Získavanie metadát o vzťahoch a obsahu na webe

Tomáš Uherčík
Ing. Marián Šimko

Motivácia

- Problém

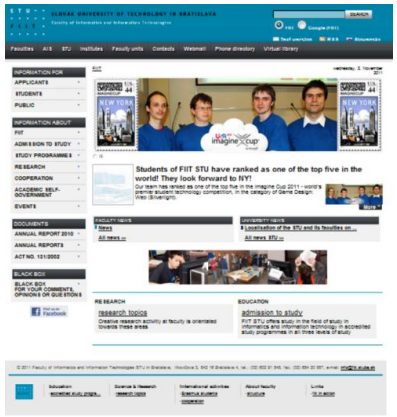
  automatické získavanie metadát o internetových zdrojoch (prepojeniach) na webe

- Kde hľadať metadáta ?
  - Obsah
  - Prepojenia
  - Záznamy o používaní webu
  - Externé zdroje, ktoré obsahujú referenciu na *URL*
    - mikroblogy

Návrh metódy

- Získanie kľučových slov z mikroblogov
- Získanie kľúčových slov z textovej analýzy zdroja
- Výber tých najlepších z nich

+



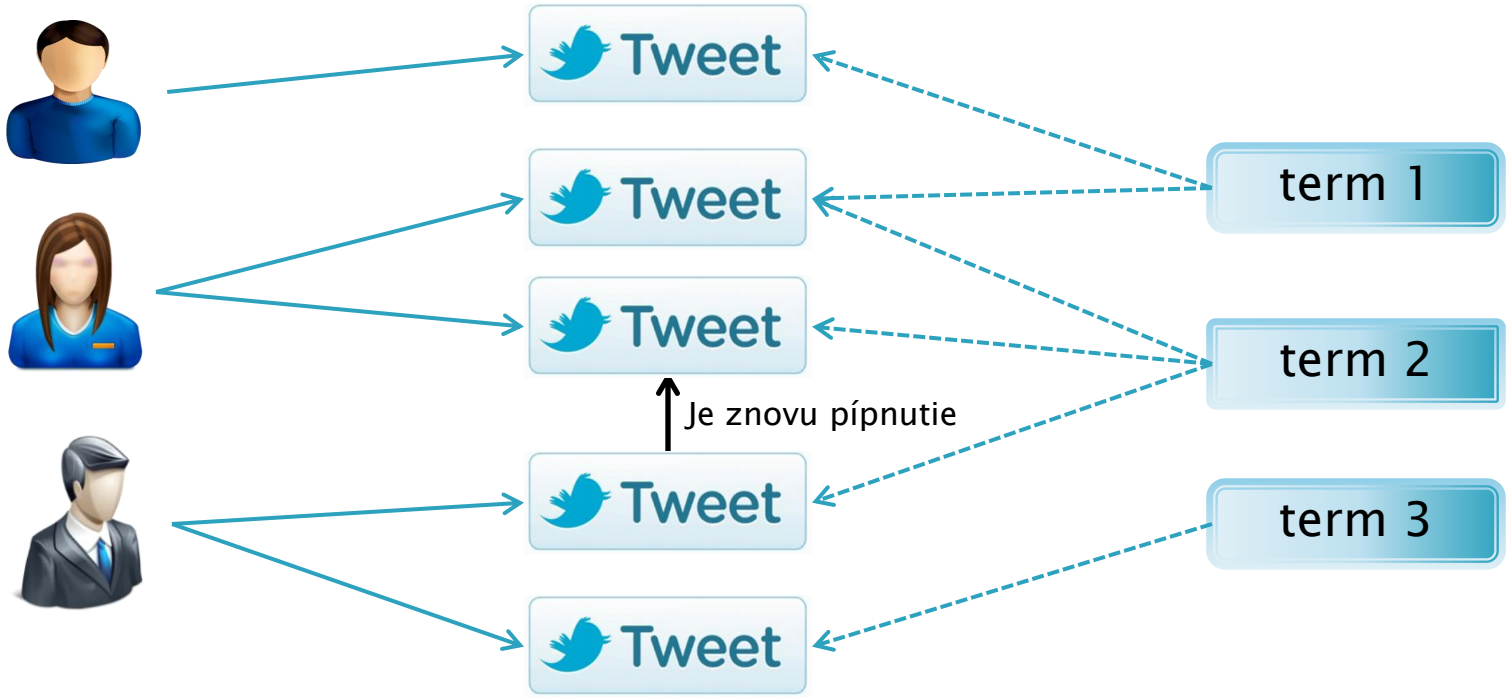
=

Metadáta

- Kľúčové slovo 1
- Kľúčové slovo 2
- Kľúčové slovo 3

Kľúčové slová z mikroblogov

- Vyhľadanie zadanej *URL* v mikroblogu*:*



Klúčové slová z mikroblogov

- Relevancia kľúčového slova:
  - Relevancia v texte pípnutí
    - Text Rank
    - Alchemy API
  - Ohodnotenie používateľa, ktorý pípnutie napísal
    - Tunk Rank

