Získavanie metadát o vzťahoch a obsahu na webe

Tomáš Uherčík
Ing. Marián Šimko

Motivácia

- Problém – Automatické získavanie anotácií o internetových zdrojoch (prepojeniach) na webe

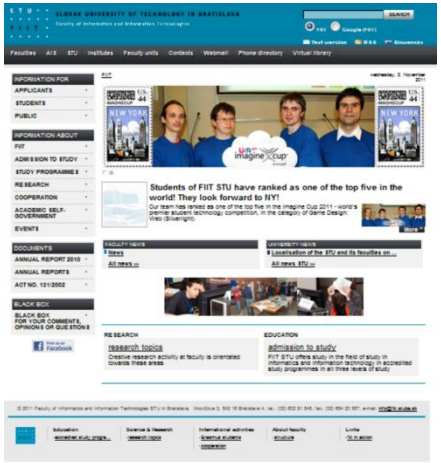
- **Kde sú metadáta ?**
  - **Mikroblogy**
  - **Priamo v obsahu**

Motivácia

- Informácie z mikroblogov:
  - dezinformácie
  - zavádzajúce informácie
  - reklamní boti
  - nálady, pocity, subjektívne názory používateľov
- Z obsahu nezískame niektoré metadáta
  - obrázky
  - videá
  - kvalita (hodnotenie) dobrý? zlý?

Návrh metódy

- Získať kľučové slová z mikroblogov
- Získať kľúčové slová z analýzy zdroja
- Vybrať z nich tie najlepšie



Klúčové slová z mikroblogov

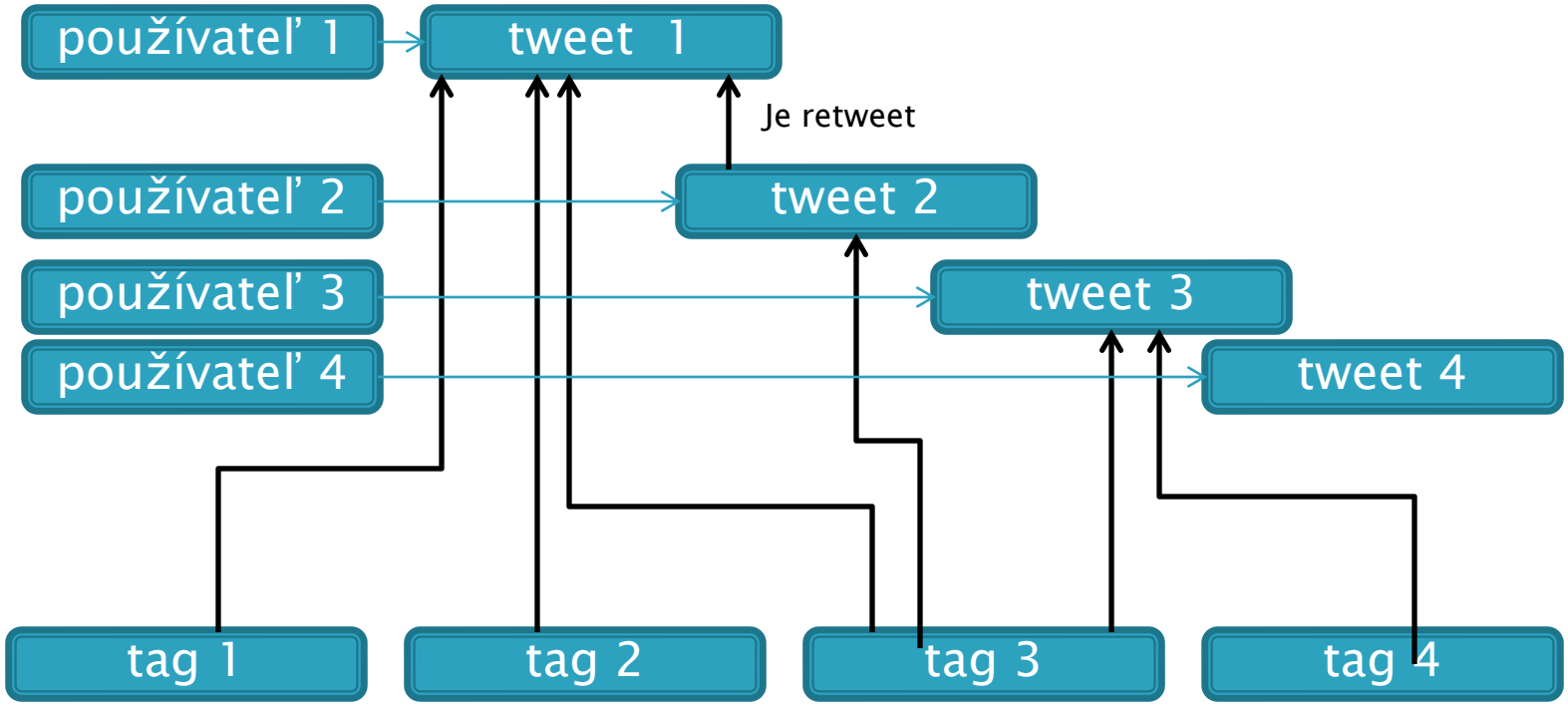
- Relevancia tagu:
  - Ohodnotenie používateľa
  - Tweet, z ktorého pochádza
  - Hashtag
  - Počet výskytov slova,TF-IDF, TextRank

