WTUM: Przewidywanie "śmieszności" żartu - Plan pracy

Maciej Łukasik, Kajetan Larwa, Anton Yushkevich, Maksim Makaranka

March 2022

1 Kontrukcja pod-modelu określającego podobieństwo dwóch dowcipów

Osoby odpowidzialne:

- Maciej Łukasik
- Kajetan Larwa

1.1 Założenia

Zostanie zbudownych kilka wersji pod-modelu mogącego określić podobieństwo między dwoma żartami. Punktem oparcia dla naszych rozważań jest następujący artykuł: https://medium.com/@adriensieg/text-similarities-da019229c894. Wstępnie rozważamy następujące metody określania podobieństwa między tekstami:

- Jaccard Similarity + lemmatization
- K-means
- Cosine Similarity
- Latent Semantic Indexing
- Word Mover's Distance (jeśli zdążymy)

W przypadku niektórych z podanych metod konieczna jest wcześniejsza wektoryzacja tekstu, tutaj także planujemy sprawdzić kilka różnych metod/pretrenowanych modeli:

- Bag of Words
- Term Frequency - Inverse Document Frequency
- SentenceBERT

- InferSent
- Universal Sentence Encoder

Dokładny podział pracy będzie już na bierząco omawiany w dwuosobowym podzespole (wstępnie Maciej zajmie się pierwszą częścią a Kajetan drugą).

2 Kontrukcja modeli wykorzystujących (lub nie) opracowane metryki

Osoby odpowidzialne:

- Anton Yushkevich
- Maksim Makaranka

2.1 Założenia

Następujące modele zostaną zbudowane i przetestowane:

- Model oparty tylko na opiniach innych użytkowników (model popularnościowy)
- Model oparty tylko na opiniach konkretnego użytkownika (model filtrowania zespołowego)
- Model Hybrydowy (zostanie zaprojektowany później, po tym jak zostaną opracowane pojdenyncze modele)

W początkowym etapie prac możliwe, że członkowie tej podgrupy będą również zaangażowani w poprzedni punkt.

Statistical report

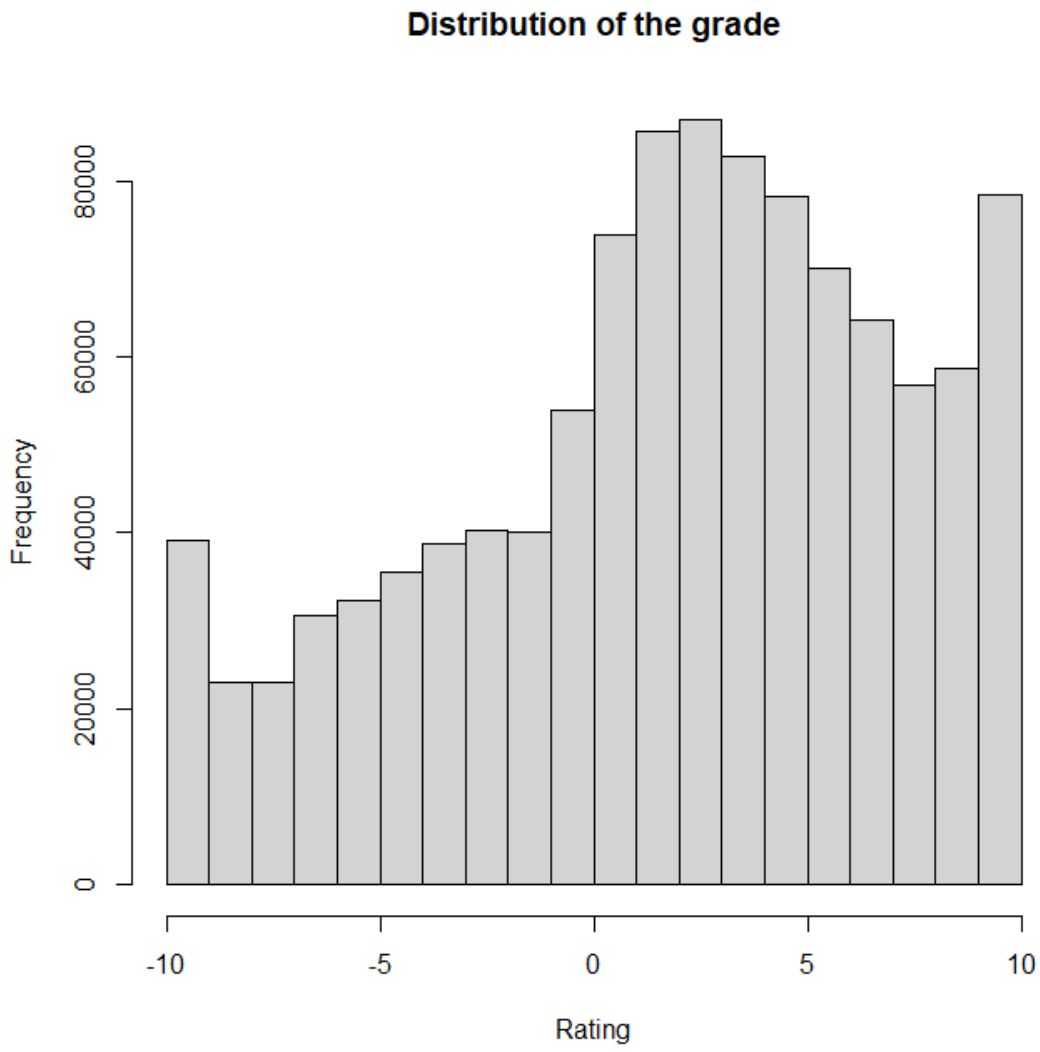
This document is a statistical report on the dataset used in our project.
The dataset contains anonymous ratings(-10 to 10) provided by a total of 41,000 users. Train file contains 1.1 million ratings for 139 jokes.

