korelovaných a edukačných čŕt,

Pri vyhodnocovaní sme nastavovali parametre jednotlivých klasifikačných modelov zvlášť pre každý z variantov. Ďalším parametrom, ktorý sme nastavovali je minimálna pravdepodobnosť $\gamma$, ktorú klasifikátor priradí pri predikcii podobnosti otázky a ktorá musí byť splnená, aby sme daný pár otázok považovali za podobný.

**Selekcia dát z dátovej sady**

V dátovej sade máme 1 172 manuálne oanotovaných otázok (celkovo tvoriacich 14 000 oanotovaných párov). Z tohto počtu 302 otázok má aspoň jednu podobnú predchádzajúcu otázku (celkovo tvoriacich 695 podobných párov). Z toho vyplýva, že naša dátová sada je nevyvážená a ak by sme ju rozdelili priamo na trénovaciu a testovaciu množinu, klasifikátor by pravdepodobne podobné páry vo výslednom natrénovanom modely ignoroval a všetky páry klasifikoval ako nie podobné. Problém nevyváženej dátovej množiny je možné riešiť dvoma spôsobmi: 1) znásobiť výskyty záznamov z menšej množiny (angl. oversampling) alebo 2) zmenšiť početnosť záznamov z väčšej skupiny (angl. undersampling). Z dôvodu veľkého nepomeru medzi podobnými a nie podobnými pármi otázok sme sa rozhodli pre zmenšenie poctu záznamov z väčšej skupiny.

Dátovú sadu delíme na trénovaciu časť a testovaciu časť nasledovne:

1. Z počtu 302 otázok s podobnou predchádzajúcou otázkou vyberieme náhodne 75% do trénovacej množiny (t.j. 226) a 25% do testovacej množiny (76 otázok).
2. Z celkového počtu 870 otázok bez podobnej predchádzajúcej otázky náhodne vyberieme 149 otázok do trénovacej množiny (tým zabezpečíme, že finálne páry otázok sa budú vzťahovať na menší počet otázok). Rovnako z týchto 870 otázok vyberieme 25% otázok do testovacej množiny (218 otázok, vďaka čomu pomer otázok s/bez podobnej prechádzajúcej otázky v testovacej množine zodpovedá reálnemu nevyváženému pomeru).
3. V oboch množinách pre každú otázku identifikujeme všetky relevantné páry s predchádzajúcimi otázkami. V trénovacej množine tým získame opäť nevyvážený pomer medzi podobnými/nepodobnými pármi (512 vs. 4459 párov). Opätovne preto znížime počet nepodobných párov náhodným výberom 512 z nich.

Motiváciou pre takéto dvojkrokové rozdelenie dátovej sady (po otázkach a následne po pároch otázok) bolo, aby sa v trénovacej množine vyskytovalo viacej párov (podobných aj nepodobných) pre menší počet otázok. Predpokladáme, že to povedie k lepšej rozlišovacej schopnosti modelu, ako keby mal pre každú otázku menší počet párov, ktorý by vznikol ich náhodným výberom zo všetkých otázok.

**Vyhodnocovacie metriky**

Úspešnosť klasifikácie podobných párov otázok sme sa rozhodli vyhodnocovať pomocou metrík *správnosť*, *presnosť* a *úplnosť*. Tieto metriky sme opísali v kapitole 2.3.8. Úspešnosť klasifikovania párov otázok považujeme za sekundárne overenie, pretože nepredstavuje užitočný pohľad o použití v skutočnom systéme. Z tohto dôvodu uvádzame úspešnosť klasifikátor pri klasifikovaní párov otázok v prílohe C.

Viac ako klasifikovanie párov otázok nás zaujíma presnosť metódy vo forme presnosti odporúčania otázky, ktorá bola odporúčaná ako podobná ak k takémuto odporúčaniu došlo. Vzťah medzi počtom správne odporúčaných a nesprávne odporúčaných, prípadne neodporúčaní v prípade, keď mala byť nejaká otázka odporúčaná zachytávame prostredníctvom metriky *úspešnosť@n*. V našom prípade vyhodnocujeme len *úspešnosť@1*, čo znamená, že sa pozeráme len na najpodobnejšiu odporúčanú otázku. Rozhodli sme sa tak preto, lebo mnoho otázok má len jednu podobnú otázku.

### 7.3.2 Naivný bayesov klasifikátor

Naivný bayesov klasifikátor (NB) predpokladá nezávislosť jednotlivých vstupných čŕt a preto je potrebné, aby v dátovej sade neexistovali korelované črty. Odstránenie korelovaných čŕt sme opísali v kapitole 6.3.4. Pri vyhodnocovaní naivného bayesovho klasifikátora sme ponechali nastavenia parametrov na predvolených hodnotách.

Väčšiu úspešnosť dosiahol naivný bayesov klasifikátor pri použití RFE čŕt a odstránení korelovaných čŕt. Na základe našej hypotézy v tabuľke 7.4 porovnávame úspešnosť odporúčania najpodobnejšej otázky s variantom, kde sú odstránené edukačné črty. Ako môžeme vidieť v tabuľke, naivný bayesov klasifikátor s veľkou istotou predikoval podobnú otázku pre otázky, ktoré nemajú žiadnu podobnú otázku (stĺpec tabuľky *Nemalo odporúčať*). Na základe tohto pozorovania môžeme prehlásiť použitie naivného bayesovho klasifikátora v rámci našej metódy za nevhodné.

**Tabuľka 7.4:** Úspešnosť naivného bayesovho klasifikátora vzhľadom na najpodobnejšiu predikovanú otázku. Pravdepodobnosť $\gamma = 0{,}91$. V testovacej množine je 76 otázok, ktoré majú podobnú otázku a 218 otázok, ktoré nemajú podobnú otázku. Otázku, ktorej hľadáme podobnú označujeme ako *Otazka*. V prvom stĺpci je počet otázok, ktoré boli podobné s otázkou *Otazka*. V druhom počet otázok, ktoré neboli podobné, avšak pre otázku *Otazka* existujú podobné otázky. V treťom stĺpci je počet otázok, ktoré boli predikované pre otázky, ktoré nemajú žiadnu podobnú otázku. V štvrtom stĺpci je zachytený vzťah medzi počtom správnych a nesprávnych predikcií. V piatom stĺpci je presnosť ako počet podobných predikovaných otázok ku počtu všetkých predikovaných otázok pre otázky, ktoré majú podobnú otázku. V šiestom stĺpci je zachytené množstvo otázok, ktoré majú podobnú otázku, pre ktoré bola predikovaná podobná otázka.

| | Dobré odp. | Zlé odp. | Nemalo odporúčať | Success@1 | Presnosť pre pod. otázky | Úplnosť pre pod. otázky |
|---|---|---|---|---|---|---|
| NB bez korelovaných | 13 | 3 | 67 | 0,5544 | 0,8125 | 0,2105 |
| NB bez korel. a edu. | 11 | 3 | 71 | 0,5374 | 0,7857 | 0,1842 |

### 7.3.3 Náhodný les

Implementácia klasifikátora *náhodný les* (NL) (angl. random forest) z knižnice *randomForest*[3] nám okrem iného umožňuje nastaviť:

- počet rozhodovacích stromov, ktoré tvoria les;
- počet premenných, ktoré sa náhodne vyberú ako uzol v strome (mtry);
- minimálny počet listov stromu.

Nájdenie vhodných parametrov sme realizovali pre *počet rozhodovacích stromov* a *mtry* pomocou vyhľadávania mriežkou (angl. grid search), čo predstavuje skúšanie všetkých kombinácií z preddefinovaných množín parametrov s cieľom nájsť kombináciu, ktorá dosiahne najlepšiu úspešnosť. Po nájdení najlepších parametrov pre klasifikáciu párov otázok sme ešte následne systematicky upravovali parametre s cieľom dosiahnuť čo najlepšiu úspešnosť pre otázku s najväčšou pravdepodobnosťou, že je pár otázok podobný.

V prípade klasifikátora *náhodný les* sme dosiahli väčšiu úspešnosť na črtách identifikovaných pomocou GBM. Použili sme črty, kde bola dôležitosť čŕt väčšia ako 0,2. Dôležitosť všetkých čŕt je uvedená v prílohe D. Odstránenie korelovaných atribútov znížilo úspešnosť. Najlepšiu úspešnosť sme dostali pri minimálnej pravdepodobnosti $\gamma = 0{,}92$. Finálne parametre pre model s edukačnými črtami sú: *počet stromov* = 300, *mtry* = 5, *minimálny počet listov* = 10. Pre model bez edukačných čŕt sú finálne parametre: *počet stromov* = 400, *mtry* = 5, *minimálny počet listov* = 10.

Úspešnosť odporúčanej najpodobnejšej otázky v tabuľke 7.5. Na výsledkoch môžeme vidieť, že klasifikátor *náhodný les* už nepredikuje podobné otázky aj pre otázky, ktoré nemajú podobnú otázku.

[3] `https://cran.r-project.org/package=randomForest`

**Tabuľka 7.5:** Úspešnosť klasifikátora náhodný les vzhľadom na najpodobnejšiu predikovanú otázku. Pravdepodobnosť $\gamma = 0{,}92$. V testovacej množine je 76 otázok, ktoré majú podobnú otázku a 218 otázok, ktoré nemajú podobnú otázku. Stĺpce predstavujú rovnaké atribúty ako v tabuľke 7.4.

| | Dobré odp. | Zlé odp. | Nemalo odporúčať | Success@1 | Presnosť pre pod. otázky | Úplnosť pre pod. otázky |
|---|---|---|---|---|---|---|
| NL GBM črty | 13 | 1 | 1 | 0,7823 | 0,9285 | 0,1842 |
| NL GBM bez edu. | 13 | 2 | 2 | 0,7789 | 0,8666 | 0,1973 |

### 7.3.4 SVM

Klasifikátor *support vector machines* (SVM) z knižnice e1071 nám umožňuje nastaviť:

- jadro (angl. kernel) (štandardne radiálne);
- parameter *cost* - cenu porušenia obmedzení (štandardne 1);
- parameter *epsilon* (štandardne 0,1),
- parameter *tolerance* (štandardne 0,001),
- váhy klasifikačných tried.

Ako prvé sme nastavili jadro na lineárne, čím sme dosiahli prvotné výrazné zvýšenie úspešnosti klasifikácie. Nastavenie parametrov *cost* a *epsilon* sme realizovali pomocou vyhľadávania mriežkou. Následne sme nastavili váhy tried na 1 pre nie podobné páry otázok a 10 pre podobné páry otázok. Napriek tomu, že je trénovacia množina vyvážená, chceme dať väčšiu váhu na správne klasifikovanie podobných párov. Počas vyhodnocovania modelu sme systematicky menili parametre *cost* a *epsilon* s cieľom nájdenia najvhodnejšej kombinácie.

