FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Pokročilé vyhledávání v datech ze zpravodajských portálů

Pavel Přibáň

Diplomová práce

Inženýrská informatika
Softwarové inženýrství
2016/2017

Vedoucí práce:
doc. Ing. Josef Steinberger, Ph.D

<KIV> KATEDRA INFORMATIKY
A VÝPOČETNÍ TECHNIKY

Motivace

Každý den je vygenerováno obrovské množství zpravodajských dat a vyhledávání informací v těchto datech se v dnešní době stalo pro většinu populace rutinní záležitostí. Většina uživatelů si už ale neuvědomuje, jak složitým problémem je implementace vyhledávání v tak velkém množství zpravodajských dat.

Existuje řada technologií, které implementaci vlastního vyhledávání zjednodušují a urychlují. Tyto technologie obsahují mnoho nastavení a nástrojů pro předzpracování dat (textu), jenž následně významně ovlivňují kvalitu vyhledávání.

Úvod a cíle práce

MediaGist je online systém pro kroslinguální analýzu agregovaných zpráv a komentářů založený převážně na technologiích sumarizace a analýze sentimentu. Systém MediaGist vytváří souhrny zpravodajských článků v pěti jazycích (angličtina, čeština, francouzština, němčina a italština).

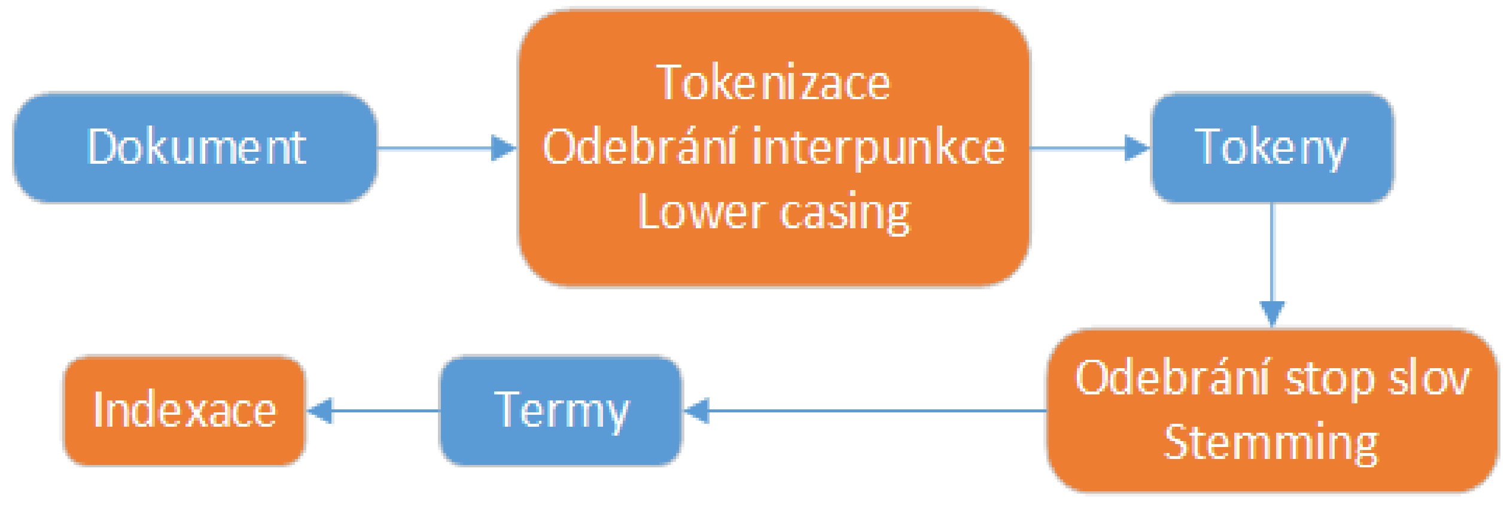
Jedním ze dvou hlavních cílů diplomové práce bylo realizovat jednoduché a rozšířené vyhledávání v těchto datech. Druhým cílem diplomové práce bylo následně otestovat vytvořené vyhledávání a **provést experimenty s předzpracováním textu** s nástroji a nastaveními, které poskytuje vybraná technologie vyhledávání (**Elasticsearch**).

