Full-textové vyhledávání s podporou porozumění textu dotazu

Adam Mištera[1]

1 Úvod

Cílem práce je prostudovat dostupné metody sémantické reprezentace textu a na základě této studie navrhnout a implementovat full-textové vyhledávání s podporou porozumění textu dotazu, které se následně integruje do předem zvoleného full-textového vyhledávače *Apache Solr*. Výstupem práce bude zhodnocení získaných výsledků, respektive analýza účinnosti daných metod vzhledem ke zvýšení kvality vyhledávání ve vybrané datové kolekci.

2 Sémantická reprezentace

Pro vytvoření sémantické reprezentace textu byly zvoleny dva modely *Word2vec* a *FastText*. **Word2vec** (Mikolov et al. (2013)) je skupina modelů založených na umělých neuronových sítích určených pro vytvoření slovních vektorů (anglicky *word embedding*). Vstupem modelu *Word2vec* je obsáhlý soubor textů vybraného jazyka, jinak také zvaný jazykový korpus. Příkladem vhodného datového korpusu je internetová encyklopedie *Wikipedia*, která v anglické mutaci obsahuje několik milionů různých článků. Proces vytvoření slovních vektorů spočívá v převedení frází, popřípadě slov jazykového korpusu na vektory umístěné ve vektorovém prostoru většinou čítajícím stovky dimenzí. Jednotlivým slovům poté odpovídá vždy jeden vektor z tohoto prostoru. Slovní vektory sdílející podobný či blízký význam jsou zpravidla umístěny blízko sebe. Naproti tomu nesouvisející slova či slova opačného významu jsou ve vektorovém prostoru daleko od sebe.

FastText (Ryan (2016)) je rozšíření modelu *Word2vec* navržené v roce 2016 firmou *Facebook* provozující největší sociální síť na světě. Na rozdíl od modelu *Word2vec*, který považuje každé slovo v korpusu za atomickou jednotku, *FastText* považuje slovo za množinu několika n-gramů. Pojem n-gram je definován jako řetězec prvků (jako například slova či písmena), které se objevují v delší posloupnosti. Například trigramy slova *jablko* jsou *jab*, *abl*, *blk* a *lko*. Výsledný slovní vektor slova *jablko* je poté součtem vektorů jednotlivých n-gramů.

3 Apache Solr

Apache Solr (Grainger a Potter (2014)) je open-source platforma napsaná v jazyce *Java*, která byla navržená a optimalizovaná pro full-textové prohledávání a indexování rozměrných textových dat.

Na rozdíl od většiny běžných databázových systémů, kde je k jednotlivým dokumentům přiřazeno pole slov, která jsou součástí dokumentu, *Apache Solr* používá odlišnou metodu, takzvaný **invertovaný index** (Grainger a Potter (2014)). Invertovaný index ke každému slovu nebo výrazu obsaženému v korpusu mapuje všechny dokumenty, ve kterých je obsažen.

[1] student bakalářského studijního programu Inženýrská informatika, obor Informatika, e-mail: amistera@students.zcu.cz

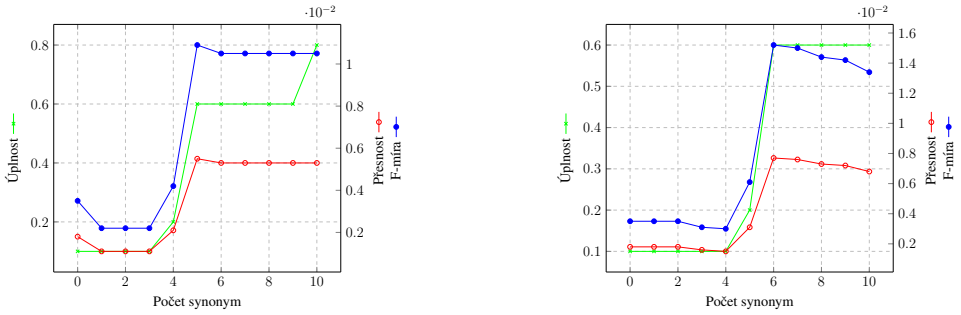
4 Řešení a experimenty

Řešením bylo sestavení rozšířeného dotazu, který vznikne přidáním synonym získaných z natrénovaného modelu k textu dotazu. Uvažujme například, že se uživatel pokouší vyhledat následující dotaz: „Nehody v zaměstnání“. Rozšířený dotaz sestavíme přidáním slov *havárie* a *neštěstí* ke slovu *nehody*. Do druhé části dotazu, konkrétně ke slovu *zaměstnání*, přidáme výrazy *povolání* a *podnikání*.

Funkčnost navrženého řešení byla ověřena na datové kolekci **CLEF** [1] obsahující více než 80 000 novinových článků z periodik Mladá Fronta a Lidové noviny. K těm bylo přiřazeno 50 různých témat, které byla použita pro ověření funkčnosti navrženého řešení. Při vyhodnocení výsledků kvality námi navrženého systému jsme sledovali tři hlavní míry, které jsou v praxi běžně používané především v oblasti získávání informací. Jedná se o *přesnost*, *úplnost* a *F-míru*.

4.1 Téma *Ohrožené druhy*

Na obrázku 1 můžeme vidět výsledky pro téma *425-AH Ohrožené druhy* současně pro oba modely *Word2vec* i *FastText*. Na první pohled je zcela zřejmé, že došlo k velmi výraznému zlepšení *přesnosti*, *úplnosti* i *F-míry*.



(a) *Word2Vec* **(b)** *FastText*

Obrázek 1: Výsledky pro téma *425-AH*

5 Závěr

S použitím modelů *Word2vec* a *FastText* se ve všech experimentech podařilo zvýšit hodnoty *úplnosti* vyhledávání pouze s mírným snížením *přesnosti*. K dalšímu navýšení hodnot *přesnosti* by téměř jistě došlo za předpokladu, že bychom použili pro dané výrazy výhradně synonyma.

Literatura

Graigner, T. a Potter, T. (2014) *Solr in Action*. New York, Shelter Island.

Mikolov, T., Chen, K., Corrado, G. a Dean, J. (2013) *Efficient Estimation of Word Representations in Vector Space*. Avaible from: *arXiv preprint arXiv:1301.3781*.

Ryan, K. J. (2016) *Facebook's New Open Source Software Can Learn 1 Billion Words in 10 Minutes*.

[1] `http://catalog.elda.org/en-us/repository/browse/ELRA-E0036/`