SVM klasifikátor dosiahol lepšiu úspešnosť na RFE črtách. Odstránenie korelovaných atribútov znížilo úspešnosť. Pravdepodobnosť $\gamma$ má hodnotu 0,93 pre RFE črty a 0,92 pre RFE črty bez edukačných čŕt. Najlepšie parametre modelu pre RFE črty boli: *cost* = 50; *epsilon* = 0,1; *tolerance* = 0,5 a váhy tried 1 a 10. Parametre modelu bez edukačných čŕt boli: *cost* = 50; *epsilon* = 0,1; *tolerance* = 0,1 a váhy tried 1 a 10.

Úspešnosť odporúčanej najpodobnejšej otázky uvádzame v tabuľke 7.6. Oproti klasifikátoru náhodný les môžeme pozorovať, že úspešnosť klasifikácie sa bez použitia edukačných čŕt znížila výraznejšie.

**Tabuľka 7.6:** Úspešnosť klasifikátora SVM vzhľadom na najpodobnejšiu predikovanú otázku. Pravdepodobnosť $\gamma = 0{,}93$ pre variant s edukačnými črtami a $\gamma = 0{,}92$ bez edukačných. V testovacej množine je 76 otázok, ktoré majú podobnú otázku a 218 otázok, ktoré nemajú podobnú otázku. Stĺpce predstavujú rovnaké atribúty ako v tabuľke 7.4.

| | Dobré odp. | Zlé odp. | Nemalo odporúčať | Success@1 | Presnosť pre pod. otázky | Úplnosť pre pod. otázky |
|---|---|---|---|---|---|---|
| SVM RFE črty | 15 | 1 | 2 | 0,7857 | 0,9375 | 0,2105 |
| SVM RFE bez edu. | 11 | 1 | 2 | 0,7721 | 0,9166 | 0,1578 |

### 7.3.5 Zhrnutie vyhodnotenia úspešnosti klasifikátorov

V rámci overenia identifikovania podobných otázok sme pozorovali, že klasifikátor *naivný bayesov klasifikátor* je nevhodný pre použitie na takúto úlohu. Náš záver vyplýva z faktu, že na množstvo otázok, ktoré nemajú podobnú otázku, dokázal nájsť otázky, o ktorých si s veľkou pravdepodobnosťou myslel, že sú podobné. V prípade klasifikátorov *náhodný les* a *SVM* sme pozorovali pomerne veľkú presnosť, pričom *SVM* dosiahlo najväčšiu presnosť aj pokrytie pre podobné otázky. V prípade oboch klasifikátorov pozorujeme, že odstránenie edukačných čŕt viedlo k zníženiu úspešnosti.

## 7.4 Overenie zoradenia odpovedí

V tejto kapitole opisujeme vyhodnotenie kroku *zoradenie odpovedí* (kapitola 5.1.5) navrhnutej metódy realizovaný pre dáta z dvoch iterácií kurzu *MITx: 6.00.2x Introduction to Computational Thinking and Data Science*. Overenie vykonávame v dvoch krokoch. V prvom kroku vyhodnocujeme zoradenie odpovedí pre identifikované podobné otázky zo zlatého štandardu. V druhom kroku pomocou klasifikátora natrénovaného v predchádzajúcej časti overenia získame identifikované podobné páry otázok a zoraďujeme odpovede pre takto identifikované páry. Vďaka tomuto prístupu v druhom kroku overenia zoraďujeme aj odpovede, ktoré pochádzajú z otázok, ktoré nie sú vôbec podobné s novou otázkou.

### 7.4.1 Metodológia overenia identifikovania podobných otázok

Zoradenie odpovedí overujeme pomocou techniky *učenie sa zoraďovať*, ktorú sme opísali v kapitole 5.1.5. Používame algoritmus $SVM^{rank}$ opísaný v kapitole 6.5.4. Pri vyhodnocovaní porovnávame variant s črtami špecifickými pre doménu vzdelávania s variantom bez takýchto čŕt. Pre algoritmus $SVM^{rank}$ sme použili predvolené nastavenie parametrov.

Pre prvý krok overenia nad odpoveďami z podobných otázok je v dátovej sade 1342 párov otázka-odpoveď. Dátovú sadu delíme na trénovaciu a testovaciu časť tak, že v trénovacej časti je náhodných 75% otázok (226 otázok) a v testovacej 25% (76 otázok).

Úspešnosť zoradenia odpovedí sme sa rozhodli vyhodnotiť pomocou metrík presnosť@1, MRR a nDCG. Všetky metriky sme opísali v kapitole 2.3.8.

### 7.4.2 Vyhodnotenie pre odpovede z podobných otázok

V prvom kroku vyhodnocujeme úspešnosť metódy pre odpovede z podobných otázok identifikovaných na základe manuálneho anotovania opísaného v kapitole 7.2. Väčšiu úspešnosť sme dosiahli pri použití čŕt identifikovaných pomocou RFE. V tabuľke 7.7 uvádzame výslednú úspešnosť. Ako môžeme vidieť, variant bez edukačných čŕt bol úspešnejší.

**Tabuľka 7.7:** Úspešnosť algoritmu $SVM^{rank}$ pri zoraďovaní odpovedí z podobných otázok.

| | P@1 | MRR | nDCG |
|---|---|---|---|
| S edukačnými črtami | 0,7115 | 0,7485 | 0,8929 |
| Bez edukačných čŕt | 0,7500 | 0,7783 | 0,9041 |

### 7.4.3 Vyhodnotenie pre odpovede z párov podobných otázok identifikovaných klasifikátorom

V druhom kroku vyhodnocujeme zoradenie otázok na pároch otázok identifikovaných pomocou klasifikátora náhodný les. Náhodný les sme uprednostnili pred SVM, pretože pri klasifikácii párov otázok dosahoval náhodný les lepšiu úspešnosť. Pre identifikovanie párov otázok sme parameter minimálnej potrebnej pravdepodobnosti $\gamma$ nastavili na 0,88. Takto vytvorená dátová sada obsahovala 688 párov otázok v trénovacej časti a 230 párov otázok v testovacej časti.

Výber najvhodnejších čŕt sme vykonali len pomocou RFE. Vybratých bolo 7 čŕt (z 73):

1. pomer najlepších odpovedí ku všetkým odpovediam používateľa;
2. dĺžka odpovede;
3. rozdiel času vytvorenia odpovede a otázky;
4. informácia, či je autor odpovede učiteľ;
5. počet unikátnych stop slov;
6. počet stop slov;
7. skóre vo forme hlasovania.

Pre RFE bez edukačných čŕt to boli črty:

1. pomer najlepších odpovedí ku všetkým odpovediam používateľa;
2. dĺžka odpovede;
3. rozdiel času vytvorenia odpovede a otázky;
4. skóre vo forme hlasovania.
5. počet unikátnych stop slov;
6. počet stop slov.

Výslednú úspešnosť uvádzame v tabuľke 7.8. Môžeme pozorovať zhoršenie úspešnosti voči zoraďovaniu odpovedí len z podobných otázok (tabuľka 7.7). Zaujímavým pozorovaním je, že rozdiel medzi úspešnosťou s a bez edukačných čŕt je v tomto prípade minimálna.

Pri vyhodnotení naučeného modelu nad testovacou množinou z párov podobných otázok (tabuľka 7.9) vidíme, že úspešnosť pre variant *s edukačnými črtami* je vyššia než v tabuľke 7.8 avšak stále je nižšia než v prípade učenia a vyhodnotenia nad odpoveďami z podobných otázok (tabuľka 7.7).

**Tabuľka 7.8:** Úspešnosť algoritmu $SVM^{rank}$ pri zoraďovaní odpovedí z podobných otázok aj z nie podobných párov otázok. Zoraďovali sa odpovede pre 29 otázok.

| | P@1 | MRR | nDCG |
|---|---|---|---|
| S edukačnými črtami | 0,6896 | 0,7637 | 0,8688 |
| Bez edukačných čŕt | 0,6896 | 0,7563 | 0,8670 |

**Tabuľka 7.9:** Úspešnosť algoritmu $SVM^{rank}$ pri naučení sa aj na nerelevantných odpovediach a vyhodnotení nad dátami z podobných párov otázok. Zoraďovali sa odpovede pre 76 otázok.

| | P@1 | MRR | nDCG |
|---|---|---|---|
| S edukačnými črtami | 0,7307 | 0,7463 | 0,8923 |
| Bez edukačných čŕt | 0,6923 | 0,7724 | 0,8952 |

### 7.4.4 Zhrnutie vyhodnotenia zoradenia odpovedí

Najväčšiu presnosť sme dosiahli pri natrénovaní modelu zoraďovania otázok pre odpovede z párov podobných otázok. V tomto prípade bola úspešnosť pre variant bez edukačných čŕt o 3,9% väčšia pre metriku P@1. Pri spojení identifikovania podobných otázok a zoradenia odpovedí zoraďujeme aj odpovede z nie podobných otázok a preto sme natrénovali model na takýchto odpovediach. Vyhodnotenie úspešnosti modelu však preukázalo nižšiu úspešnosť.

## 7.5 Celkové overenie metódy

Po natrénovaní modelov identifikovania podobných otázok a zoradenia odpovedí môžeme vyhodnotiť celkovú úspešnosť metódy. Z testovacej množiny párov otázok sme odporučili všetky otázky spĺňajúce príslušnú minimálnu podobnosť $\gamma$, tak ako bola nastavená pre jednotlivé varianty opísané v kapitolách 7.3.3 a 7.3.4. Identifikované páry sme posunuli ako vstup pre natrénovaný model $SVM^{rank}$. Ako natrénovaný model sme použili model z prvého kroku vyhodnotenia zoradenia odpovedí - naučený na odpovediach z párov podobných otázok. Tento model dosiahol väčšiu úspešnosť a bol naučený a vyhodnotený na väčšom množstve dát a preto je vhodnejší na použitie.

V celkovom overení metódy vyhodnocujeme dve kombinácie variantov identifikovania podobných otázok a zoradenia odpovedí:

1. Použitie edukačných čŕt pri identifikácii podobných otázok (IPO) a použitie edukačných čŕt pri zoraďovaní otázok (ZO).
2. Nepoužitie edukačných čŕt pri IPO a nepoužitie edukačných čŕt pri ZO.

