

RSS čtečka s inteligentní filtrací

Projekt z umělé inteligence

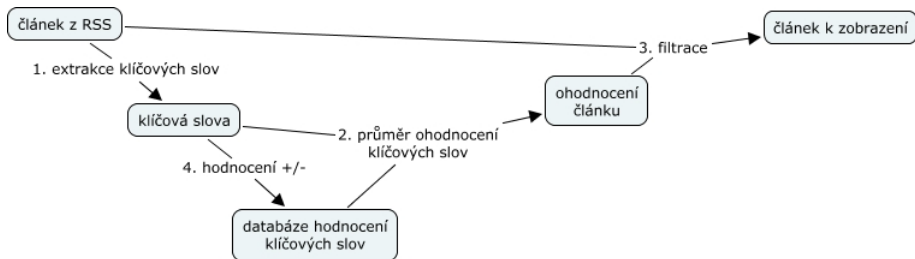
Jiří Procházka

FI MUNI

8. března 2013

Koncept

- Problém moderní informační společnosti je přehlcenost informacemi
- Mnoho novinových článků je pro uživatele nezajímavých, proto by bylo vhodné aby RSS čtečka na základě jeho preferencí články filtrovala





Většina podobných projektů buď filtruje:

- 1 Pomocí ručně zadaných filtrů (např. regulární výrazy)
- 2 Ručním zařazením do taxonomie administrátory RSS agregátorů
- 3 Neznámým proprietárním způsobem

Implementace

Python

- flexibilní mnohostranný jazyk
- nltk - Natural Language Toolkit

Tiny Tiny RSS

- vyspělá RSS čtečka
- projekt koncipovat jako propojený s ní (REST interface)

The screenshot displays the Tiny Tiny RSS web interface. On the left is a sidebar with a tree view of categories: Special (All articles (4206), Fresh articles (94), Starred articles, Published articles, Archived articles, Recently read), Labels (2929), Blogs (13), Comics (1), Development (1), Games (3), Linux (3), Android (1) (LWN.net, Debian SA, LOR (2), Loving the Penguin), News (3446), Reddit (3443), and Science (3). The main content area is titled 'News' and features a list of articles. Each article entry includes a star icon, a title, a snippet of text, and metadata like the author and comment count. The first article is 'Just so everyone knows how fake the "Press Release" pictures of the Nexus 4 are...' by roninb. Other articles include 'Purchase YouTube Views', 'BootMetro: Metro style web framework', 'Building a JSON webservice in R', 'Why "singing" sand dunes hum certain notes', 'Family's house ransacked as a result of a poorly worded craigslist ad', 'Polish drunk driver dies in car accident having a record 22.3% (2.23% of alcohol in blood)', 'APIliner - Documenting APIs with source code examples', 'Google to release "Nexus Keyboard" for Android', 'Kost di Surabaya Barat murah, kost harian surabaya | Info Kost Murah Surabaya, Tarif Surabaya Barat Bisa harian', and 'Causal Link between Solar Variability and Climate Anomalies in East Asia during the Minimum'.

Extrakce klíčových slov

- Je třeba získat strukturovanou sumarizaci článku pro ohodnocení, což budou klíčová slova.
- Protože projekt by měl fungovat na generickém obsahu, nelze předem vydedukovat cílové charakteristiky a podle toho navrhnout specifitější strukturu a způsob extrakce informací.
- S využitím knihovny nltk lze jednoduše najít vhodné fráze jako kolokace (slova která se často vedle sebe vyskytují) na základě jejich distribuce pravděpodobnosti v obecném korpusu (což je postup pro získávání klíčových frází), nebo najít slova z článku jejichž rozdíl frekvence výskytu v článku a korpusu je velký.
Také lze oba postupy kombinovat, a kombinovat extrakci z článku a nadpisu.

Průměr ohodnocení klíčových slov

V databázi se vybere (pokud existuje) ohodnocení extrahovaných klíčových z minulých hodnocení uživatelem a jejich průměr se použije k filtraci.

- Je také možné využít váženého průměru, na základě hodnot z extrakce klíčových slov.
- Využitím ontologie Lexvo, která vychází mimojiné z Wordnetu, lze využít informací o slovech, typu synonyma, hyponym a hypernym (“rudá” hypernymum: “červená”), a tím rozšířit množinu extrahovaných klíčových slov a zvětšit tak šanci, že budou nalezeny klíčová slova dříve hodnocená uživatelem, což urychlí “učení” programu preferencí uživatele.

Filtrace a hodnocení

Filtrace probíhá porovnáním průměru ohodnocení s uživatelem nastavenou prahovou hodnotou, po kterém čtečka článek buď zahodí, přesune do jiné složky, nebo zvýrazní jako doporučný.

Hodnocení by mělo být tlačítka + a - a možné jak u klíčových slov jednotlivě, nebo celého článku, rozdělením hodnocení klíčovým slovům.

Konec prezentace

Děkuji za pozornost