Here we load the data and create the dataframe.

```
library("ggplot2")
trainSet <- read.csv("train.csv")
df <- data.frame(trainSet)
```

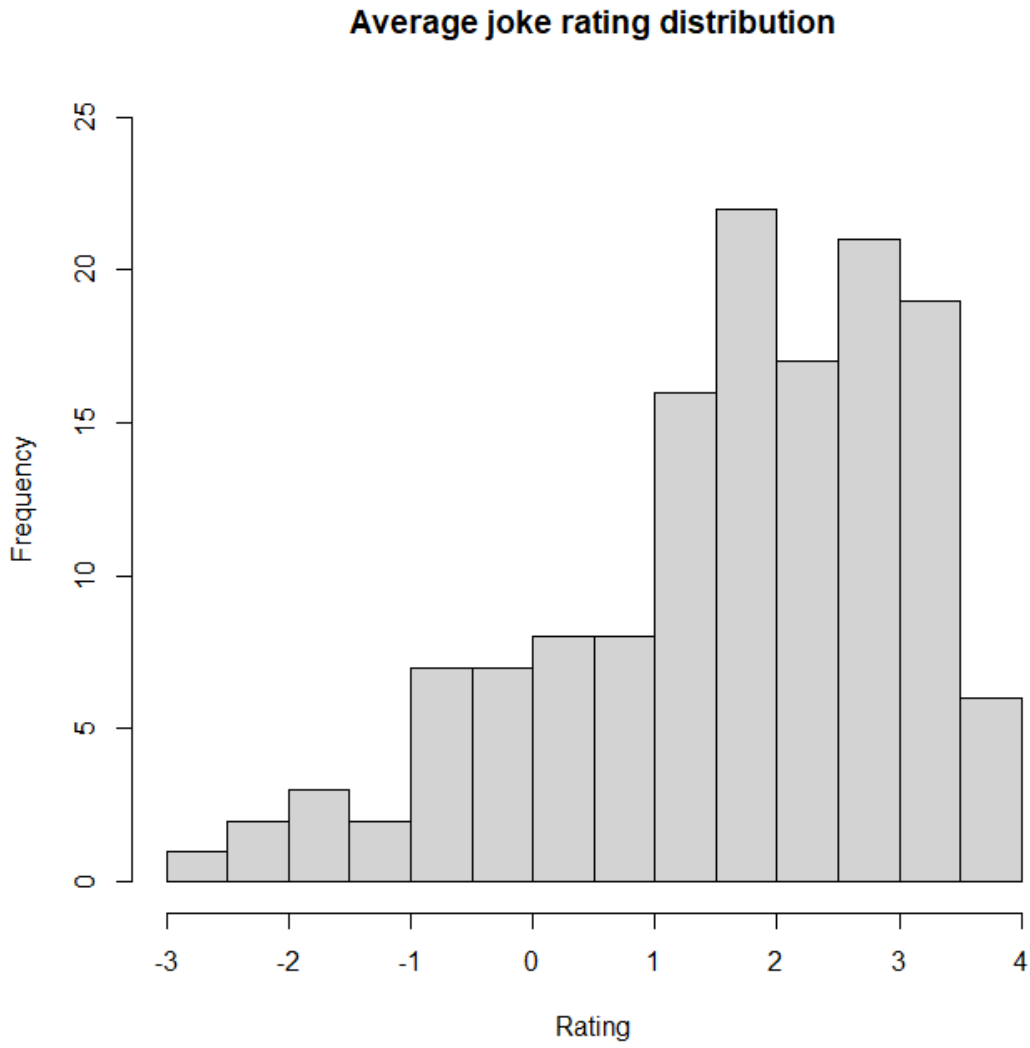
Here is a histogram representing distribution of the grade. The distribution is little skewed to the right.

```
hist(df$Rating,main = "Distribution of the grade", xlab = "Rating")
```