Úspešnosť pre klasifikátor náhodný les uvádzame v tabuľke 7.10 a pre SVM v tabuľke 7.11. V prvom stĺpci tabuliek je počet otázok, pre ktoré boli zoraďované odpovede a v druhom

**Tabuľka 7.10:** Celková úspešnosť navrhnutej metódy vo forme odporúčania otázok. Hodnoty v tabuľke sú pre identifikovanie podobných otázok pomocou klasifikátora náhodný les.

| | Všetky | Z pod. | S 1 odp. | Z pod. s 1 odp. | P@1 | MRR | nDCG |
|---|---|---|---|---|---|---|---|
| S edukačnými črtami | 18 | 16 | 7 | 7 | 0,7222 | 0,8472 | 0,8667 |
| Bez edukačných čŕt | 15 | 12 | 7 | 5 | 0,6666 | 0,7333 | 0,7720 |

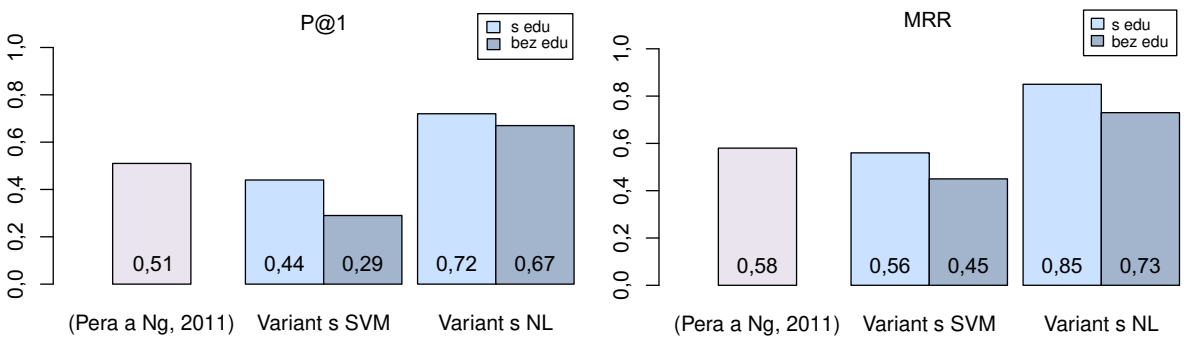
**Tabuľka 7.11:** Celková úspešnosť navrhnutej metódy vo forme odporúčania otázok. Hodnoty v tabuľke sú pre identifikovanie podobných otázok pomocou klasifikátora SVM.

| | Všetky | Z pod. | S 1 odp. | Z pod. s 1 odp. | P@1 | MRR | nDCG |
|---|---|---|---|---|---|---|---|
| S edukačnými črtami | 18 | 13 | 7 | 5 | 0,4444 | 0,5555 | 0,6483 |
| Bez edukačných čŕt | 17 | 11 | 6 | 3 | 0,2941 | 0,4495 | 0,5421 |

počet otázok, ktorých zoraďované odpovede pochádzali z aspoň jednej podobnej otázky. Pri interpretovaní výsledkov metrík v posledných troch stĺpcoch je potrebné si uvedomiť, že pre niektoré otázky mohla byť zoraďovaná len jedna odpoveď. Počet takýchto otázok uvádzame v treťom stĺpci. Druhou dôležitou informáciou je, že v prípade ak bola zoraďovaná len jedna odpoveď a tá pochádzala z páru nie podobných otázok (nesprávne pozitívny výsledok pri klasifikácii podobných otázok), tak sme úspešnosť daného zoradenia určili ako 0. V stĺpci 4 tabuliek 7.10 a 7.11 uvádzame počet otázok s jednou odpoveďou, ktoré pochádzali z páru podobných otázok - v takom prípade bola úspešnosť zoradenia automaticky 1.

Z výsledkov vidíme, že rozdiel úspešnosti pre zoradenie odpovedí identifikovanými pomocou náhodný les a SVM je výrazný. Variant s klasifikátorom náhodný les dosiahol o 28% väčšiu úspešnosť pre metriku P@1, ktorá reflektuje odporúčanie jednej odpovede. Oba varianty dosiahli nižšiu úspešnosť pri odstránení edukačných čŕt, čo potvrdzuje našu hypotézu.

Hoci žiadna z predchádzajúcich prác sa nevenuje priamo odpovedaniu na otázky v doméne vzdelávania, môžeme porovnať naše výsledky s autormi Pera a Ng (2011). Autori Pera a Ng (2011) dosiahli P@1 = 0,51, MRR = 0,58 na dátach z CQA systému Yahoo! Answers. Úspešnosť autorov Pera a Ng (2011) sa nám podarilo výrazne prekonať. Porovnanie úspešnosti našej metódy s prácou (Pera a Ng, 2011) vizualizujeme na obr. 7.4.




**Obr. 7.4:** Porovnanie úspešnosti celkového overenia metódy v práci (Pera a Ng, 2011) a našej metódy. Na obrázku vľavo je úspešnosť podľa metriky P@1, vpravo podľa MRR.

## 7.6 Sumarizácia výsledkov

V tejto kapitole sme opísali realizáciu offline experimentu nad dátami z kurzu, ktorý sa konal na edX platforme. Nedostupnosť zlatého štandardu podobných otázok v doméne CQA systémov a rovnako MOOC kurzov reprezentuje veľký problém, ktorý sme sa rozhodli vyriešiť stiahnutím dát z webu a následného manuálneho anotovania s účelom vytvorenia zlatého štandardu podobných otázok a zoradenia odpovedí.

Slabým miestom nášho overenia navrhnutej metódy je manuálne vytvorenie zlatého štandardu. Iba malá časť dátovej sady bola vytvorená ako výsledok práce dvoch anotátorov. Väčšina dátovej sady bola oanotovaná len autorom tejto práce.

Pomocou experimentálneho overenia sa nám podarilo identifikovať minimálnu potrebnú pravdepodobnosť na to, aby bol pár otázok klasifikátorom označený ako podobný. Tento parameter $\gamma$ využívame v kroku *orezanie zoznamu podobných otázok*. Ako najvhodnejšia hodnota $\gamma$ sa ukázala hodnota 0,92.

Dva hlavné kroky navrhnutej metódy sme najskôr overili samostatne a na záver sme overili aj ich spojenie. Zistili sme, že pri nepoužití čŕt špecifických pre doménu vzdelávania dochádza k zníženiu úspešnosti, čo potvrdzuje našu hypotézu. Výsledky celkového overenia metódy ukazujú, že s pomerne vysokou presnosťou vieme zodpovedať novú otázku (metrika P@1 = 0,72) - približne na 3 otázky zo 4 by naša metóda odporučila najvhodnejšiu odpoveď. S takouto úspešnosťou by sme zodpovadali 18% otázok, ktoré majú v archíve podobnú otázku, čo potvrdzuje našu hypotézu o zodpovedaní signifikantného počtu otázok.

Pre overenie navrhnutej metódy sme zvažovali živý (angl. online) experiment, kde by sme v reálnom čase odpovedali na otázky v online kurze. Aposteriórne vyhodnotenie naučeného modelu na dátach z kurzu zameraného na výučbu chémie však ukázalo, že náš model bol veľmi nepresný a bez doučenia modelu na nových oanotovaných dátach by jeho použitie nebolo vhodné. Z tohto dôvodu sme živý experiment nezrealizovali. Prenositeľnosť modelu nebola možná pravdepodobne kvôli charakteru dát z kurzu. Tento kurz používal iba dve úrovne hierarchie kategórií a na základe textov otázok, ktoré sme videli, usudzujeme že bol viac zameraný len na jednu konkrétnu tému.

# Kapitola 8

# Zhodnotenie

V tejto práci sme sa venovali podpore používateľov v online komunitách študentov s využitím archívu otázok a odpovedí. Najväčšími archívmi otázok a odpovedí disponujú v súčasnosti CQA systémy. Z tohto dôvodu sme analyzovali CQA systémy a podporu spolupráce používateľov v týchto systémoch. Identifikovali sme niekoľko spôsobov podpory spolupráce používateľov a rozhodli sme sa zamerať na využitie archívu otázok a odpovedí na automatické zodpovedanie nových otázok.

Zvolený spôsob podpory používateľov v našej práci aplikujeme vo vzdelávacej doméne. Za účelom lepšieho porozumenia špecifík online komunít sme tieto komunity bližšie analyzovali, pričom sme sa zamerali predovšetkým na hromadné otvorené online kurzy (angl. massive open online courses - MOOC). Pozorovali sme, že skoro žiadne systémy poskytujúce online kurzy nedisponujú CQA systémami a na komunikáciu študentov využívajú diskusné fóra. Fóra však nedisponujú tak dobrou štruktúrou otázka-odpovede-komentáre, akou sú známe CQA systémy na otvorenom webe.

Podpora používateľov prostredníctvom automatického odpovedania na otázky je dôležitá, nakoľko slúži ako prevencia vytvárania duplicitných otázok. Problém s duplicitnými otázkami majú rovnako CQA systémy ako aj online komunity študentov. V prípade online kurzov, kde sa často nachádza 1000 študentov na jedného učiteľa, duplicitné otázky prispievajú k preťaženiu učiteľa, ktorý sa potom nemá čas venovať prioritnejším otázkam.

Na základe analýzy sme navrhli metódu automatického odpovedania na otázky s využitím špecifických informácií pre online komunity študentov. Výslednú podobnosť otázok určujeme na základe textovej podobnosti otázok a časovej a kategorickej príslušnosti otázok. Textovú podobnosť sme realizovali na základe TF-IDF, LDA a GloVe. Pri výbere odpovede, ktorú použijeme na automatické odpovedanie zvažujeme kvalitu odpovede a expertízu používateľa a textovú podobnosť medzi novou otázkou a odpoveďou (na základe TF-IDF). V procese návrhu metódy sme zohľadnili špecifiká domény vzdelávania a na ich základe definovali črty pre identifikáciu podobných otázok a zoradenie odpovedí, ktoré sa vo všeobecnej doméne nenachádzajú.

Navrhnutú metódu sme realizovali a vyhodnotili na dátach z kurzu týkajúceho sa informačných technológií, ktorý bol vyučovaný na platforme edX. Dáta sme získali vytvorením skriptu, ktorý ich stiahol počas toho ako prechádzal diskusie na fóre. Zlatý štandard sme vytvorili manuálne anotovaním podobnosti otázok a vhodnosti odpovedí. Anotovanie sme vykonali pred samotným experimentovaním. Časť dátovej sady podobnosti otázok bola oanotovaná dvoma anotátormi. Za účelom získania anotácií od viacerých anotátorov sme vytvorili online nástroj na anotovanie podobnosti otázok.

Metódu sme overili zvlášť pre dva hlavné komponenty našej metódy - identifikovanie podobných otázok a zoradenie odpovedí. Z dôvodu veľkého počtu čŕt sme použili výber najdôležitejších čŕt pomocou rekurzívnej eliminácie čŕt a modelov založených na gradientovom zosilnení. Klasifikáciu podobných otázok sme vykonali troma klasifikátormi: naivným bayesovým klasifikátorom, náhodným lesom a SVM. SVM dosiahlo najlepšiu úspešnosť pri odporúčaní najpodobnejšej otázky, avšak náhodný les bol úspešnejší pri klasifikovaní jednotlivých párov otázok.

Zoradenie odpovedí sme overovali pre zoradenie odpovedí z len podobných otázok a aj pre zoradenie odpovedí z nie podobných otázok. Výsledky preukázali, že model natrénovaný na odpovediach z len podobných otázok bol úspešnejší.