$$TunkRank(X) = \sum_{Y=nasledovníci(X)} \frac{1 + p.TunkRank(Y)}{|nasledovníci(Y)|}$$

kde p je pravdepodobnosť znovu pípnutia

Úpravený Tunk Rank – ARank

- Tunk Rank obohatený o závislosť od periódy publikovania pípnutí:

$$ARank(X) = \sum_{Y=nasledovníci(X)} \frac{1 + \frac{p}{log(T)} . ARank(Y)}{|nasledovníci(Y)|}$$

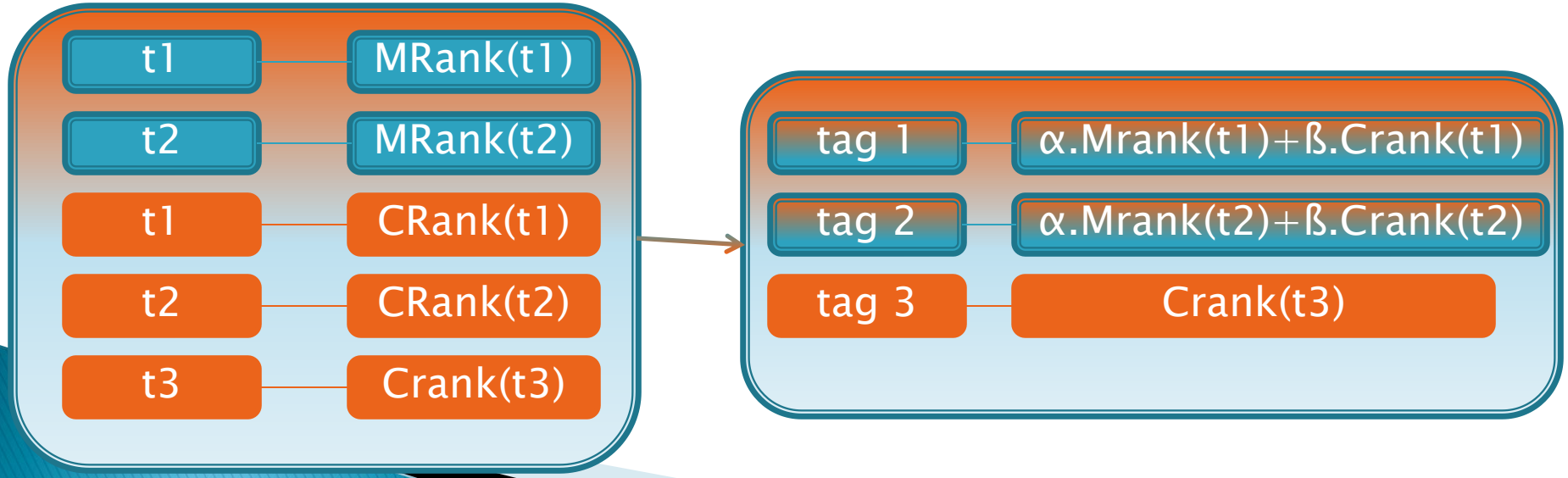
kde:
p – pravdepodobnosť znovu pípnutia
T – perióda ako často používateľ publikuje pípnutia
X – používateľ