Používatelia mikroblogov

- PageRank
- HITS
- TrustRank
- TunkRank
- NodeRank

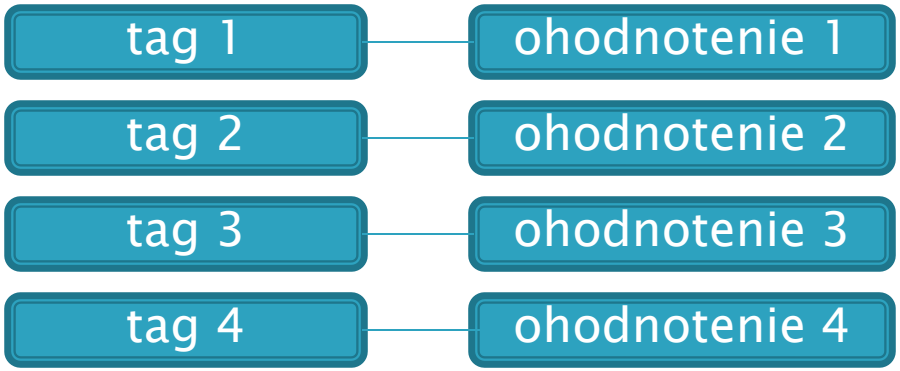
Kľúčové slová z mikroblogov

- Príklad pre vyhľadávanie zdroja z



Klúčové slová z mikroblogov

- Množina tagov spolu s ich ohodnotením



Analýza textu

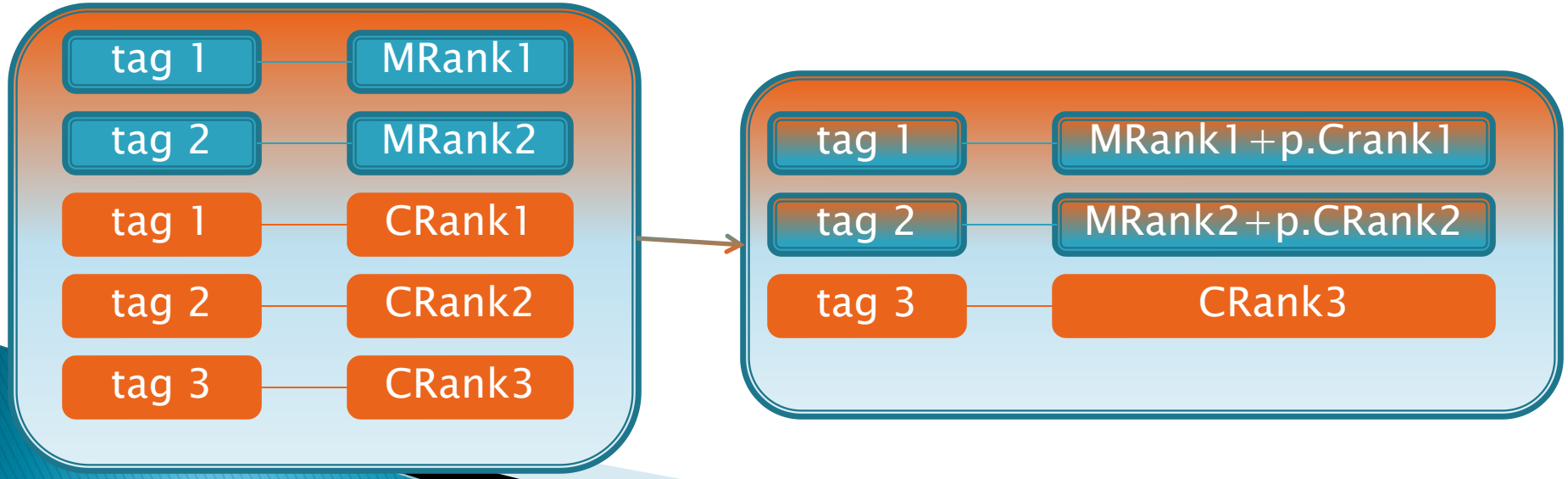
- Získanie množiny tagov spolu s ich rankingom