Celkové overenie metódy (overenie spojenia algoritmov pre identifikáciu podobných otázok a zoradenie odpovedí) ukázalo pomerne vysokú úspešnosť pri odporúčaní jednej odpovede (P@1 = 0,72). Doméne automatického odpovedania na otázky vo vzdelávacej doméne sa podľa nášho vedomia venujeme ako prví. Práca najbližšia k tej našej (Pera a Ng, 2011) sa venovala odpovedaniu na otázkach zo systému Yahoo! Answers. Úspešnosť našej metódy automatického odpovedania na otázky je lepšia než úspešnosť autorov (Pera a Ng, 2011), ktorí dosiahli úspešnosť 0,51 pre metriku P@1.

Ako možnosti ďalšej práce vidíme overenie metódy aposteriórne nad neoanotovanou časťou druhej iterácie kurzu, s ktorým sme pracovali. Druhou možnosťou ďalšieho overenia je realizácia živého (angl. online) experimentu. Ako ďalšie možnosti zlepšenia modelu vidíme vylepšenie čŕt, ktoré sme použili pri klasifikácii otázok a zoradení odpovedí. Napriek tomu, že sme identifikovali veľký počet čŕt, nie všetky sú zvolené vhodne a napríklad črty zachytávajúce počet určitých slov by bolo lepšie počítať relatívne voči otázke, z ktorej pochádzajú.

# Literatúra


1. ADAMSON, David; DYKE, Gregory; JANG, Hyeju; ROSÉ, Carolyn Penstein, 2014. Towards an Agile Approach to Adapting Dynamic Collaboration Support to Student Needs. *International Journal of Artificial Intelligence in Education*. Roč. 24, č. 1, s. 92–124. ISSN 1560-4292.
2. ALARIO-HOYOS, Carlos; PÉREZ-SANAGUSTÍN, Mar; DELGADO-KLOOS, Carlos; PARADA G., Hugo A.; MUÑOZ-ORGANERO, Mario; HERAS, Antonio Rodríguez-de-las, 2013. Analysing the Impact of Built-In and External Social Tools in a MOOC on Educational Technologies. In: HERNÁNDEZ-LEO, Davinia; LEY, Tobias; KLAMMA, Ralf; HARRER, Andreas (ed.). *Proceedings of 8th European Conference, on Technology Enhanced Learning - EC-TEL '13*. Paphos, Cyprus: Berlin, Heidelberg, s. 5–18.
3. ALARIO-HOYOS, Carlos; PEREZ-SANAGUSTIN, Mar; DELGADO-KLOOS, Carlos; PARADA G., Hugo A.; MUNOZ-ORGANERO, Mario, 2014. Delving into Participants' Profiles and Use of Social Tools in MOOCs. *IEEE Transactions on Learning Technologies*. Roč. 7, č. 3, s. 260–266. ISSN 1939-1382.
4. ANDERSON, Ashton; HUTTENLOCHER, Daniel; KLEINBERG, Jon; LESKOVEC, Jure, 2014. Engaging with massive online courses. In: *Engaging with massive online courses. Proceedings of the 23rd international conference on World wide web - WWW '14*. New York, New York, USA: ACM Press, s. 687–698. ISBN 9781450327442.
5. ARITAJATI, Chulakorn; NARAYANAN, N. Hari, 2013. Facilitating Students' Collaboration and Learning in a Question and Answer System. In: *Facilitating Students' Collaboration and Learning in a Question and Answer System. Proceedings of the 2013 conference on Computer supported cooperative work companion - CSCW '13*. New York, New York, USA: ACM Press, s. 101–106. ISBN 9781450313322.
6. BELINKOV, Yonatan; MOHTARAMI, Mitra; CYPHERS, Scott; GLASS, James, 2015. VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems. In: *VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, s. 282–287.
7. CAO, Xin; CONG, Gao; CUI, Bin; JENSEN, Christian Søndergaard; ZHANG, Ce, 2009. The Use of Categorization Information in Language Models for Question Retrieval. In:

*The Use of Categorization Information in Language Models for Question Retrieval*. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. Hong Kong, China: ACM, s. 265–274. CIKM '09. ISBN 978-1-60558-512-3.

8. CAO, Xin; CONG, Gao; CUI, Bin; JENSEN, Christian S., 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In: *A generalized framework of exploring category information for question retrieval in community question answer archives*. *Proceedings of the 19th international conference on World wide web - WWW '10*. New York, New York, USA: ACM Press, s. 201–210. ISBN 9781605587998.
9. CHEN, Long; ZHANG, Dell; LEVENE, Mark, 2013. Question retrieval with user intent. In: *Question retrieval with user intent*. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*. New York, New York, USA: ACM Press, s. 973–976.
10. COETZEE, Derrick; FOX, Armando; HEARST, Marti A.; HARTMANN, Bjoern, 2014. Chatrooms in MOOCs. In: *Chatrooms in MOOCs*. *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*. New York, New York, USA: ACM Press, s. 127–136. ISBN 9781450326698.
11. CUI, Yi; WISE, Alyssa Friend, 2015. Identifying Content-Related Threads in MOOC Discussion Forums. In: *Identifying Content-Related Threads in MOOC Discussion Forums*. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*. New York, New York, USA: ACM Press, s. 299–303. ISBN 9781450334112.
12. DUAN, Huizhong; CAO, Yunbo; CHIN-YEW, Lin; YU, Yong, 2008. Searching questions by identifying question topic and question focus. In: *Searching questions by identifying question topic and question focus*. *Proceedings of ACL-08: HLT*, s. 156–164.
13. FENG, Yunping; CHEN, Di; ZHAO, Zihao; CHEN, Haopeng; XI, Puzhao, 2015. The Impact of Students And TAs' Participation on Students' Academic Performance in MOOC. In: *The Impact of Students And TAs' Participation on Students' Academic Performance in MOOC*. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*. New York, New York, USA: ACM Press, s. 1149–1154.
14. HECKING, Tobias; HOPPE, H. Ulrich; HARRER, Andreas, 2015. Uncovering the Structure of Knowledge Exchange in a MOOC Discussion Forum. In: *Uncovering the Structure of Knowledge Exchange in a MOOC Discussion Forum*. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*. New York, New York, USA: ACM Press, s. 1614–1615.
15. HUANG, Jonathan; DASGUPTA, Anirban; GHOSH, Arpita; MANNING, Jane; SANDERS, Marc, 2014. Superposter behavior in MOOC forums. In: *Superposter behavior in MOOC forums*. *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*. New York, New York, USA: ACM Press, s. 117–126. ISBN 9781450326698.

16. HUNA, Adrian; SRBA, Ivan; BIELIKOVA, Maria, 2016. Exploiting Content Quality and Question Difficulty in CQA Reputation Systems. In: *Exploiting Content Quality and Question Difficulty in CQA Reputation Systems. Proceedings of International Conference on Nework Science - NetSciX '16*. Berlin, Heidelberg: Springer Berlin Heidelberg, s. 68–81. Lecture Notes in Computer Science.

17. JI, Zongcheng; XU, Fei; WANG, Bin; HE, Ben, 2012. Question-answer topic model for question retrieval in community question answering. In: *Question-answer topic model for question retrieval in community question answering. Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*. New York, New York, USA: ACM Press, s. 2471–2474. ISBN 9781450311564.

18. JOACHIMS, Thorsten, 2006. Training Linear SVMs in Linear Time. In: *Training Linear SVMs in Linear Time. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Philadelphia, PA, USA: ACM, s. 217–226. KDD '06. ISBN 1-59593-339-5.

19. LI, Shuguang; MANANDHAR, Suresh, 2011. Improving question recommendation by exploiting information need. In: *Improving question recommendation by exploiting information need. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, s. 1425–1434. ISBN 978-1-932432-87-9.

20. LIU, Tie-Yan, 2009. Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval*. Roč. 3, č. 3, s. 225–331. ISSN 1554-0669.

21. MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey, 2013. Efficient Estimation of Word Representations in Vector Space.

22. PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D., 2014. GloVe: Global Vectors for Word Representation. In: *GloVe: Global Vectors for Word Representation. Empirical Methods in Natural Language Processing (EMNLP)*, s. 1532–1543. Dostupné tiež z: `http://www.aclweb.org/anthology/D14-1162`.

23. PERA, Maria Soledad; NG, Yiu-Kai, 2011. A Community Question-answering Refinement System. In: *A Community Question-answering Refinement System. Proceedings of the 22Nd ACM Conference on Hypertext and Hypermedia*. Eindhoven, The Netherlands: ACM, s. 251–260. HT '11. ISBN 978-1-4503-0256-2. Dostupné z DOI: `10.1145/1995966.1995999`.

24. RAM, Ashwin; AI, Hua; RAM, Preetha; SAHAY, Saurav, 2011. Open Social Learning Communities. In: *Open Social Learning Communities. Proceedings of the International Conference on Web Intelligence, Mining and Semantics - WIMS'11*. New York, New York, USA: ACM Press. ISBN 9781450301480.

25. SHATNAWI, Safwan; GABER, Mohamed Medhat; COCEA, Mihaela, 2014. Automatic Content Related Feedback for MOOCs Based on Course Domain Ontology. In: *Automatic Content Related Feedback for MOOCs Based on Course Domain Ontology. Intelligent Data*

*Engineering and Automated Learning – IDEAL 2014: 15th International Conference*. Cham: Springer International Publishing, zv. 8669, s. 27–35. Lecture Notes in Computer Science.

26. SHTOK, Anna; DROR, Gideon; MAAREK, Yoelle; SZPEKTOR, Idan, 2012. Learning from the past: Answering New Questions with Past Answers. In: *Learning from the past: Answering New Questions with Past Answers. Proceedings of the 21st international conference on World Wide Web - WWW '12*. New York, New York, USA: ACM Press, s. 759–768. ISBN 9781450312295.
27. SRBA, Ivan; BIELIKOVA, Maria, 2015. Askalot. In: *Askalot. Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing - CSCW'15 Companion*. New York, New York, USA: ACM Press, s. 179–182. ISBN 9781450329460.
28. SRBA, Ivan; BIELIKOVA, Maria, 2016a. A Comprehensive Survey and Classification of Approaches for Community Question Answering. *ACM Trans. Web*. Roč. 10, č. 3, s. 18:1–18:63. ISSN 1559-1131. Dostupné z DOI: 10.1145/2934687.
29. SRBA, Ivan; BIELIKOVA, Maria, 2016b. Design of CQA Systems for Flexible and Scalable Deployment and Evaluation. In: *Design of CQA Systems for Flexible and Scalable Deployment and Evaluation. Proceedings of the 16th International Conference on Web Engineering - ICWE '16*, To appear.
30. SURYANTO, Maggy Anastasia; LIM, Ee Peng; SUN, Aixin; CHIANG, Roger H. L., 2009. Quality-aware Collaborative Question Answering: Methods and Evaluation. In: *Quality-aware Collaborative Question Answering: Methods and Evaluation. Proceedings of the Second ACM International Conference on Web Search and Data Mining*. Barcelona, Spain: ACM, s. 142–151. WSDM '09. ISBN 978-1-60558-390-7. Dostupné z DOI: 10.1145/1498759.1498820.
31. WANG, Jun; HU, Xia; LI, Zhoujun; CHAO, Wenhan; HU, Biyun, 2011. Learning to recommend questions based on public interest. In: *Learning to recommend questions based on public interest. Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, New York, USA: ACM Press, s. 2029–2032.
32. WANG, Kai; MING, Zhaoyan; CHUA, Tat-Seng, 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In: *A syntactic tree matching approach to finding similar questions in community-based qa services. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*. New York, New York, USA: ACM Press, s. 187–194. ISBN 9781605584836.
33. WU, Haocheng; WU, Wei; ZHOU, Ming; CHEN, Enhong; DUAN, Lei; SHUM, Heung-Yeung, 2014. Improving search relevance for short queries in community question answering. In: *Improving search relevance for short queries in community question answering. Proceedings of the 7th ACM international conference on Web search and data mining - WSDM '14*. New York, New York, USA: ACM Press, s. 43–52. ISBN 9781450323512.