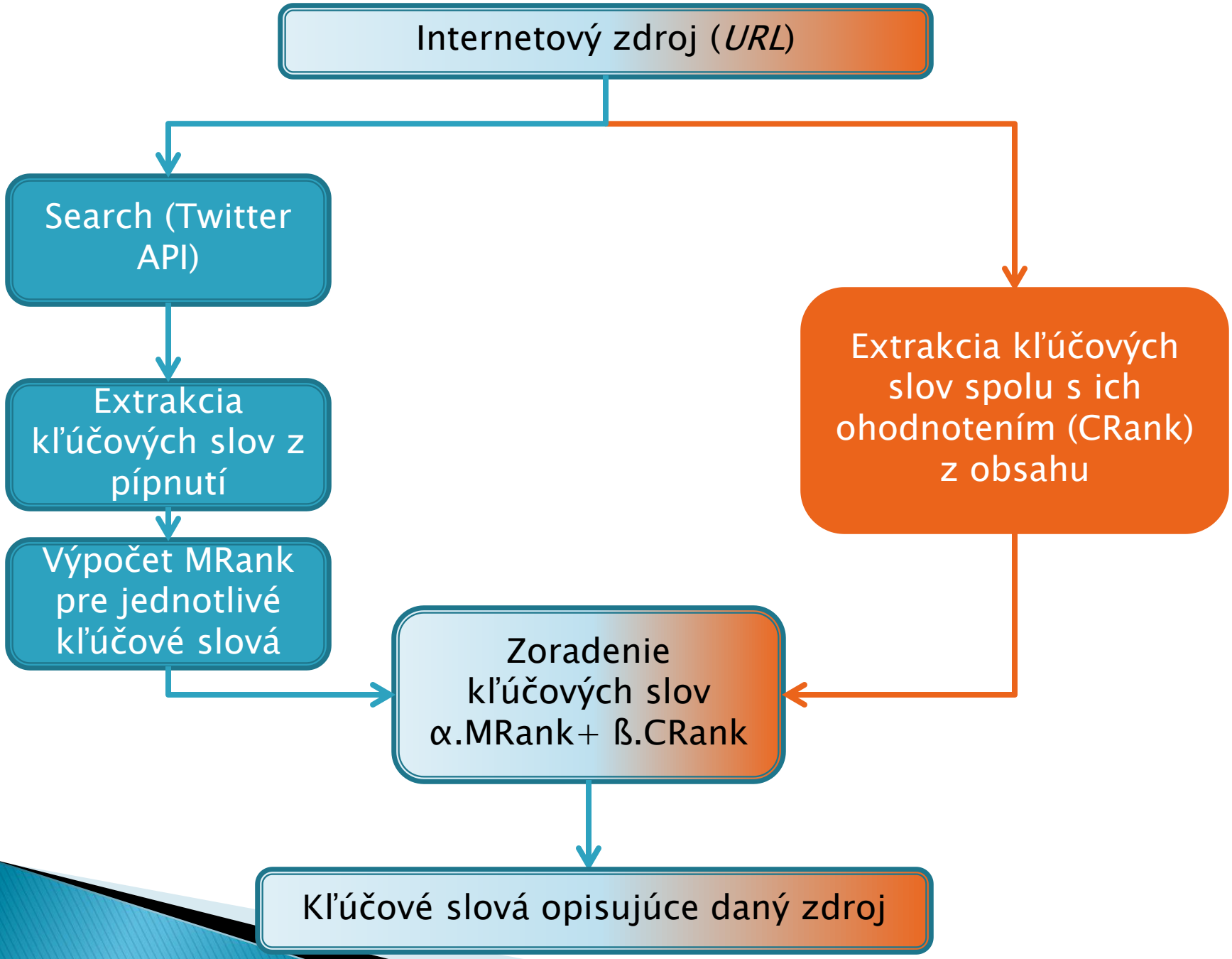
Výsledná množina kľúčových slov

- Zoradená množina z oboch zdrojov:
  - Pre *mikroblogy* MRank(t,ARank(t))
  - Pre *obsah(content)* CRank(t)

$$\mathbf{Rank(t) = \alpha.MRank(t, ARank(t)) + ß.CRank(t)}$$

kde *t* – kľúčové slovo, α, ß – normalizačné konštanty




Internetový zdroj (*URL*)
Search (Twitter API)
Extrakcia kľúčových slov z pípnutí
Výpočet MRank pre jednotlivé kľúčové slová
Extrakcia kľúčových slov spolu s ich ohodnotením (CRank) z obsahu
Zoradenie kľúčových slov α.MRank+ ß.CRank
Kľúčové slová opisujúce daný zdroj

Overenie

- Vyhodnotiť prínos obohatenia kľúčovými slovami z mikroblogov:
  - ľudskí experti
  - porovnanie s kľúčovými slovami z tagovacích systémov
  - porovnanie s kľúčovými slovami z analýzy textu

- Overiť na stálom datasete

Ďalšia práca

- Vystavenie webovej služby na extrakciu kľúčových slov o *URL*
  - Možnosť rozšírenia služby *metall*

