Systém pro extrakci informací z kriminalistických textů

Marek Naggy vedoucí práce: prof. Ing. Karel Ježek, CSc.

KIV KATEDRA INFORMATIKY A VÝPOČETNÍ TECHNIKY

FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

Motivace

20 miliard korun. Na tolik se odhadují škody spojené s 300 tisíci kriminálními činy, které se průměrně v České republice ročně odehrají. Cílem práce bylo prozkoumat možnosti predikce a prevence zločinů na základě existujících nestrukturovaných dat (např. záznamů výslechů nebo sledování) a vytvoření části systému, který by toto umožňoval. Takový systém by mohl pomoci např. polici nebo investigativním reportérům s automatickým zpracováním nestrukturovaných dokumentů a s následným doporučením zájmových osob, míst nebo předmětů, kterým se dále blíže věnovat.

Hlavní cíle

Největší pozornost byla v práci věnována extrakci osobních jmen, ze kterých je následně vytvářena sociální (kriminální) síť osob na základě jejich společného výskytu v dokumentech. Po vytvoření této sítě a simulaci její destabilizace pomáhá systém odpovědět na otázky:

- Která osoba je v sociální síti významná?
- Odstraněním kterých osob lze síť efektivně destabilizovat?
- Pomocí kterých osob síť infiltrovat?
- Která spojení jsou významná a měla by být monitorována?

Pro zodpovězení těchto otázek byl představen model bezškálové sítě, vhodný způsob její destabilizace a metody výpočtu centralit, které umožňují určení významných uzlů v síti. Jejich postupným odstraňováním na základě centralit je simulováno zatčení nebo výslech podezřelých osob a je sledován vliv odebírání uzlů na strukturu této sítě.

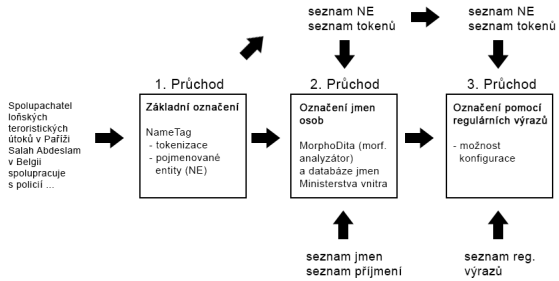
Dále je v práci představeno určení nejvýznamnějších entit, které se v textech vyskytují nebo detekce často se spolu vyskytujících osob. A jelikož zkoumané dokumenty obsahují množství osobních dat, byl též vytvořen anonymizátor pojmenovaných entit pro volné experimentování nepověřenými osobami.

Workflow systému

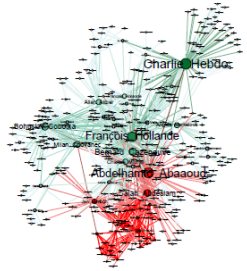
Data: Téměř 200 novinových článků o teroristických útocích v Paříži a Bruselu ze zpravodajského serveru iDnes. Rovněž testováno na 200 reálných dokumentech PČR.

iDNES.cz

Rozpoznání pojmenovaných entit: Rozpoznání jmen osob, míst, předmětů, rodných čísel, SPZ, dat apod. Trojfázový průchod.



Tvorba sociální sítě: Z rozpoznaných osob je na základě společného výskytu v dokumentu vytvořena sociální síť. Při její tvorbě se zohledňuje počet výskytů a koreferencí v dokumentu.



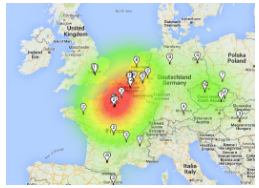
Výpočet centralit a destabilizace sítě: určení významných osob na základě centrality. Simulace a sledování efektivity jejich odebrání (zatčení/výslechu). Optimalizace destabilizace sítě (rozpad na komponenty).

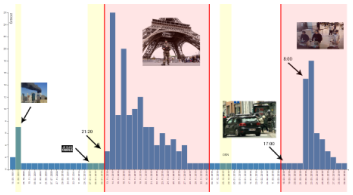
Anonymizace pojmenovaných entit: volitelná anonymizace osobních údajů zachovávající morfologii. Pověřené osoby mohou pracovat i s neanonymizovanými údaji. Anonymizace generované sociální sítě.

Zvýraznění pojmenovaných entit: z dokumentů jsou generovány HTML výstupy obsahující zvýrazněné entity. Na tyto soubory je pak odkazováno z dalších souborů pro usnadnění dalších analýz.

upozorňovaly také na další účastníky teroristických útoků včetně bratrů **Brahima** a **Salaha Abdeslamových** či **Mohameda Abriniho** . S odvoláním na místní úřady o tom informovala agentura **AFP** . Kromě starostky Molenbeeku **Françoise Schepmansové** dostali seznam radikálů i místní policisté . **Abaaoud** byl podle seznamu v té době v **Syrii** , **Abdeslamové** byli součástí „ islamistického hnutí “ a Abrini se ze **Syrie** do **Belgie** údajně vrátil . Tajné služby seznam zřídily po odhalení teroristické buňky ve GEOGRAPHICAL_CITY . Podobné seznamy zaslaly údajně i jiným belgickým obcím . Cílem bylo zajistit lidi radikály vyšší dohled a případně jim zabránit odnětím pasu v odchodu do **Syrie** . Policie čtvrti **Molenbeek** , která zaměstnává na **100 000** obyvatel asi stovku policistů , nemá podle starostky

Určení prominentních komunit: často se spolu vyskytující osoby. **Prostorová a časová analýza:** tvorba tepelné mapy a grafů s frekventovanými daty.





Výsledky

Pomocí simulace destabilizace sociální sítě bylo ukázáno, na které osoby se zaměřit, aby se v síti efektivně znemožnila spolupráce. Oproti náhodnému odebírání osob se při cíleném ušetřilo 69% nákladů při stejné míře destabilizace.

Rozpoznání osob dosahuje pokrytí 84% s přesností 75%. Navzdory vysokému výskytu politiků a vyšetřovatelů v testovacích datech se podařilo určit osoby, které dle dostupných informacích v těchto událostech byly hlavními účastníky nebo je významně propojují, jako jsou např. S. Abdeslam, A. Abaaoud, či N. Laachraoui.