34. XUE, Xiaobing; JEON, Jiwoon; CROFT, W. Bruce, 2008. Retrieval models for question and answer archives. In: *Retrieval models for question and answer archives. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*. New York, New York, USA: ACM Press, s. 475–482. ISBN 9781605581644.

35. YANG, Diyi; ADAMSON, David; ROSÉ, Carolyn Penstein, 2014. Question Recommendation with Constraints for Massive Open Online Courses. In: *Question Recommendation with Constraints for Massive Open Online Courses. Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14*. New York, New York, USA: ACM Press, s. 49–56. ISBN 9781450326681.

36. YANG, Diyi; WEN, Miaomiao; HOWLEY, Iris; KRAUT, Robert; ROSE, Carolyn, 2015. Exploring the Effect of Confusion in Discussion Forums of Massive Open Online Courses. In: *Exploring the Effect of Confusion in Discussion Forums of Massive Open Online Courses. Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*. New York, New York, USA: ACM Press, s. 121–130. ISBN 9781450334112.

37. ZHANG, Kai; WU, Wei; WU, Haocheng; LI, Zhoujun; ZHOU, Ming, 2014. Question Retrieval with High Quality Answers in Community Question Answering. In: *Question Retrieval with High Quality Answers in Community Question Answering. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*. New York, New York, USA: ACM Press, s. 371–380. ISBN 9781450325981.

38. ZHENG, Saijing; ROSSON, Mary Beth; SHIH, Patrick C.; CARROLL, John M., 2015. Understanding Student Motivation, Behaviors and Perceptions in MOOCs. In: *Understanding Student Motivation, Behaviors and Perceptions in MOOCs. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*. New York, New York, USA: ACM Press, s. 1882–1895.

39. ZHOU, Guangyou; CAI, Li; ZHAO, Jun; LIU, Kang, 2011. Phrase-based translation model for question retrieval in community question answer archives. In: *Phrase-based translation model for question retrieval in community question answer archives. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Association for Computational Linguistics, s. 653–662. ISBN 978-1-932432-87-9.

# Príloha A

# Inštalačná príručka

## Požiadavky

Systém Askalot, ktorého databázovú štruktúru sme využili pre stiahnutie obsahu diskusných fór online kurzov v systéme edX má nasledujúce požiadavky na systém:

- Ruby 2.3
- Ruby on Rails 4.2
- PostgreSQL 9.3
- Elasticsearch 1.7

V rámci nášho riešenia sme pridali systémovú požiadavku na jazyk R (verzia 3.3) a prepojenie medzi jazykmi R a Ruby pomocou Rserve[1].

## Inštalácia

Po nainštalovaní všetkých požiadaviek samotný Askalot nainštalovať nasledovným spôsobom:

1. Skopírovať archív `askalot-metoda.zip` z priečinka `implementacia` a rozzipovať archív.
2. Importovať databázový súbor `metoda.sql` z priečinka `databaza`.
3. Spustiť v priečinku príkaz `bundle install`.
4. Upraviť konfiguračný súbor na databázové pripojenie `database.yml` v priečinku `config`.

[1] `https://rforge.net/Rserve/`

# Inštalačná príručka pre nástroj na anotovanie otázok

## Požiadavky

Nástroj na anotovanie otázok je napísaný v jazyku PHP a požiadavky na systém sú nasledovné:

- PHP >= 5.6.4
- PHP rozšírenia: OpenSSL, PDO, Mbstring, Tokenizer, XML
- PostgreSQL 9.3

## Inštalácia

1. Skopírovať archív `nastroj-na-anotovanie.zip` z priečinka `implementacia` a rozzipovať archív.
2. Importovať databázový súbor `nastroj-na-anotovanie.sql` z priečinka `databaza`.
3. Spustiť webserver, ktorý bude mať prístup do priečinka s nástrojom na anotovanie.

# Príloha B

# Návod na reprodukciu výsledkov

Navrhnutú metódu sa nám nepodarilo zrealizovať ako komplexný program, ktorému dáme vstup a na výstupe získame odporúčané odpovede. Z tohto dôvodu opisujeme komplexný postup príkazov, ktoré treba spustiť, ak chceme zreprodukovať výsledky opísané v tejto práci. Na spúšťanie R skriptov odporúčame použiť program Rstudio[1]. Proces postupu za účelom reprodukcie výsledkov vizualizujeme na obr. B.1.

1. Spustiť príkaz `export RAILS_ENV=development`, čím sa nastaví premenná prostredia.
2. Spustiť príkaz `rake dp:all_text_profiles`, čím sa v databáze lematizovaná reprezentácia textu otázok, odpovedí a komentárov.
3. Spustiť príkaz `rake dp:all_text_file`, čím sa vytvoria súbory z lematizovanej reprezentácie textov uloženej v databáze.
4. Vytvoriť súbory pre podobnosť podľa TF-IDF.
   - Krok treba vykonať ak pracujeme s iným kurzom než *MITx: 6.00.2x*, na ktorom sme overovali metódu v tejto práci. V takom prípade je potrebné zmazať súbory v priečinku `r/saved` a spustiť R skript `r/tfidf_similarity.r`.
5. Spustiť príkaz `rake dp:r_similarity_input`, čím sa vytvorí CSV súbor s pármi otázok a ich črtami.
6. Spustiť R skript `r/add_glove_columns.r`, čím sa vypočítajú črty založené na GloVe.
7. Identifikovať dôležité črty pomocou RFE a GBM.
   - Krok treba vykonať ak pracujeme s iným kurzom a chceme identikovať dôležité črty. V takom prípade je potrebné spustiť R skripty `r/gbm.r` a `r/rfe.r`. Na základe identifikovaných čŕt upraviť zoznam čŕt v súboroch `r/rf,svm_new,NaiveBayes.r`.
8. Identifikovať korelované črty pomocou R skriptu `r/correlation_heatmap.r`.

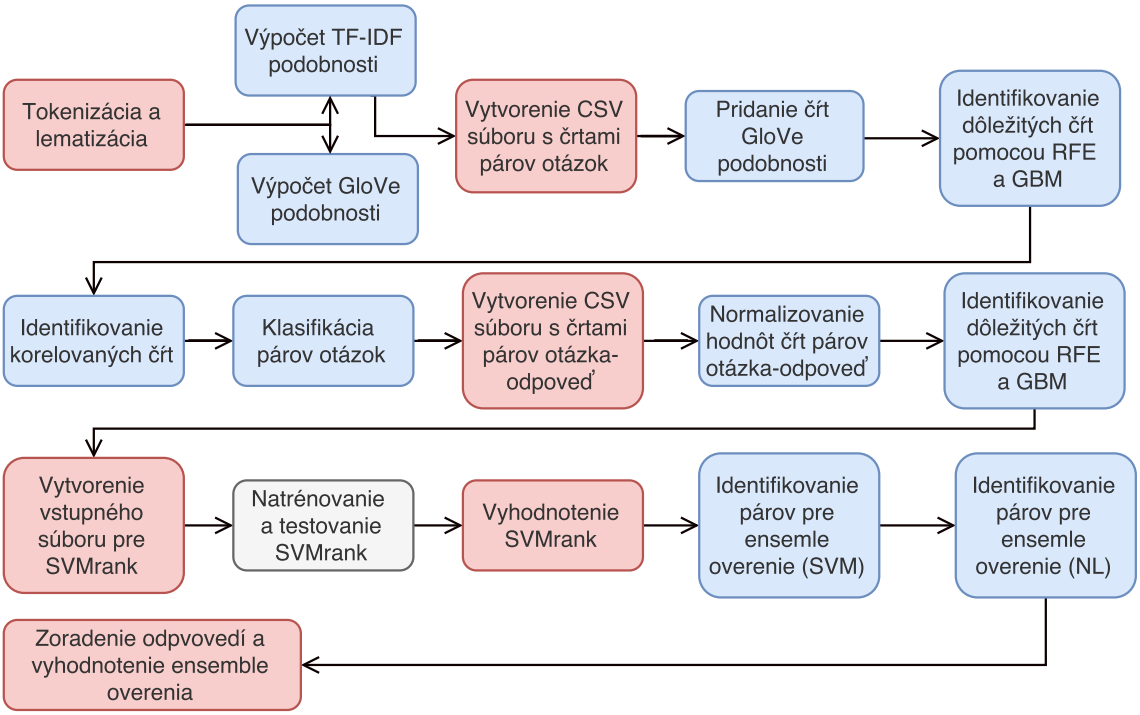
[1] `https://www.rstudio.com/`

- Krok treba vykonať ak pracujeme s iným kurzom a chceme identikovať korelované črty.

9. Vyhodnotiť klasifikáciu podobných otázok pomocou súborov `r/{rf,svm_new,NaiveBayes}.r`.
10. Spustiť príkaz `rake dp:learning_to_rank_csv_generator`, čím sa vytvorí CSV súbor s črtami pre zoraďovanie odpovedí.
11. Spustiť R skript `r/ltr_normalizacia.R`, čím sa znormalizujú dáta pre $SVM^{rank}$.
12. Identifikovať dôležité črty pomocou RFE a GBM.
    - Krok treba vykonať ak pracujeme s iným kurzom a chceme identikovať dôležité črty. V takom prípade je potrebné spustiť R skripty `r/ltr_gbm_feature_selection.r` a `r/ltr_rfe.r`. Na základe identifikovaných čŕt upraviť zoznam čŕt v súbore `components/shared/app/services/shared/dp/learning_to_rank_input_generator.rb` na riadkoch 15, 18 a 21.
13. Spustiť príkaz `rake dp:ltr_input_generator`, čím sa vytvorí vstup pre $SVM^{rank}$.
14. Prejsť do priečinka `r/ltr` a spustiť príkaz `sh script.sh`.
15. Spustiť príkaz `rake dp:ltr_evaluation`, čím sa vyhodnotí úspešnosť zoradenia odpovedí.
16. Spustiť R skript `r/rf_ensemble.r`, čím sa vygenerujú páry otázok pre overenie celkovej metódy podľa klasifikátora náhodný les.
17. Spustiť R skript `r/svm_ensemble.r`, čím sa vygenerujú páry otázok pre overenie celkovej metódy podľa klasifikátora SVM.
18. Spustiť príkaz `rake dp:ensemble_evaluation`, čím dostaneme úspešnosť pre spojenie identifikovania podobných otázok a zoradenia odpovedí.

