Eksploracja tekstu i wyszukiwanie informacji w mediach społecznościowych

ZADANIA

- Zadanie 1 zajęcia z dn 29.10.2018
- Zadanie 2 zajęcia z dn 05.11.2018
- Zadanie 3 zajęcia z dn 19.11.2018
- Zadanie 4 zajęcia z dn 26.11.2018
- Zadanie 5 zajęcia z dn 03.12.2018
- Zadanie 6 zajęcia z dn 10.12.2018
- Zadanie 7 zajęcia z dn 07.01.2019
- Zadanie 8 zajęcia z dn 14.01.2019

ZADANIE 1

Należy pobrać za pomocą biblioteki gutenbergr książki *20000 mil podmorskiej żeglugi* Juliusza Verne'a oraz *Ulyssesa* Jamesa Joyce'a a następnie wykonać wykresy kolumnowe częstości występowania słów zarówno w przypadku pozostawienia jak i usunięcia słów kluczowych. Skorzystać z wbudowanego w bibliotece **tidytext** zbioru **stop_words**. Aby wykresy były czytelne, ograniczyć się tylko do najczęsciej występujących słów (użyć funkcji **filter()** lub innej). Kod wysłać na adres julian.sienkiewicz@pw.edu.pl (mailto:julian.sienkiewicz@pw.edu.pl), wpisując w temacie "Nazwisko, TEXT, Zad 1". Kilka obrazów z ggplota na jednej stronie można wykonać za pomocą funkcji **grid.arrange()** z pakietu gridExtra.

ZADANIE 2

Wykorzystać wbudowany zbiór **acq** (**data(acq)**, 50 przykładowych dokumentów z Reutersa w formie korpusu) z pakietu **tm** do wykonania analizy podobieństwa tekstów przy wykorzystaniu macierzy termów-dokumentów. Usunąć znaki interpunkcyjne oraz przeprowadzić steming i usunięcie słów funkcyjnych. Policzyć dla każdej pary dokumentów iloczyn skalarny oraz cosinus podobieństwa i wykonać histogramy takich wartości - zarówno dla zwykłego sumowania słów jak i dla transformacji TF-IDF. Sprawdzić, które dwa dokumenty są najbliżej siebie w każdej z metod.

ZADANIE 3

Zadanie polega na porównaniu wykładnikow skalowania w prawie Heapa dla różnych języków. W tym celu należy pobrać za pomocą biblioteki gutenbergr kilka (najlepiej więcej niż 10) pozycji występujących w dwóch językach, w każdym przypadku policzyć wykładnik β w prawie Heapa oraz wyznaczyć korelacje tak otrzymanych serii, wraz z p-wartością. Dodatkowo proszę sprawdzić jakie są wykładniki β w przypadku gdy dokonamy losowego przetasowania słów w książkach.

ZADANIE 4

Dokonać porównania *Skelpów cynamonowych* Brunona Schultza (id=8119) oraz *Sonetów krymskich* Adama Mickiewicza (id = 27081) na podstawie bigramów: wykreślić po 20 najpopularnijeszych bigramów w obu pozycjach jak również sieci słów (samoodzielnie wybrać poziom odcięcia).

ZADANIE 5

Na bazie dowolnej książki otrzymać bigramy, a następnie dla 3 wybranych rzeczowników (najlepiej z jak największą liczbą wystąpień) policzyć wartości emocjonalne słów stojących z lewej strony. Wykonać chmury

słów.

ZADANIE 6

Korzystając ze omawianego zbioru BBC, połącz klasę 1 i -1 w jedną i dokonaj klasyfikacji za pomocą metody SVM (jądro liniowe), porównując klasę subiektywną (1 i -1) z obiektywną (0). Badania przeprowadź dla w miarę podobnych liczb postów z każdej klasy (np. 500 i 500).

ZADANIE 7

Korzystając ze zbioru Gutenberga pobrać po trzy książki Juliusza Verne'a oraz Jane Austen i wykonać model tematyczny LDA: wyszukać po 2 tematy zarówno w książkach Verne'a jak i Austen i zwizualizowac je tak jak w przykładzie 8.15, a następnie połączyć oba zbiory i jeszcze raz wykonać tę analizę. Sprawdzić jaki efekt da rozdzielenie na 3 tematy (w przypadku scalonego zbioru).

ZADANIE 8

Wykonac wykresy takie, jak na Wykładzie 7 (http://www.if.pw.edu.pl/~julas/TEXT/pliki/TEXT7.pdf), slajdy 17 i 18. W tym celu należy zebrać dane dla wybranej linii tramwajowej dla 30-40 punktów, zachowując odstęp 1 minuty pomiędzy kolejnymi pobraniami. Następnie wykreślić położenia, używając Google Map oraz policzyć prędkość tramwaju na kolejnych odcinkach. Odległość pomiędzy kolejnymi punktami można policzyć korzystając z biblioteki **geosphere**.