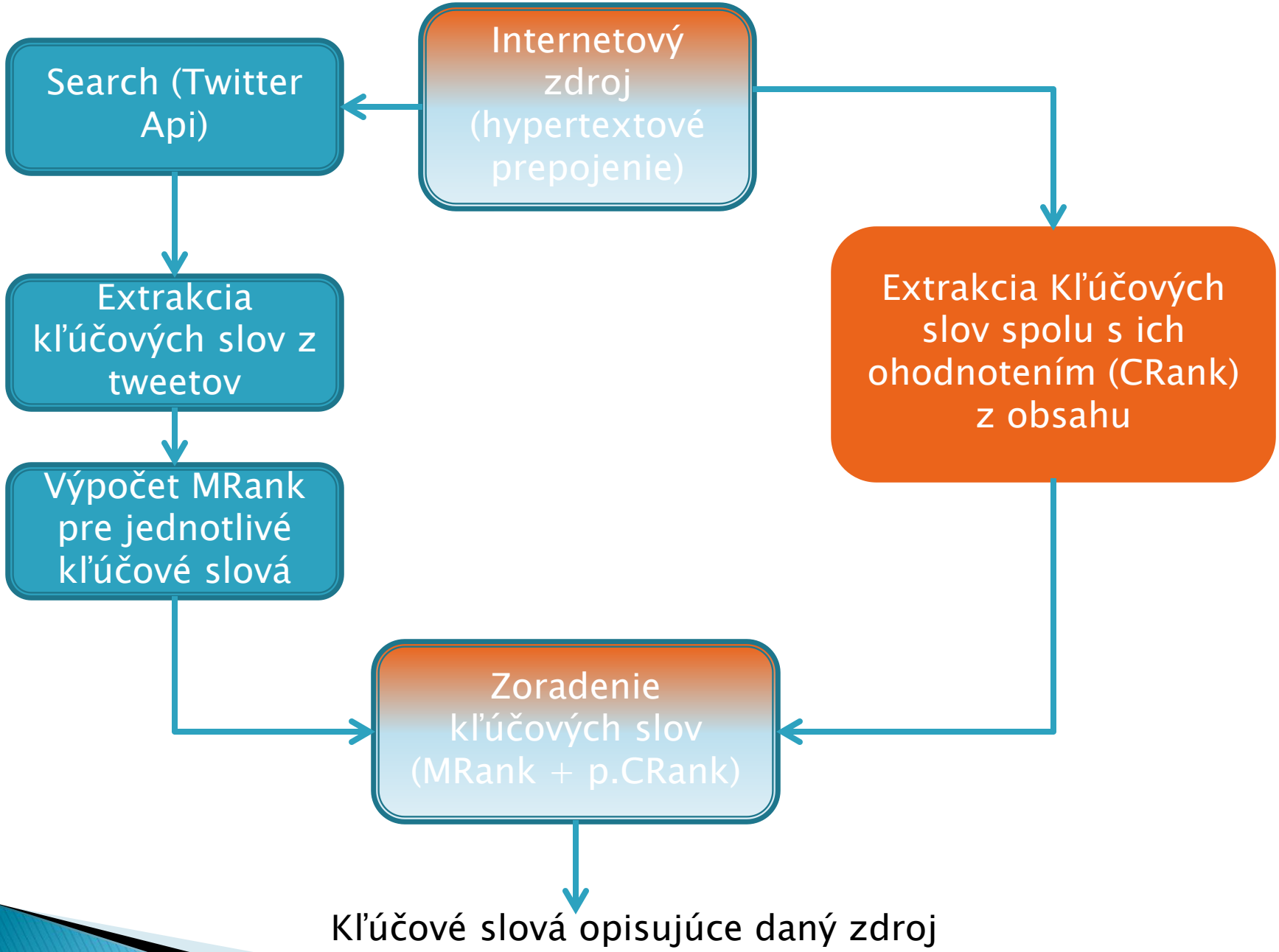
tag 1	ohodnotenie 1
tag 2	ohodnotenie 2
tag 3	ohodnotenie 3

Výsledná množina kľúčových slov

- Zoradená množina z oboch zdrojov
- Zoraďovacie funkcie:
  - Pre mikroblogy MRank(t)
  - Pre obsah(content) CRank(t)