Ak by sme chceli anotovať dáta, tak slúžia na to príkazy:

- `rails runner scripts/question_or_discussion.rb` - anotovanie typu príspevku (otázka alebo diskusia),
- `rake dp:similarity` - anotovanie typu podobnosti otázok,
- `rake dp:suitable_answers` - anotovanie vhodnosti odpovedí.




**Obr. B.1:** Schéma postupnosti jednotlivých krokov za účelom zreprodukovania výsledkov. Červeno podfarbené bloky predstavujú časti metódy realizované v jazyku R a modré v jazyku R.

# Príloha C

# Úspešnosť klasifikácie párov otázok

V tejto prílohe uvádzame úspešnosť klasifikácie párov otázok, ktorú vykonajú klasifikátory v kroku *identifikovanie podobných otázok* navrhnutej metódy.

### Naivný bayesov klasifikátor

Úspešnosť klasifikácie párov otázok pre naivný bayesov klasifikátor uvádzame v tabuľke C.1.

**Tabuľka C.1:** Úspešnosť naivného bayesovho klasifikátora pri klasifikácii párov otázok s črtami identifikovanými pomocou RFE. Pravdepodobnosť $\gamma = 0{,}91$. Tabuľka (a) predstavuje úspešnosť pre klasifikáciu s použitím edukačných čŕt a tabuľka (b) pre klasifikáciu bez použitia edukačných čŕt.

**(a)**

| | Správnosť | Presnosť | Úplnosť |
|---|---|---|---|
| Celkovo | 0,9176 | - | - |
| Podobné | - | 0,3034 | 0,4808 |
| Nie podobné | - | 0,9713 | 0,9409 |

**(b)**

| | Správnosť | Presnosť | Úplnosť |
|---|---|---|---|
| Celkovo | 0,9162 | - | - |
| Podobné | - | 0,2955 | 0,4699 |
| Nie podobné | - | 0,9707 | 0,9400 |

### Náhodný les

Úspešnosť klasifikácie párov otázok pre klasifikátor náhodný les (angl. random forest) uvádzame v tabuľke C.2.

**Tabuľka C.2:** Úspešnosť klasifikátora náhodný les pri klasifikácii párov otázok s črtami identifikovanými pomocou GBM. Pravdepodobnosť $\gamma = 0{,}92$. Tabuľka (a) predstavuje úspešnosť pre klasifikáciu s použitím edukačných čŕt a tabuľka (b) pre klasifikáciu bez použitia edukačných čŕt.

**(a)**

| | Správnosť | Presnosť | Úplnosť |
|---|---|---|---|
| Celkovo | 0,9542 | - | - |
| Podobné | - | 0,7647 | 0,1420 |
| Nie podobné | - | 0,9560 | 0,9976 |

**(b)**

| | Správnosť | Presnosť | Úplnosť |
|---|---|---|---|
| Celkovo | 0,9528 | - | - |
| Podobné | - | 0,6969 | 0,1256 |
| Nie podobné | - | 0,9551 | 0,9970 |

**SVM**

Úspešnosť klasifikácie párov otázok pre klasifikátor SVM uvádzame v tabuľke C.3.

**Tabuľka C.3:** Úspešnosť klasifikátora SVM pri klasifikácii párov otázok s črtami identifikovanými pomocou RFE. Tabuľka (a) predstavuje úspešnosť pre klasifikáciu s použitím edukačných čŕt $\gamma = 0{,}93$. V tabuľke (b) je úspešnosť pre klasifikáciu bez použitia edukačných čŕt a $\gamma = 0{,}92$.

**(a)**

| | Správnosť | Presnosť | Úplnosť |
|---|---|---|---|
| Celkovo | 0,9514 | - | - |
| Podobné | - | 0,6538 | 0,0928 |
| Nie podobné | - | 0,9535 | 0,9973 |

**(b)**

| | Správnosť | Presnosť | Úplnosť |
|---|---|---|---|
| Celkovo | 0,9514 | - | - |
| Podobné | - | 0,7500 | 0,0655 |
| Nie podobné | - | 0,9523 | 0,9988 |

# Príloha D

# Črty použité pre klasifikáciu podobných otázok

V tejto prílohe uvádzame zoznam použitých čŕt pre identifikovanie podobných otázok. Črty môžeme rozdeliť do dvoch kategórií: 1) črty špecifické pre otázky a 2) črty spoločné pre jednotlivé časti otázok (nadpis otázky (N), text otázky (T), nadpis+text (NT), časť s opytovacími vetami (OV)). Črty patriace do prvej kategórie sú:

- podobnosti ako umiestnenia kategórie v hierarchii kategórií (`hierarchia_kategorii`);
- binárna informácia, či sú otázky z rovnakej kategórie (`rovnaka_kategoria`);
- binárna informácia, či sú otázky z rovnakej kategórie podľa identifikátora zo systému edX. (`rovnaka_kategoria_edx`);
- binárna informácia, či sú otázky z rovnakej iterácie kurzu (`rovnaka_iteracia`);
- rozdiel v čase pridania prvej otázky v kategóriách, do ktorých otázky patria (`cas_kategoria`);
- rozdiel v čase pridania otázok relatívne ku prvej otázke v kurze/predmete (`cas`).

Všetky tieto črty sú špecifické pre doménu vzdelávania. Nasledujúce črty boli aplikované pre všetky časti otázok (`N`, `T`, `NT`, `OV`):

- podobnosť na základe TF-IDF reprezentácie (`TFIDF`);
- podobnosť na základe GloVe vektorov (`GloVe`);
- pomer dĺžky textov (`pomer_dlzky`);
- pomer dĺžky textov bez stop slov (`pomer_dlzky_bez_stop_slov`);
- pomer počtu stop slov (`pomer_stop_slov`);
- pomer počtu unikátnych stop slov (`pomer_unikatnych_stop_slov`);
- pomer počtu otáznikov (`pomer_otaznikov`);
- pomer zhodných unikátnych slov (`unikatne_slova`);
- pomer počtu zhodných unikátnych slov ku väčšiemu počtu unikátnych slov v textoch (`unikatne_slova_pomer`);

- počet zhodných bigramov (`bigramy`);
- pomer počtu zhodných bigramov ku väčšiemu počtu unikátnych bigramov v textoch (`bigramy_pomer`);
- počet zhodných podstatných mien (`podstatne_mena`);
- počet zhodných podstatných mien ku väčšiemu počtu unikátnych podstatných mien v textoch (`podstatne_mena_pomer`);
- počet zhodných podstatných mien ku väčšiemu počtu unikátnych slov v textoch (`podstatne_mena_pomer_text_viac`);
- počet zhodných podstatných mien ku menšiemu počtu unikátnych slov v textoch (`podstatne_mena_pomer_text_menej`);
- počet zhodných slovies (`slovesa`);
- počet zhodných slovies ku väčšiemu počtu unikátnych slovies v textoch (`slovesa_pomer`);
- počet zhodných slovies ku väčšiemu počtu unikátnych slov v textoch (`slovesa_pomer_text_menej`);
- počet zhodných slovies ku menšiemu počtu unikátnych slov v textoch (`slovesa_pomer_text_viac`);
- počet zhodných opytovacích zámien (`opytovacie_zamena`);
- pomer počtu zhodných opytovacích zámien ku väčšiemu počtu unikátnych opytovacích zámien v textoch (`opytovacie_zamena_pomer`).

Celkovo tak pracujeme s 90 črtami.

## Zoznam korelovaných čŕt pre identifikovanie podobných otázok

Predovšetkým z dôvodu použitia naivného bayesovho klasifikátora sme v procese výbere dôležitých čŕt eliminovali korelované črty. Ich zoznam je nasledovný:

- NT TFIDF,
- hierarchia_kategorii,
- rovnaka_kategoria,
- NT pomer_dlzky_bez_stop_slov,
- NT bigramy,
- NT bigramy_pomer,
- NT podstatne_mena,
- NT podstatne_mena_pomer,
- NT podstatne_mena_pomer_text_viac,
- NT slovesa_pomer,
- NT slovesa_pomer_text_menej,
- NT slovesa_pomer_text_viac,
- T podstatne_mena_pomer,
- T slovesa_pomer,
- N podstatne_mena,
- N podstatne_mena_pomer,
- N podstatne_mena_pomer_text_menej,
- NT GloVe.

**Tabuľka D.1:** Dôležitosť čŕt (stĺpec Dôl.) pri identifikovaní párov podobných otázok podľa RFE.

| Črta | Dôl. | Črta | Dôl. |
|---|---|---|---|
| NT TFIDF | 21,44 | T GloVe | 6,01 |
| NT podstatne_mena_pomer_text_menej | 14,27 | N podstatne_mena | 5,98 |
| T TFIDF | 12,36 | **cas_kategoria** | 5,81 |
| NT podstatne_mena_pomer_text_viac | 12,15 | N unikatne_slova_pomer | 5,80 |
| NT slovesa_pomer | 11,44 | T bigramy_pomer | 5,76 |
| NT bigramy_pomer | 11,43 | NT pomer_stop_slov | 5,74 |
| N TFIDF | 10,28 | T pomer_unikatnych_stop_slov | 5,54 |
| T GloVe | 10,11 | NT unikatne_slova | 5,37 |
| NT GloVe | 9,47 | OV pomer_stop_slov | 5,24 |
| T podstatne_mena_pomer_text_menej | 9,41 | T unikatne_slova_pomer | 5,22 |
| **rovnaka_kategoria_edx** | 9,01 | OV GloVe | 5,20 |
| NT slovesa_pomer_text_menej | 8,95 | NT pomer_dlzky_bez_stop_slov | 5,19 |
| NT slovesa_pomer_text_viac | 8,69 | OV slovesa_pomer_text_viac | 5,16 |
| T slovesa_pomer_text_menej | 7,93 | OV pomer_dlzky | 5,11 |
| NT slovesa_pomer | 7,87 | NT podstatne_mena | 5,10 |
| N podstatne_mena_pomer_text_viac | 7,70 | T unikatne_slova | 5,04 |
| NT bigramy | 7,61 | NT pomer_dlzky | 5,02 |
| N slovesa_pomer | 7,59 | T podstatne_mena | 5,00 |
| T slovesa_pomer | 7,56 | NT pomer_stop_slov | 4,84 |
| NT unikatne_slova_pomer | 7,45 | T pomer_dlzky_bez_stop_slov | 4,82 |
| T slovesa_pomer_text_viac | 7,11 | OV unikatne_slova | 4,76 |
| T podstatne_mena_pomer_text_viac | 6,93 | NT slovesa | 4,69 |
| N podstatne_mena_pomer_text_menej | 6,89 | N pomer_dlzky_bez_stop_slov | 4,59 |
| **hierarchia_kategorii** | 6,68 | T slovesa | 4,46 |
| T slovesa_pomer | 6,58 | OV unikatne_slova_pomer | 4,43 |
| rovnaka_kategoria | 6,48 | OV pomer_dlzky_bez_stop_slov | 4,38 |
| T bigramy | 6,46 | OV pomer_unikatnych_stop_slov | 4,28 |
| N bigramy_pomer | 6,22 | T pomer_stop_slov | 4,20 |

## Dôležitosť čŕt vypočítaná pomocou RFE

V tabuľke D.1 uvádzame zoznam čŕt identifikovaných ako najvhodnejšie pomocou RFE aj s ich dôležitosťou. Prvých 45 bolo odporúčaných ako črty, ktoré majú byť použité. V tabuľke sa nachádza len 56 čŕt, pretože sa nám nepodarilo získať dôležitosť ďalších čŕt. Črty špecifické pre doménu vzdelávania sú zvýraznené hrubým písmom.

