FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2017

Pokročilé vyhledávání v datech ze zpravodajských portálů

Pavel Přibáň[1]

1 Úvod

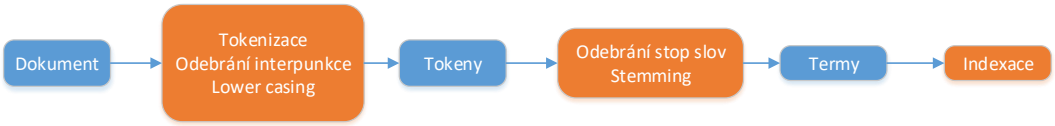
Každý den je na internetu vygenerováno obrovské množství dat a vyhledávání informací v těchto datech se v dnešní době stalo pro většinu populace rutinní záležitostí. Většina uživatelů si už ale neuvědomuje, jak složitým problémem je vyhledávání v tak velkém množství dat.

Systém MediaGist, viz Steinberger (2016), pravidelně seskupuje nové články v několika jazycích ze zpravodajských portálů do clusterů (skupina článků s podobným či stejným tématem) a vytváří z nich souhrny obsahu. Ve vytvořených souhrnech ale není možné vyhledávat. Cílem práce bylo implementovat vyhledávání v datech systému MediaGist a otestovat úspěšnost vytvořeného vyhledávání.

2 Vytvořené řešení

Pro implementaci vyhledávání byl použit nástroj **Elasticsearch**, který byl vybrán na základě porovnání s **Apache Solr**. Vytvořené řešení obsahuje fulltextové a pokročilé textové vyhledávání.

Pro vyhledávání v textových datech je klíčové předzpracování textu a jeho indexace. Pro předzpracování textu byly použity standardní postupy (odebrání stop slov, stemming aj.), viz obr. 1. Důležitou částí řešení byl také návrh indexů pro jednotlivé jazyky, který bylo potřeba vytvořit tak, aby bylo možné zadat dotaz, který bude vykonán nad všemi jazyky najednou.



Obrázek 1: Znázornění postupu předzpracování dokumentů před indexací

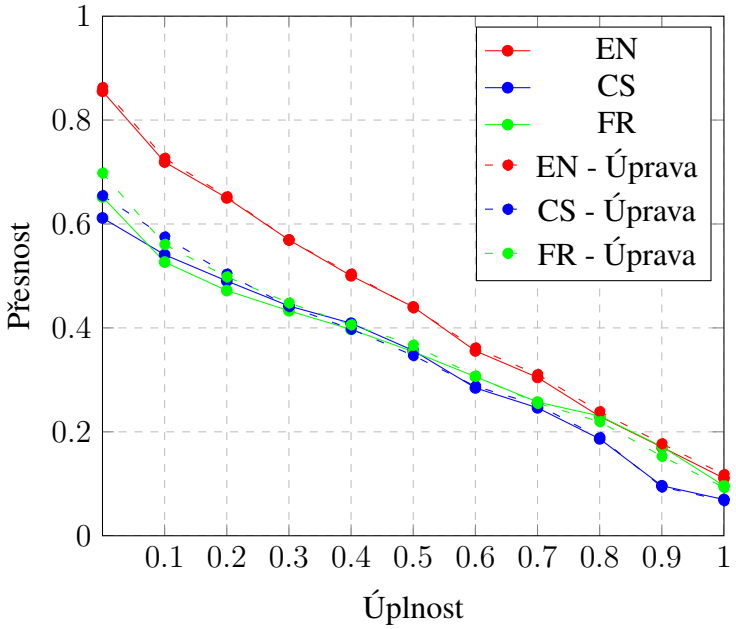
3 Testování a experimenty

Cílem testování bylo ověřit úspěšnost vytvořeného vyhledávání a na základě provedených experimentů 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, viz Croft et al. (2010), 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.

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

[1] student navazujícího studijního programu Inženýrská informatika, obor Softwarové inženýrství, e-mail: pribanp@students.zcu.cz

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 Di Nunzio et al. (2008) a roku 2006 Di Nunzio et al. (2006).



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

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 řešení CLEF AdHoc úloh a dosažené výsledky MAP míry (tučně) v této práci

4 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. Porovnání s řešeními z CLEF AdHoc úloh ukazuje, že vytvořené řešení je podle MAP míry téměř na stejné úrovni a nijak výrazně nezaostává.

Literatura

Steinberger, J. (2016). MediaGist: A cross-lingual analyser of aggregated news and commentaries. *ACL*.

Croft, W. B., Metzler, D., Strohmann, T. (2010). *Search engines*. Pearson Education.

Di Nunzio, G. M., Ferro, N., Mandl, T., Peters, C. (2006, September). CLEF 2006: Ad hoc track overview. In *Workshop of the CLEF for European Languages*. Springer Berlin Heidelberg.

Di Nunzio, G., Ferro, N., Mandl, T., Peters, C. (2008). Clef 2007: Ad hoc track overview. *Advances in Multilingual and Multimodal Information Retrieval*, 13-32.

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