Analýza a vizualizace dat ze sociálních médií

Dalibor Toth

Sociální média

- Nejen sociální sítě
- Diskuzní fóra
- Blogy
- Zprávy a jejich komentáře

- Zajímavá data
  - Obsah příspěvků, článků
  - Vazby mezi účastníky
  - Vzory chování

Využití?

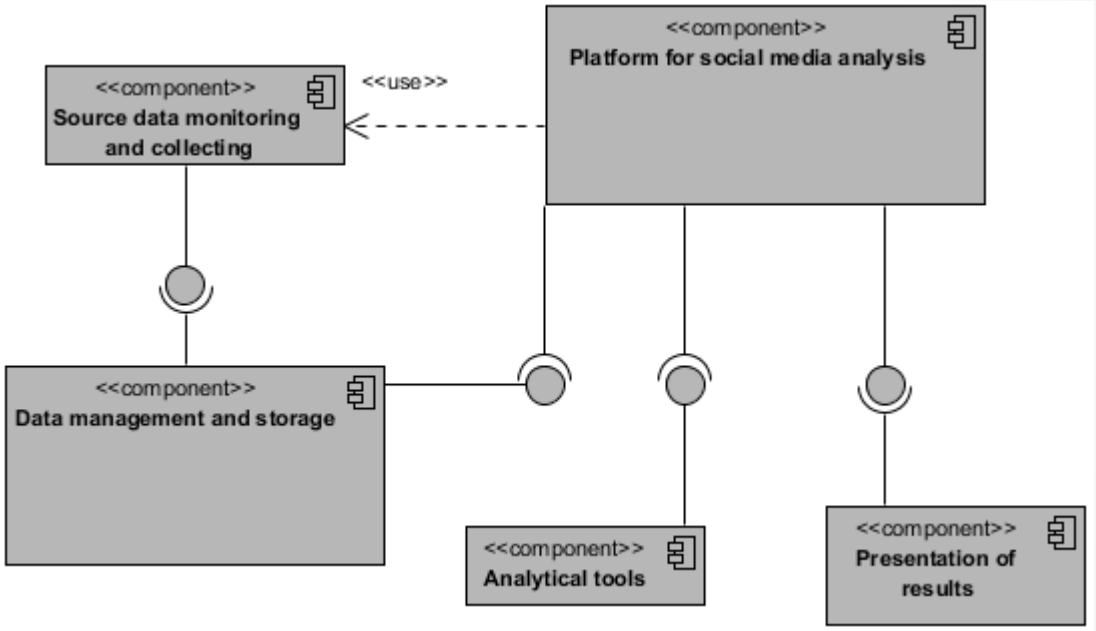
- Marketing
  - Upozornění na důležité události
  - Automatizovaná vyhodnocení (jak je hodnocen produkt v diskuzích)
- Bezpečnost
  - Identifikace vzorů podezřelého chování a upozorňování
  - Monitoring podezřelých oblastí (warez fóra apod.)
  - Vyhledávání nebezpečných osob
- Veřejné mínění
  - Odhady dalšího vývoje, např. výsledků voleb (oblíbenost politiků apod.)
- Žurnalismus
  - Rychlé a automatizované upozornění na nové události, trendy

Dekompozice problému

- 4 základní části + mezičlánek

- **Sběr dat**
- **Ukládání dat**
- **Analýza**
- **Vizualizace**

- **Platforma?**



Sběr dat

- Diskuzní fóra
  - Chronologicky řazené příspěvky
  - Vlákna v diskuzi
- Blogy
  - Obsah samotných příspěvků, případně i připojené reakce
  - Různorodé cíle
- Zprávy
  - Delší úvodní text a na něj reagující příspěvky
- Sociální sítě
  - Příspěvky a komentáře
  - Sdílení obsahu
  - Obecné vs. Profesionální
- Intranet

Ukládání dat

- Jak data rozumně ukládat?
- Obrovské množství relativně různorodých dat
- Využití nástrojů poskytujících částečné předzpracování nestrukturovaných dat
  - Indexace textů
  - Fulltextové vyhledávání
  - Možnost už částečně profiltrovat data pro další zpracování
  - **Apache Solr**

Analýza

- Obsah
  - Základ pro další práci s textem
  - Identifikace klíčových pojmů, které se v textu objevují
- Vazby
  - Monitorování vazeb mezi jednotlivci
  - Jak to ukládat? Co všechno evidovat?
- Sentiment
  - Jednotlivé příspěvky
  - Celé diskuze
  - Postoj autora z jeho příspěvků
  - Trend a upozornění na změny
- Identifikace jednotlivce
  - Vzory chování, časové souvislosti příspěvků, překlepy, slovní spojení
  - Identifikace jedné osoby pod různými pseudonymy

Vizualizace

- Výsledky analýz
- Detaily jednotlivých záznamů
- Skládání komplexního obrazu z více pohledů

- Obrazový i textový výstup v závislosti na kontextu

Platforma

- Propojení jednotlivých částí (komponent) do funkčního celku
- Možnost konfigurace systému podle zaměření na cílová data

- Poskytování rozhraní pro další zpracování výsledků
  - Pro uživatele přes vizualizační část
  - Pro další SW, který výstupní data může dále zpracovávat

Kde jsme?

- Analýza obsahu
- Analýza sentimentu

- Návrh platformy - idea

Kudy dál?

- I po dekompozici příliš široký záběr témat

- Možnosti uplatnění částí systému v rámci FI/labu, kde figuruje velké množství dat?
  - Analýza
    - Zpracování interních dat, např. v rámci diskuzí/anket v IS
  - Vizualizace
    - Kypo
    - CEP

Díky za pozornost.

Nějaké otázky?