Smyslem experimentů bylo zjistit vliv dostupných nástrojů a nastavení pro předzpracování textu (ve vybrané technologii) na výslednou kvalitu vyhledávání.

Předzpracování textu

Pro vyhledávání v textových datech je klíčové předzpracování textu a jeho indexace. Nejdůležitější částí předzpracování je *stemming*, tzn. redukce různých tvarů slova, ve kterých se dané slovo může vyskytnout, na společný základní tvar.

Mezi další používané postupy při předzpracování textu patří např. tokenizace nebo odebírání stop slov, jež mají také významný vliv na kvalitu vyhledávání. Obecný postup při předzpracování textu před indexací lze vidět na obr. 1.



Obrázek 1: Obecný postup při předzpracování textu před indexací

Realizace

Pro realizaci vyhledávání byla na základě porovnání technologií **Apache Solr** a **Elasticsearch** vybrána technologie **Elasticsearch**. Obě technologie jsou založeny na knihovně Apache Lucene, jejíž nástroje pro předzpracování textu využívá Apache Solr i Elasticsearch.

Testování a experimenty

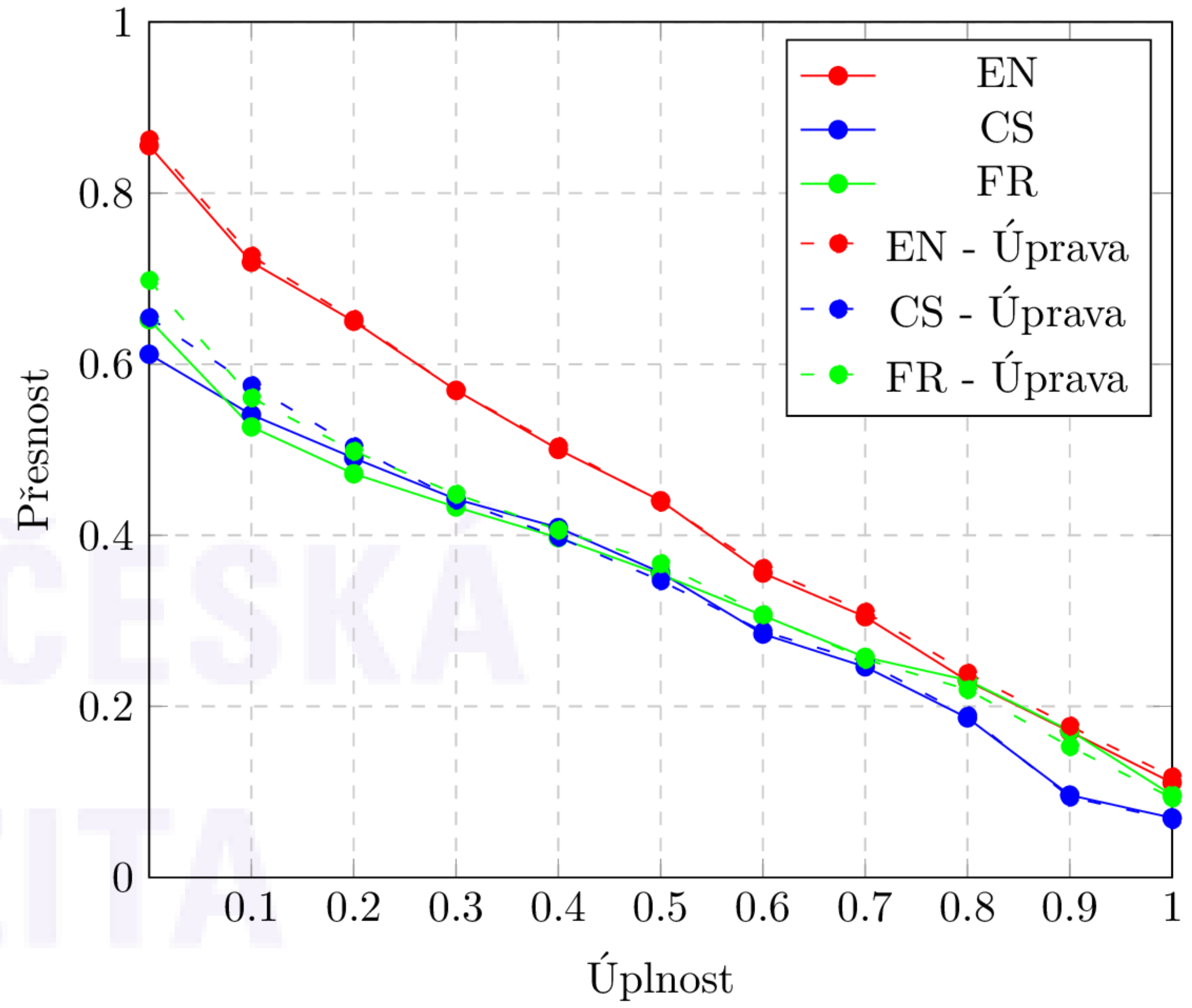
Cílem experimentů a testování bylo zjistit, jaký vliv mají jednotlivé kroky předzpracování textu na kvalitu výsledků implementovaného vyhledávání. Pro porovnání výsledků jednotlivých experimentů byla použita **MAP** (Mean Average Precision) míra a velikosti indexů pro jednotlivé jazyky. Testování probíhalo nad daty z balíčku **CLEF AdHoc - News 2004-2008** a bylo provedeno pro tři jazyky – češtinu, angličtinu a francouzštinu.