## Dôležitosť čŕt vypočítaná pomocou GBM

V tabuľke D.2 uvádzame zoznam čŕt identifikovaných ako najvhodnejšie pomocou RFE aj s ich dôležitosťou. V našej práci sme použili črty s dôležitosťou väčšou ako 0,2. Črty špecifické pre doménu vzdelávania sú zvýraznené hrubým písmom.

**Tabuľka D.2:** Dôležitosť čŕt pri identifikovaní párov podobných otázok podľa GBM.

| Črta | Dôl. | Črta | Dôl. |
|---|---|---|---|
| NT TFIDF | 11,85 | T unikatne_slova_pomer | 0,45 |
| NT podstatne_mena_pomer_text_viac | 8,89 | OV pomer_stop_slov | 0,42 |
| N GloVe | 7,13 | OV slovesa_pomer_text_viac | 0,40 |
| N podstatne_mena_pomer_text_viac | 5,84 | rovnaka_kategoria | 0,34 |
| T TFIDF | 3,91 | T verb_intersection_ratio | 0,29 |
| T podstatne_mena_pomer_text_viac | 3,88 | N pomer_unikatnych_stop_slov | 0,29 |
| T podstatne_mena_pomer_text_menej | 3,65 | NT verb_intersection_ratio | 0,26 |
| NT podstatne_mena_pomer_text_menej | 3,46 | T pomer_stop_slov | 0,24 |
| N TFIDF | 2,91 | N bigramy_pomer | 0,19 |
| NT bigramy_pomer | 2,70 | OV pomer_dlzky_bez_stop_slov | 0,18 |
| T bigramy | 2,43 | OV podstatne_slova_pomer | 0,14 |
| T slovesa_pomer_text_menej | 2,41 | NT pomer_otaznikov | 0,13 |
| NT podstatne_slova_pomer | 2,41 | OV pomer_unikatnych_stop_slov | 0,12 |
| NT slovesa_pomer_text_menej | 2,22 | OV TFIDF | 0,12 |
| T pomer_dlzky | 2,21 | OV pomer_dlzky | 0,11 |
| NT GloVe | 1,98 | OV podstatne_mena_pomer_text_menej | 0,10 |
| T bigramy_pomer | 1,89 | OV bigramy | 0,04 |
| T slovesa_pomer_text_viac | 1,78 | T pomer_dlzky_bez_stop_slov | 0,03 |
| NT pomer_unikatnych_stop_slov | 1,58 | OV pomer_unikatnych_stop_slov | 0,03 |
| N pomer_dlzky | 1,33 | OV slovesa_pomer_text_menej | 0,01 |
| NT podstatne_mena | 1,25 | OV podstatne_mena_pomer_text_viac | 0,01 |
| N podstatne_slova_pomer | 1,24 | N opytovacie_zamena_pomer | 0,01 |
| N pomer_unikatnych_stop_slov | 1,21 | N podstatne_mena | 0,00 |
| **cas** | 1,11 | T opytovacie_zamena | 0,00 |
| **hierarchia_kategorii** | 1,09 | **rovnaka_iteracia** | 0,00 |
| N unikatne_slova_pomer | 1,01 | NT slovesa | 0,00 |
| **cas_kategoria** | 1,00 | NT opytovacie_zamena | 0,00 |
| T GloVe | 0,99 | NT opytovacie_zamena_pomer | 0,00 |
| NT pomer_dlzky_bez_stop_slov | 0,91 | T pomer_otaznikov | 0,00 |
| T pomer_unikatnych_stop_slov | 0,87 | T slovesa | 0,00 |
| N pomer_dlzky_bez_stop_slov | 0,86 | T opytovacie_zamena_pomer | 0,00 |
| NT pomer_stop_slov | 0,85 | N pomer_otaznikov | 0,00 |
| NT unikatne_slova_pomer | 0,83 | N bigramy | 0,00 |
| OV unikatne_slova_pomer | 0,81 | N slovesa | 0,00 |
| **rovnaka_kategoria_edx** | 0,80 | N slovesa_pomer | 0,00 |
| NT pomer_unikatnych_stop_slov | 0,79 | N slovesa_pomer_text_menej | 0,00 |
| OV GloVe | 0,74 | N slovesa_pomer_text_viac | 0,00 |
| NT slovesa_pomer_text_viac | 0,73 | N opytovacie_zamena | 0,00 |
| T podstatne_mena | 0,73 | OV pomer_otaznikov | 0,00 |
| NT bigramy | 0,68 | OV bigramy_pomer | 0,00 |
| T podstatne_slova_pomer | 0,63 | OV podstatne_mena | 0,00 |
| T pomer_unikatnych_stop_slov | 0,62 | OV slovesa | 0,00 |
| NT pomer_dlzky | 0,53 | OV slovesa_pomer | 0,00 |
| N podstatne_mena_pomer_text_menej | 0,51 | OV opytovacie_zamena | 0,00 |
| N pomer_stop_slov | 0,47 | OV opytovacie_zamena_pomer | 0,00 |

# Príloha E

# Črty použité pre zoradenie odpovedí

V tejto prílohe uvádzame zoznam použitých čŕt pre zoraďovanie odpovedí. Črty môžeme rozdeliť do dvoch kategórií: 1) črty špecifické pre odpovede a 2) črty spoločné pre jednotlivé časti otázok (nadpis otázky (N), text otázky (T), nadpis+text (NT), časť s opytovacími vetami (OV)). Črty patriace do prvej kategórie sú:

- binárna informácia, či je autor odpovede učiteľ (`ucitel`);
- binárna informácia, či je autor odpovede rovnaký ako autor pôvodnej otázky; (`pytajuci`);
- binárna informácia, či je odpoveď označená ako najlepšia; (`najlepsia`);
- binárna informácia, či odpoveď z rovnakej iterácie kurzu ako nová otázka; (`rovnaka_iteracia`);
- skóre vo forme hlasovania (`skore`);
- dĺžka odpovede (`dlzka`);
- počet otáznikov (`otazniky`);
- počet hypertextových odkazov (`url`);
- počet obrázkov (`obrazky`);
- počet stop slov (`stop_slova`);
- počet unikátnych stop slov (`unikatne_stop_slova`);
- pomer najlepších odpovedí ku všetkým odpovediam používateľa (`pomer_naj_odpovedi`);
- rozdiel času vytvorenia odpovede a pôvodnej otázky (`cas`);

Nasledujúce črty boli aplikované pre všetky časti otázok (`N`, `T`, `NT`, `OV`):

- podobnosť na základe TF-IDF reprezentácie (`TFIDF`);
- pomer dĺžky textov (`pomer_dlzky`);
- pomer dĺžky textov bez stop slov (`pomer_dlzky_bez_stop_slov`);
- počet zhodných unikátnych slov (`unikatne_slova`);

- pomer počtu zhodných unikátnych slov ku väčšiemu počtu unikátnych slov v textoch (`unikatne_slova_pomer`);
- počet zhodných bigramov (`bigramy`);
- pomer počtu zhodných bigramov ku väčšiemu počtu unikátnych bigramov v textoch (`bigramy_pomer`);
- počet zhodných podstatných mien (`podstatne_mena`);
- počet zhodných podstatných mien ku väčšiemu počtu unikátnych podstatných mien v textoch (`podstatne_mena_pomer`);
- počet zhodných podstatných mien ku väčšiemu počtu unikátnych slov v textoch (`podstatne_mena_pomer_text_viac`);
- počet zhodných podstatných mien ku menšiemu počtu unikátnych slov v textoch (`podstatne_mena_pomer_text_menej`);
- počet zhodných slovies (`slovesa`);
- počet zhodných slovies ku väčšiemu počtu unikátnych slovies v textoch (`slovesa_pomer`);
- počet zhodných slovies ku väčšiemu počtu unikátnych slov v textoch (`slovesa_pomer_text_menej`);
- počet zhodných slovies ku menšiemu počtu unikátnych slov v textoch (`slovesa_pomer_text_viac`);

Celkovo tak pracujeme s 73 črtami.

## Dôležitosť čŕt vypočítaná pomocou RFE

V tabuľke E.1 uvádzame zoznam čŕt identifikovaných ako najvhodnejšie pre zoradenie odpovedí z podobných otázok pomocou RFE aj s ich dôležitosťou. V tabuľke E.1 sú črty, identifikované pri použití edukačných čŕt. Prvých 10 bolo odporúčaných ako črty, ktoré majú byť použité. V tabuľke sa nachádza len 15 čŕt, pretože sa nám nepodarilo získať dôležitosť ďalších čŕt. Črty špecifické pre doménu vzdelávania sú zvýraznené hrubým písmom.

V tabuľke E.2 uvádzame zoznam čŕt identifikovaných pomocou RFE, keď edukačné črty neboli uvažované. Prvých 7 bolo odporúčaných ako črty, ktoré majú byť použité. V tabuľke sa nachádza len 11 čŕt, pretože sa nám nepodarilo získať dôležitosť ďalších čŕt. Črty špecifické pre doménu vzdelávania sú zvýraznené hrubým písmom.

## Dôležitosť čŕt vypočítaná pomocou GBM

V tabuľke E.3 uvádzame zoznam čŕt identifikovaných ako najvhodnejšie pomocou GBM aj s ich dôležitosťou. V našej práci sme použili črty s dôležitosťou väčšou ako 0,78. Črty špecifické pre doménu vzdelávania sú zvýraznené hrubým písmom.