- Vylepšenie už implementovaných častí:
  - Sofistikovanejší spôsob porovnávania kľúč. slov
  - Vylepšenie implementácie Text Rank
  - Vyhodnotenie typu obsahu *URL*
    - Automatické nastavenie hodnoty norm. konštánt α a ß

$$\mathbf{Rank(t) = \alpha.MRank(t, ARank(t)) + ß.CRank(t)}$$

Príklad extrakcie kľúčových slov

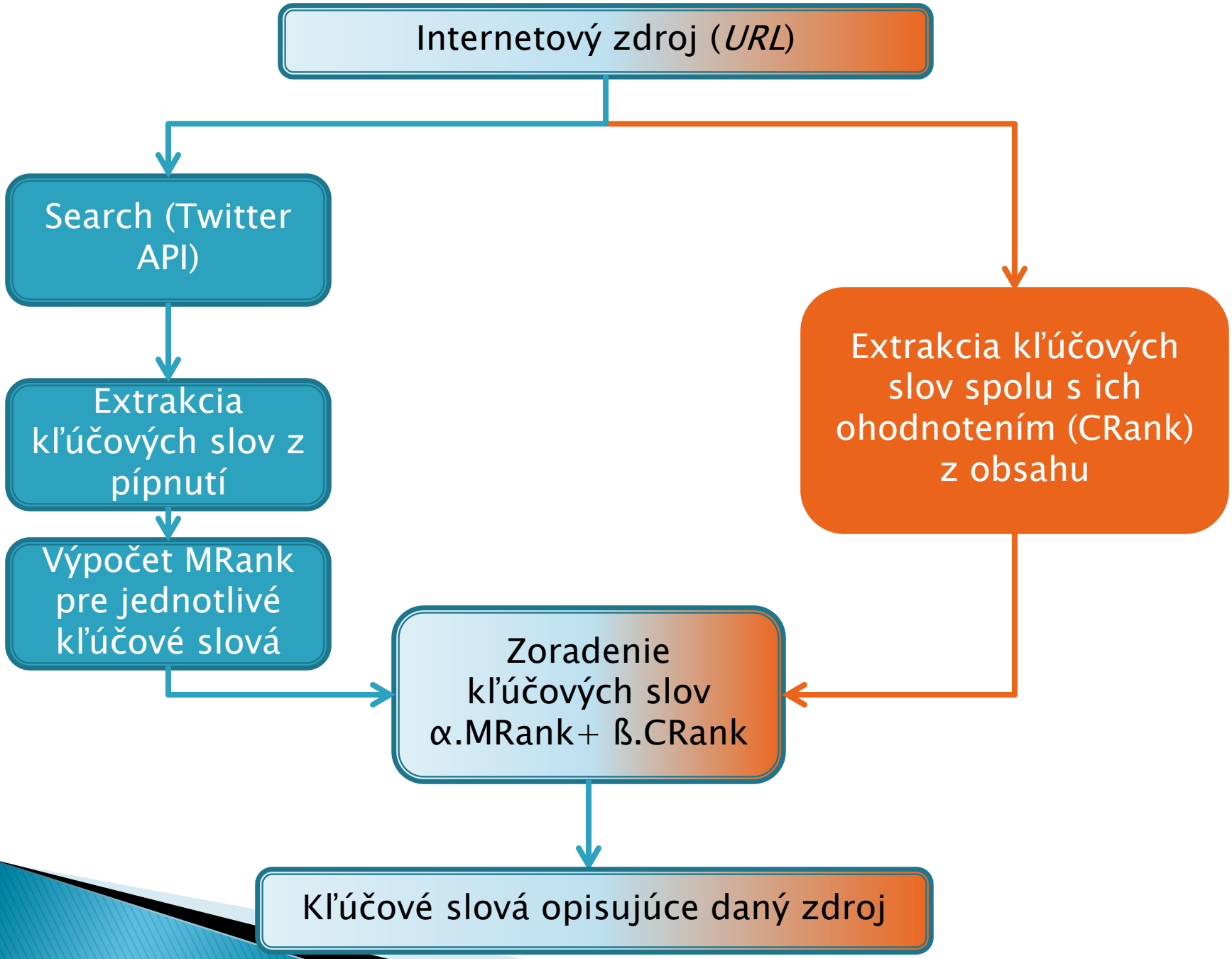
- Pre video: *President Obama Speaks about Insourcing American Jobs*

Twitter	Obsah *URL*	Zjednotené
american jobs	House business leaders	President Obama
insourcing	President Obama	insourcing
President Obama	America	american jobs
		house business leaders
		America

Príklad extrakcie kľúčových slov

- Pre video: Santigold L.E.S. Artistes

Twitter	Obsah *URL*	Zjednotené
santigold	Nima Nourizadeh	santigold
Youtube video		Youtube video
L.E.S. Artistes		L.E.S. Artistes
		Nima Nourizadeh


Internetový zdroj (*URL*)
Search (Twitter API)
Extrakcia kľúčových slov z pípnutí
Výpočet MRank pre jednotlivé kľúčové slová
Extrakcia kľúčových slov spolu s ich ohodnotením (CRank) z obsahu
Zoradenie kľúčových slov α.MRank+ ß.CRank
Kľúčové slová opisujúce daný zdroj