Pořadí	Angličtina	Čeština	Franc.
1.	0.4402	0.4242	0.4468
2.	0.4342	0.3586	0.4096
3.	**0.4317**	0.3484	0.4077
4.	0.4274	0.3419	0.3828
5.	0.4057	**0.3267**	0.3794
6. - 9.	–	–	**0.3490**

Tabulka 1: Nejlepší výsledky MAP míry pro původní řešení CLEF AdHoc úloh a dosažené výsledky (tučně) v této práci

Nejprve byla vyhodnocena MAP míra pro základní navržené předzpracování. Celkem bylo provedeno **15 experimentů** a při každém z experimentů bylo předzpracování částečně upraveno (např. změněn stemmer, vynechán některý krok předzpracování apod.).

Na základě experimentů byl upraven původní postup při předzpracování. Hodnoty MAP míry získané touto úpravou (viz tučné výsledky v tab. 1) byly porovnány s řešeními CLEF AdHoc úloh z roku 2007 a 2006.



Obrázek 2: Přesnost/úplnost graf pro testování před a po úpravě předzpracování

Výsledky

Jako nejlepší stemmer pro angličtinu se ukázal stemmer pojmenovaný v nástroji Elasticsearch **light_english**[1] a pro francouzštinu **light_french**. Dále bylo zjištěno, že nejvyšší MAP míry pro český jazyk je dosaženo při indexaci **bigramů** a při **neodstranění stop slov**. Pro anglický jazyk se jeví jako nejlepší řešení použití **trigramů** a pro francouzský použití **trigramů** a **čtyřgramů**. Na obr. 2 je zobrazen přesnost/úplnost graf před (plná křivka) a po (čárkovaná křivka) konečné úpravě předzpracování.

Po úpravě předzpracování došlo k mírnému zlepšení MAP míry, ale také k výraznému snížení velikosti jednotlivých indexů.

Získané výsledky pro jednotlivé jazyky je možné využít při řešení podobných problémů (vyhledávní ve vícejazyčných textech), tzn. použít stejné nebo podobné nástroje a nastavení pro předzpracování textu před indexací, které se ukázalo podle MAP míry jako nejlepší.

---

[1] *Podrobnosti o použitých stemmerech lze nalézt v oficiální dokumentaci Elasticsearch na* `www.elastic.co`