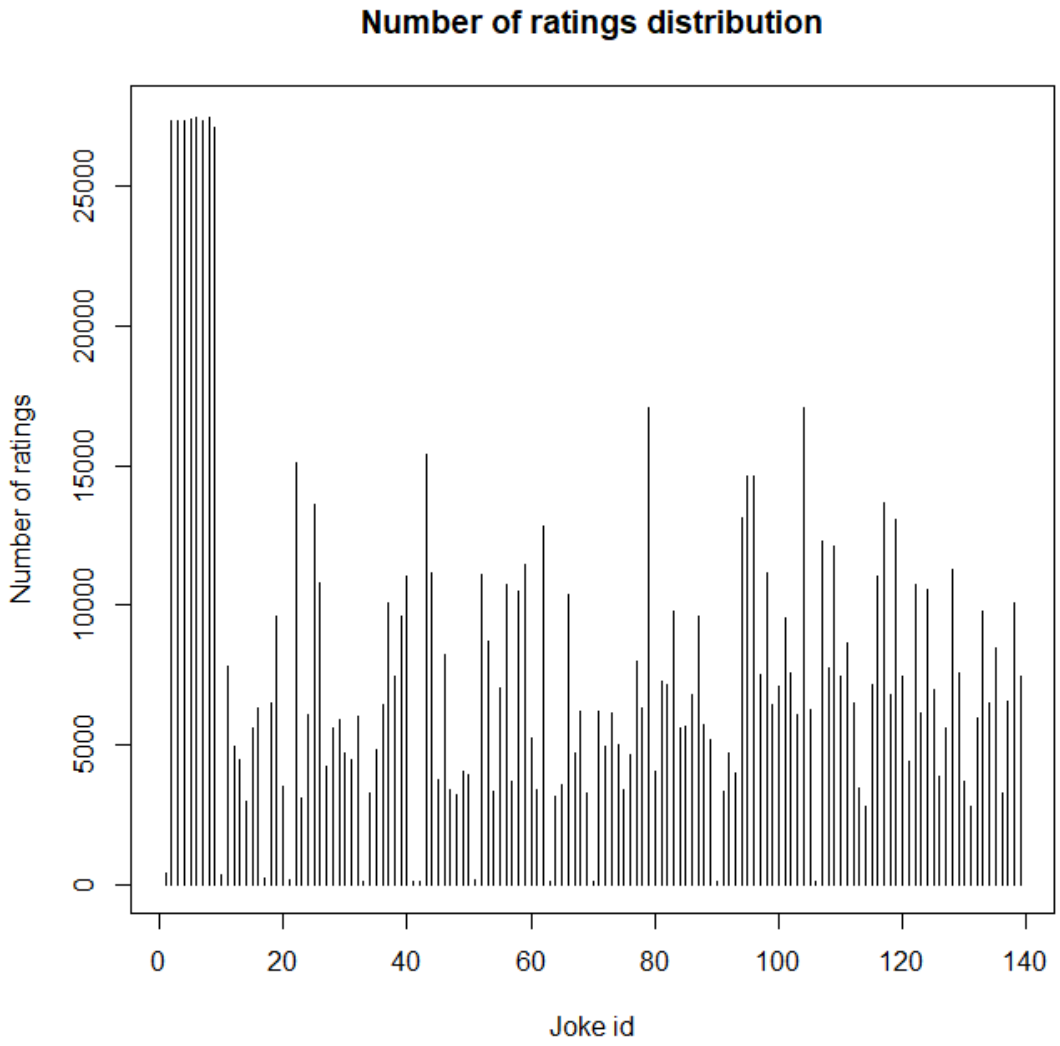
Here is a histogram representing average joke rating distribution. Average joke rating ranges from -3 to 4, the most popular are within [1.5, 2]. The distribution of average ratings is a little skewed to the right.

```
average_ratings <- aggregate(df$Rating ~ df$joke_id, df, mean)
hist(average_ratings$`df$Rating`,ylim=c(0,25), breaks = 20, main = "Average joke
rating distribution", xlab = "Rating")
```



Here is a histogram representing amount of ratings distribution. As we can see, the popularity of jokes is different. While some jokes have more than 25,000 ratings, others do not even reach 1000 ratings.

```
ratings_numbers <- aggregate(df$Rating ~ df$joke_id, df, length)
plot(ratings_numbers, type = "h", breaks = 139, main = "Number of ratings
distribution", xlab = "Joke id", ylab = "Number of ratings")
```



Here is a histogram representing standard deviations of jokes ratings distribution. As we can see, the standard deviations form a normal distribution. From this we can conclude that the ratings of jokes are compatible.

```
standard_deviations <- aggregate(df$Rating ~ df$joke_id, df, sd)
hist(standard_deviations$`df$Rating`, xlim = c(4, 6.5),ylim=c(0,20), breaks =
20, main="Standard deviations distribution", xlab = "Standard deviation",
     ylab = "Jokes number")
```