**Tabuľka E.1:** Dôležitosť čŕt pre zoraďovanie odpovedí podľa RFE s použitím edukačných čŕt.

| Črta | Dôležitosť |
|---|---|
| Pomer najlepších odpovedí ku všetkým odpovediam používateľa | 13,59 |
| Dĺžka odpovede | 11,24 |
| **Binárna informácia, či je autor odpovede učiteľ** | 10,40 |
| Rozdiel času vytvorenia odpovede a pôvodnej otázky | 9,94 |
| Skóre vo forme hlasovania | 7,15 |
| Počet hypertextových odkazov | 5,98 |
| Počet unikátnych stop slov | 5,92 |
| NT TFIDF | 5,78 |
| Počet otáznikov | 5,58 |
| N pomer_dlzky_bez_stop_slov | 5,52 |
| Počet stop slov | 5,44 |
| T unikatne_slova | 5,26 |
| N pomer_dlzky_bez_stop_slov | 5,00 |
| T TFIDF | 5,00 |
| NT unikatne_slova | 4,98 |

**Tabuľka E.2:** Dôležitosť čŕt pre zoraďovanie odpovedí podľa RFE bez použitia edukačných čŕt.

| Črta | Dôležitosť |
|---|---|
| Pomer najlepších odpovedí ku všetkým odpovediam používateľa | 16,25 |
| Dĺžka odpovede | 11,39 |
| Rozdiel času vytvorenia odpovede a pôvodnej otázky | 9,96 |
| Počet hypertextových odkazov | 7,24 |
| Skóre vo forme hlasovania | 6,89 |
| NT pomer_dlzky_bez_stop_slov | 6,26 |
| Počet otáznikov | 6,21 |
| NT TFIDF | 6,06 |
| Počet unikátnych stop slov | 5,92 |
| Počet stop slov | 5,85 |
| N pomer_dlzky_bez_stop_slov | 5,17 |

**Tabuľka E.3:** Dôležitosť čŕt pre zoraďovanie odpovedí podľa GBM.

| Črta | Dôl. | Črta | Dôl. |
|---|---|---|---|
| cas | 14,28 | OV pomer_dlzky_bez_stop_slov | 0,50 |
| dlzka | 11,17 | OV podstatne_mena_pocet | 0,50 |
| **ucitel** | 6,78 | N unikatne_slova | 0,49 |
| pomer_naj_odpovedi | 6,29 | NT slovesa_pocet | 0,47 |
| T unikatne_slova | 4,12 | N unikatne_slova_pomer | 0,40 |
| N pomer_dlzky | 3,70 | OV slovesa_pomer_text_menej | 0,40 |
| N pomer_dlzky_bez_stop_slov | 3,53 | OV bigrams_pomer | 0,40 |
| skore | 3,41 | T podstatne_mena_pomer | 0,35 |
| NT TFIDF | 3,32 | OV podstatne_mena_pomer | 0,32 |
| stop_slova | 3,11 | OV bigramy | 0,29 |
| T podstatne_mena_pomer_text_menej | 3,07 | T podstatne_mena_pocet | 0,27 |
| NT pomer_dlzky | 2,67 | T slovesa_pocet | 0,27 |
| T pomer_dlzky | 2,31 | NT slovesa_pomer_text_viac | 0,21 |
| T slovesa_pomer_text_viac | 2,14 | OV podstatne_mena_pomer_text_menej | 0,20 |
| NT podstatne_mena_pomer_text_menej | 2,10 | url | 0,20 |
| otazniky | 1,87 | NT slovesa_pomer_text_menej | 0,19 |
| T unikatne_slova_pomer | 1,83 | NT podstatne_mena_pomer_text_viac | 0,15 |
| T slovesa_pomer | 1,54 | OV TFIDF | 0,00 |
| T podstatne_mena_pomer_text_viac | 1,54 | pytajuci | 0,00 |
| T bigrams_pomer | 1,40 | obrazky | 0,00 |
| NT unikatne_slova | 1,38 | unikatne_stop_slova | 0,00 |
| NT podstatne_mena_pomer | 1,26 | najlepsia | 0,00 |
| NT pomer_dlzky_bez_stop_slov | 1,23 | **rovnaka_iteracia** | 0,00 |
| OV pomer_dlzky | 1,06 | NT bigramy | 0,00 |
| OV slovesa_pomer_text_viac | 1,06 | NT podstatne_mena_pocet | 0,00 |
| T slovesa_pomer_text_menej | 0,96 | T bigramy | 0,00 |
| T TFIDF | 0,79 | N bigramy | 0,00 |
| N slovesa_pomer | 0,78 | N bigramy_pomer | 0,00 |
| N TFIDF | 0,69 | N podstatne_mena_pocet | 0,00 |
| OV unikatne_slova_pomer | 0,66 | N podstatne_mena_pomer | 0,00 |
| NT unikatne_slova_pomer | 0,62 | N podstatne_mena_pomer_text_viac | 0,00 |
| N podstatne_mena_pomer_text_menej | 0,61 | N slovesa_pocet | 0,00 |
| OV slovesa_pocet | 0,60 | N slovesa_pomer_text_menej | 0,00 |
| NT bigrams_pomer | 0,59 | N slovesa_pomer_text_viac | 0,00 |
| NT slovesa_pomer | 0,58 | OV podstatne_mena_pomer_text_viac | 0,00 |
| T pomer_dlzky_bez_stop_slov | 0,57 | OV slovesa_pomer | 0,00 |
| OV unikatne_slova | 0,56 | | |

# Príloha F

# Plán práce na riešení projektu

### Plán práce na predmet diplomový projekt II

Po skončení akademického roka budeme pokračovať v budovaní dátovej sady obsahujúcej informácie o MOOC kurzoch, predovšetkým príspevky z diskusných fór a študijné materiály vo forme textového obsahu kurzov a prepisov videa a audia.

V priebehu nasledujúceho semestra implementujeme našu metódu a zrealizujeme offline experiment. V letnom semestri akademického roka 2016/2017 zrealizujeme online experiment a prípadne budeme dolaďovať navrhnutú metódu.

### Zhodnotenie plánu práce na predmet diplomový projekt II

Náš plán sa nám z veľkej časti podarilo splniť. Implementáciu metódy sme zrealizovali takmer kompletne (chýba jedna črta pre algoritmus *učenie sa zoraďovať*). Offline experiment sa nám tiež podarilo plne implementovať, a čiastočne vyhodnotiť. Jediný problém bol s dátami, nakoľko kurz, na ktorom sme offline experiment zrealizovali, mal len jednu iteráciu a druhá začínala až v strede semestra.

### Plán práce na predmet diplomová práca

Existujú dva smery, ktorými môžeme pokračovať v záverečnej práci na projekte. Z tohto dôvodu uvádzame dve tabuľky s alternatívnymi plánmi pre jednotlivé týždne semestra. V tabuľke F.1 uvádzame plán práce v prípade realizovania online experimentu a v tabuľke F.2 v prípade realizovania len offline experimentu. Zvyšné dáta z druhej iterácie kurzu pre potreby offline experimentu získame ešte pred začatím letného semestra.

### Zhodnotenie plánu práce na predmet diplomová práca

Na začiatku predmetu diplomová práca sa nám nepodarilo mať potvrdeného inštruktora, ktorý by bol ochotný umožniť nám zrealizovanie online experimentu na jeho kurze. Z tohto dôvodu sme sa zamerali na zrealizovanie offline experimentu, avšak stále sme počas semestra pokračovali v hľadaní vhodného kurzu na zrealizovanie online experimentu počas

**Tabuľka F.1:** Plán práce na letný semester. Verzia s online experimentom.

| Týždeň | Plán práce |
|---|---|
| 1. | Implementácia robota pre automatické odpovedanie nových otázok |
| 2. | Nasadenie robota a vyladenie prípadných problémov |
| 3. | Implementácia ďalších modelov pre výpočet podobnosti textov |
| 4. | Doanotovanie dát z druhej iterácie kurzu |
| 5. | Predbežné vyhodnotenie online experimentu pre potreby IIT.SRC |
| 6.-7. | Offline overenie metódy |
| 8.-10. | Vyhodnotenie online experimentu |
| 11.-12. | Dokončenie diplomovej práce |

**Tabuľka F.2:** Plán práce na letný semester. Verzia bez online experimentu.

| Týždeň | Plán práce |
|---|---|
| 1. | Implementácia nástroja na anotovanie podobnosti otázok |
| 2. | Implementácia nástroja na anotovanie vhodnosti odpovedí |
| 3. | Nasadenie nástroja a prípadné doladenie chýb |
| 4 | Implementácia ďalších modelov pre výpočet podobnosti textov |
| 5.-6. | Experimentovanie s použitými črtami v navrhnutej metóde |
| 7. | Vyhodnotenie dát získaných od anotátorov (napr. zhoda anotátorov) |
| 8.-10. | Offline vyhodnotenie metódy |
| 11.-12. | Dokončenie diplomovej práce |

aspoň niekoľkých týždňov. Takýto kurz sa nám napokon podarilo nájsť, no prvotný pokus o prenesenie natrénovaného modelu na dáta nového kurzu nebol úspešný. Z tohto dôvodu sme pokračovali vo vylepšovaní metódy a offline experimente. Online experiment sa nám napokon nepodarilo zrealizovať.

Alternatívna cesta offline experimentu sa nám podarila zrealizovať v takmer všetkých bodoch uvedených v tabuľke F.2. Nezrealizovali sme nástroj na anotovanie vhodnosti odpovedí. V pláne sme predpokladali získanie oanotovaných dát od anotátorov a my sme sa procesu anotovania neplánovali zúčastniť. V praxi sa však ukázalo, že anotovanie otázok je pre anotátorov časovo náročné a anotátori nemali motiváciu, aby oanotovali dostatočne veľký počet otázok a preto sme museli dáta anotovať aj my. V pláne sme predpokladali, že budeme mať oanotované dáta z dvoch iterácií, avšak časť druhého kurzu sme nestihli oanotovať. Túto časť sme nestihli ani vyhodnotiť v aposteriórnom vyhodnotení úspešnosti metódy.

# Príloha G

# Obsah elektronického média

- `/diplomova-praca/diplomova-praca.pdf`
  - Elektronická verzia dokumentu.
- `/implementacia`
  - `/askalot-metoda.zip`
    - Implementácia navrhnutej metódy vrátane zdrojových súborov systému Askalot.
  - `/askalot-stahovac.zip`
    - Implementácia sťahovača dát zo systému edX vrátane zdrojových súborov systému Askalot.
  - `/nastroj-na-anotovanie.zip`
    - Implementácia nástroja na anotovanie podobnosti otázok.
- `/databaza`
  - `/dump.sql`
    - Databáza so stiahnutými dátami z kurzov zo systému edX.
  - `/nastroj-na-anotovanie.sql`
    - Databáza s dátami získanými pomocou vytvoreného nástroja na anotovanie.
- `/iit-src`
  - `/clanok.pdf`
    - Článok z konferencie IIT.SRC 2017.
  - `/plagat.pdf`
    - Plagát z konferencie IIT.SRC 2017.