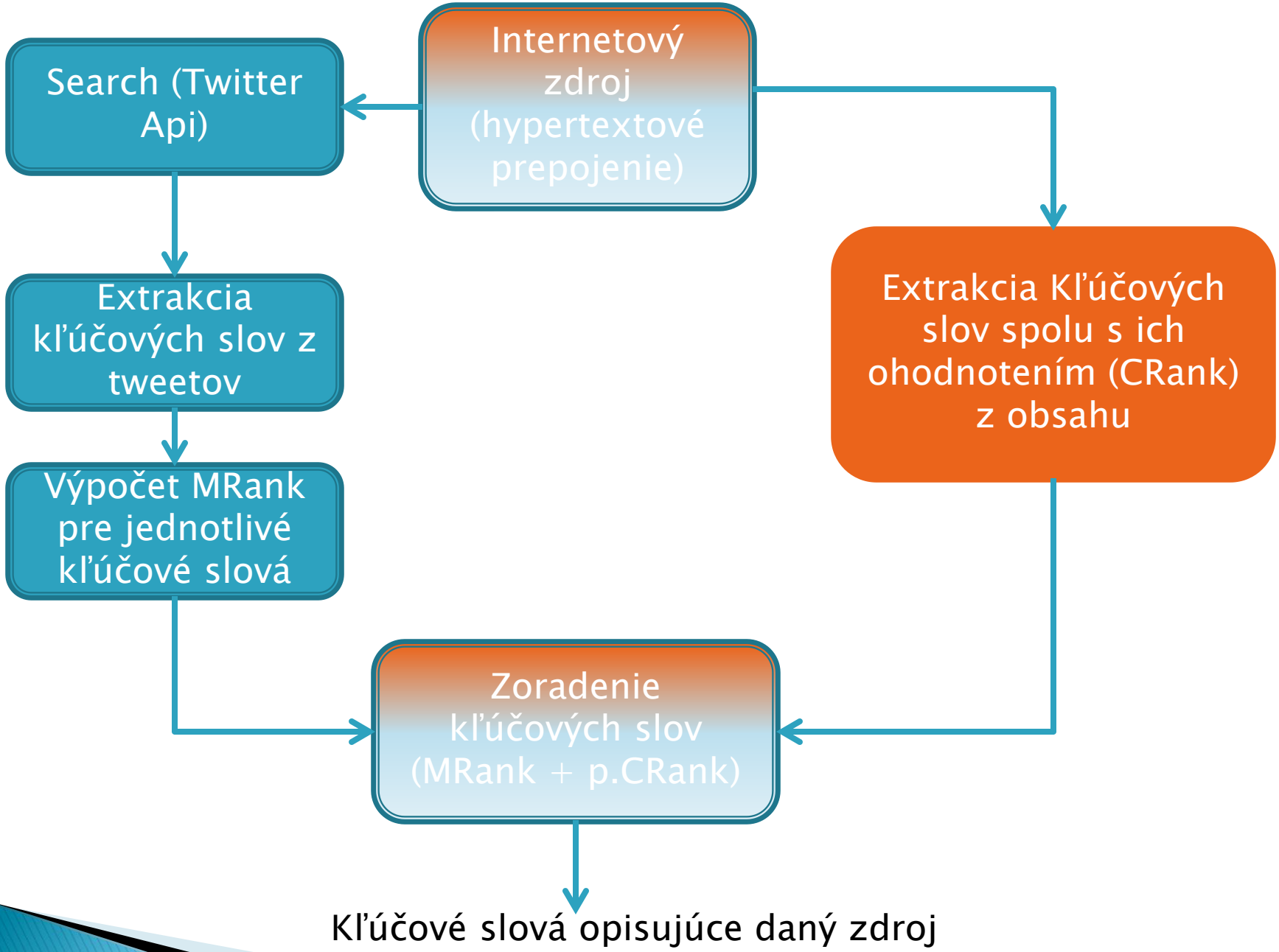
Rank(t) = MRank(t) +p.CRank(t)




Internetový zdroj (hypertextové prepojenie)
Search (Twitter Api)
Extrakcia kľúčových slov z tweetov
Výpočet MRank pre jednotlivé kľúčové slová
Extrakcia Kľúčových slov spolu s ich ohodnotením (CRank) z obsahu
Zoradenie kľúčových slov (MRank + p.CRank)
Kľúčové slová opisujúce daný zdroj

Overenie

- Použitie recent liniek z delicious.com
- Porovnanie kľúčových slov


Internetový zdroj (hypertextové prepojenie)
Search (Twitter Api)
Extrakcia kľúčových slov z tweetov
Výpočet MRank pre jednotlivé kľúčové slová
Extrakcia Kľúčových slov spolu s ich ohodnotením (CRank) z obsahu
Zoradenie kľúčových slov (MRank + p.CRank)
Kľúčové slová opisujúce daný zdroj