Odstraňování spamu z textových korpusů

Vít Suchomel

Centrum zpracování přirozeného jazyka FI MU

16. 5. 2014

Co je spam?

- Spam v emailu: nevyžádané, obtěžující zprávy.
- Spam v internetových stránkách: hlavním cílem není prezentace kvalitních/původních informací čtenářům.
- Spam v korpusech sestavených z internetových stránek: text neobsahující plynulé, smysluplné, přirozeně navazující věty.

Máme společné cíle? Částečně.

Vyhledávače

- uživatelé: hledající v internetu
- jejich cíl: rychle najít, co potřebují
- požadují: kvalitní stránky obsahující hledané informace

Textové korpusy

- uživatelé: jazykovědci, lexikografové, učitelé a studenti jazyků
- jejich cíl: zjistit, jak se slova a fráze chovají v textu (řeči), časté kontexty, gramatické relace, příklady použití ve větách
- požadují: různorodé texty, přirozené a plynulé v daném jazyce

Rozdíl: Při získávání dat do textového korpusu nevadí libovolné odstavce plynulého přirozeného textu, ani jedná-li se o

- obsahové/odkazové farmy
- rozcestníky (umíme ignorovat při stahování)
- duplicitní texty (umíme odstranit při zpracování)

Příklady webových stránek nevhodných do korpusu

- http://www.metropoliscine.com.ar/archivos/accessory/development/cashadvancedirect.html
- http://www.pay-day-loan.biz/IA/harlan.html
- http://marketing.feedfury.com/content/49344684-cashnetusa-pay-day-loans-and-price-cut-codes.html
- http://www.expert-lender.com/Latest-Mortgage-Rates.aspx

Publikace k tématu

Pomikálek, Jan a Suchomel, Vít. *Efficient Web Crawling for Large Text Corpora*. In *Proceedings of the seventh Web as Corpus Workshop*. Lyon, 2012. s. 39-43.

- vývoj vlastního crawleru zaměřujícího se na textově bohaté webové domény

Baisa, Vít a Suchomel, Vít. *Detecting Spam in Web Corpora*. In *Proceedings of the 6th Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno, 2012. s. 69-76.

- srovnání četností slovních n-gramů v čistějším a spamovějším korpuse
- přesnost 69 % na malé testovací množině

Kilgarriff, Adam a Suchomel, Vít. *Web Spam*. In *Proceedings of the 8th Web as Corpus Workshop*. Lancaster, 2013. s. 46-52.

- shrnutí našeho problému spamu ve webových korpusech

Jak dál

- Zapojit analýzy zpracování přirozeného jazyka: četnost ngramů slov, syntaktickou analýzu.
- Využít seznamy spamových stránek od Seznamu.
- Využít vyhledávače pro sestavování korpusů (převést řešení části našeho problému na vyhledávač).

Výsledek

Klasifikátor textových dokumentů:

- vhodnost pro korpus
- návaznost na nástroje k sestavení a čištění korpusů: crawling, detekce jazyka, oprava kódování, odstraňování ,,boilerplate“ v html, deduplikace podobných odstavců
- rizika: obtížné, spamovací techniky se stále vyvíjí

Dotazy, komentáře