Internetové vyhledávání založené na sumarizaci textů

Bc. Tárik Saleh Salem[1], Doc. Ing. Karel Ježek, CSc.[2]

1 Úvod

Internet je téměř neomezený zdroj informací. Žijeme ale v době informačního přehlcení a hledání požadovaných a v neposlední řadě kvalitních informací se stává čím dál náročnější úlohou. I když standardní internetový vyhledávač vrátí uživateli výsledky vyhledávání, jeho hledání není většinou u konce. Velmi často musí uživatel navštívit nalezené webové stránky, aby posoudil jejich relevantnost nebo našel požadovanou informaci. Webová aplikace ASI, produkt této práce, se snaží omezit nutnost návštěvy další webové stránky, anebo alespoň urychlit výběr relevantní webové stránky.

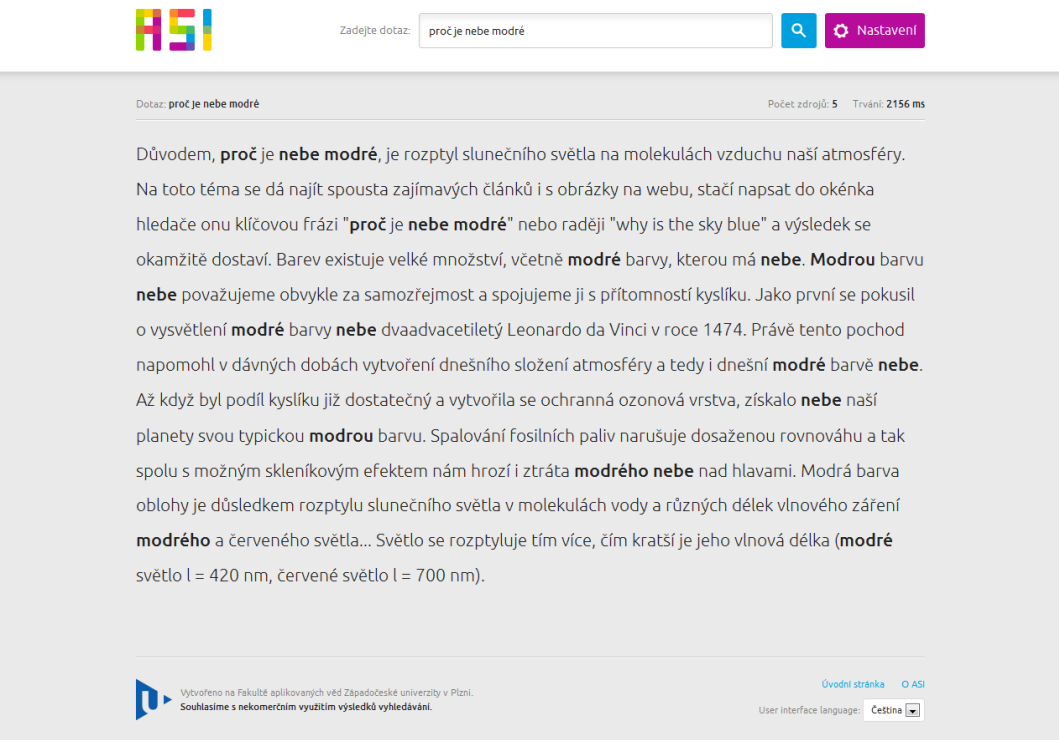
2 Postup vyhledávání

Cílem bylo vytvořit internetový vyhledávač založený na automatické dotazové multidokumentové sumarizaci textů. Celý proces vyhledávání probíhá následovně:

1. Zadání vstupních dat (dotaz, počet zdrojů a délka extraktu).
2. Vyhledání relevantních webových stránek k danému dotazu.
3. Stažení nalezených webových stránek a filtrace textů.
4. Zpracování textů z webových stránek pro vlastní proces sumarizace. Zpracováním textů je na mysli:
   a) tokenizace,
   b) identifikace vět,
   c) lemmatizace,
   d) označení stopslov,
   e) označení klíčových slov dotazu.
5. Multidokumentová na dotazu založená LSA sumarizace a vytvoření souhrnu (extraktu).
6. Předání souhrnu a dalších dat na výstup.

Využívaná sumarizační metoda založená na latentní sémantické analýze (anglicky Latent Semantic Analysis, zkráceně LSA) byla poprvé představena autory Yihong Gong a Xin Liu. Základem latentní sémantické analýzy je singulární rozklad matic (anglicky Singular Value Decomposition, zkráceně SVD), kde vstupní matice reprezentuje texty pro sumarizaci. Ze sémantického hlediska nám singulární rozklad matic umožňuje odhalit latentní (skrytou) sémantickou strukturu textů reprezentovaných maticí.

[1] student navazujícího studijního programu Inženýrská informatika, obor Softwarové inženýrství, e-mail: tarik@students.zcu.cz
[2] vedoucí práce, e-mail: jezek_ka@kiv.zcu.cz



Obrázek 1: Výsledek vyhledávání na dotaz „proč je nebe modré“.

3 Závěr

Výsledkem je webová aplikace ASI (viz obrázek 1) využívající latentní sémantickou analýzu pro sumarizaci textů z webových stránek. Aplikace ASI dokáže sumarizovat webové stránky psané v češtině a angličtině a dovoluje snadné rozšíření o další algoritmy pro vyhledávání.

Literatura

Yihong Gong, Xin Liu, 2001. *Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis.* NEC USA, C & C Research Laboratories.

Josef Steinberger, Karel Ježek, 2004. *Using Latent Semantic Analysis in Text Summarization and Summary Evaluation.* Dept. of Computer Science and Engineering, University of West Bohemia in Plzeň, Czech Republic.

Martin Křišťan, 2007. *Vícedokumentový sumarizátor textů založen na latentní sémantické analýze.* Katedra informatiky a výpočetní techniky, Fakulta aplikovaných věd, Západočeská univerzita v Plzni.