Eksploracja danych 1

25/27 października 2017

1 Raport I - zakres tematów

Najbliższe 3 spotkania będą skoncentrowane na eksploracji danych udostępnionych w ramach konkursu. Raport, który będzie podsumowaniem tych tematów, powinien zawierać wyniki eksploracji wraz z ich analizą i komentarzem. Powinien on zawierać nie tylko zaproponowane na zajęciach zaganienia, ale również dodatkowe analizy zaproponowane samodzielnie przez studenta.

Do przeprowadzenia analizy można użyć dowolnego oprogramowania. Przykładowe propozycje:

- python: pandas, sklearn
- RapidMiner
- Orange
- R

2 Wczytywanie danych

Podczas wczytywania danych mogą pojawić się problemy wynikające z nieoczekiwanego formatowania danych, przykładowo:

- puste ciągi znaków,
- ograniczniki pól wewnątrz wartości pól,
- nowe linie wewnątrz rekordów danych.

Wczytując dane za pomocą pakietu pandas mogą przydać się:

- read_csv:
  - `https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html`

  - warto między innymi zwrócić uwagę na: delimiter, names, converters,

- indeksowanie:
  - `https://pandas.pydata.org/pandas-docs/stable/indexing.html`

- selekcja:
  - `https://stackoverflow.com/questions/17071871/select-rows-from-a-dataframe-b`

3 Search Phrases

Proszę przeanalizować dane odnośnie zapytań. Można między innymi zaczać od:

- sprawdzenia liczby zapytań dla różnych kategorii oraz przedstawienia tego na wykresie,
- przygotowania wykresu liczby zapytań zależnie od daty oraz jego analizy,
- sprawdzenia czy są wyszukania/kategorie mające inną charakterystykę cyklu liczby wyszukań niż średnie dane,
- poszukania zmian trendów sezonowych (wiosna/lato/jesień/zima, początek roku akademickiego, ferie, święta), oraz pojawiających się nowych trendów (premiery produktów).

Wyniki analizy będą stanowić część raportu pierwszego.