

Narzędzia

Krzysztof Rudaś, Piotr Smuda, Karolina Wyszyńska

30 marca 2015

Nasze plany

1. Analiza wydźwięku komentarzy w postaci zależności od czasu (gdzie jednostką czasu jest odcinek)
2. Znalezienie najpopularniejszych odcinków ze względu na ilość oraz ze względu na jakość otrzymanych komentarzy
3. Najbardziej wyczekiwany odcinek ze względu na szybkość pojawiania się komentarzy względem czasu emisji
 - 1) Będą potrzebne dodatkowo dane takie jak:
 - data i godzina emisji odcinków
 - data i godziny komentarzy
4. Analiza odbioru poszczególnych/wybranych bohaterów poprzez analizę słów, które pojawiają się w odniesieniu do danego bohatera (kontekst zdaniowy - słowa w obrębie zdania z imieniem bohatera)
5. Wpływ przerw w emisji seriali (np. dzień świąteczny) na odbiór serialu przez widzów

Narzędzia

1. Obróbka tekstu
 - 1) pakiet *tm*
 - `Corpus(VectorSource(x))` - stworzenie korpusu
 - `tm_map` - interfejs do obróbki tekstu
 - `stripWhitespace` - zamiana znaków białych na spacje
 - `content_transformer(tolower)` - zamiana wielkich na małe litery
 - `removeWords` - usuwanie słów (np. stopwordsów)
 - `removePunctuation` - usuwanie znaków interpunkcyjnych
 - `stemDocument` - zamiana słów do słów bazowych
 - 2) pakiet *stringi*
 - `stri_extract_all_words` - wyciągnięcie słów
 - `stri_replace_all_regex` - podmiana słów na inne (np. linki na znak pusty)
 - `stri_detect_regex` - wykrywanie jakiegoś napisu
 - `stri_count_regex` - zliczanie ilości wystąpień danego napisu
2. Słownik
 - 1) jakiś słownik [link](#)
 - prawie 14000 słów
 - skala 0:10 (?)
 - nie wiem (Piotrek) co w nim oznaczają kolumny
 - 2) [AFINN](#)
 - prawie 2500 słów
 - skala -5:5
3. Wizualizacja wyników

- 1) aplikacja Shiny (?)
- 2) pakiet *dygraphs* - szeregi czasowe
 - dygraph - stworzenie interaktywny wykres, coś ala shiny
 - dyRangeSelector - interaktywna oś pod wykresem dla ustalenia interesującego nas czasu
- 3) pakiet *wordcloud* - chmury słów
 - textplot - wykres x,y z nienachodzącymi się słowami
 - wordlayout - położenie słów na wykresie, jeśli byśmy chcieli sami zrobić taki wykres
 - wordcloud - wykres z chmurą słów
- 4) pakiet *ggplot2* ([link2](#), [link3](#)) - wykresy różnego typu
 - qplot - szybki wykres, podobna funkcja do plot
 - ggplot - bardziej rozbudowana (profesjonalna) funkcja do wykresów
 - ggpoints - geometria pozwalająca zrobić wykres punktowy
 - geom_bar - geometria pozwalająca zrobić wykres słupkowy
 - geom_smooth - geometria do pracy z danymi ciągłymi i szukania zależności typu regresyjnego
- 5) pakiet *googleVis* - wykresy różnego typu (w tym interaktywne)
 - plot - bazowa funkcja do wywoływania różnych typów wykresów
 - gvisLineChart - tworzy wykresy krzywych zależnych od zadanych parametrów
 - gvisMotionChart - tworzy wykresy interaktywne zależne od kilku parametrów, w tym możliwość obserwowania zmian w czasie
 - gvisAnnotatedTimeLine - wykres zależności czynnika od czasu, wraz z możliwością umieszczania adnotacji w wybranych punktach wykresu
 - gvisMerge- umieszczenie kilku wykresów w jednym oknie (podobnie jak w Tableau)
 - gvisScatterChart - tworzy wykresy rozproszenia
 - gvisTable - tworzy tabele