Recommender systems - Jokes

Obtaining dataset

Dataset was scraped from http://jokes.cc.com/

Software used:

- Google Chrome extension Web Scraper: http://webscraper.io/

Obtained dataset:

- csv format
- 29 categories
- +- 9 thousand jokes

Modifying dataset

Software used:

- Python3: https://docs.python.org/3/
- CSV to JSON: http://www.csvjson.com/csv2json
- RAKE-NLTK: https://github.com/csurfer/rake-nltk
- Script: https://goo.gl/jJ2p2x

Modifying dataset

Steps taken:

- Delete empty entries, empty attribute and change attribute names with regex.
- Delete categories with less than 100 jokes.
- Extract keywords from each joke with Rake.
- In each category, see the amount of short and long jokes.
- Balance dataset getting 100 jokes of each category having in count the amount of short and long jokes to be the same.

Modifying dataset

Evolution of the dataset:

- We start with 9182 jokes.
- After deleting the small categories we end up with 9090 jokes.
- After balancing the dataset we end up with 2898 jokes.

Modifying dataset

Category	Long	Short
Animal	126	304
Blonde	54	139
Blue Collar	81	87
Dark Humor	80	102
Dirty	88	267
Doctor	137	249
Fat	8	126
Food	72	162
God	111	176
Gross	106	230
Insults	83	361
Kids	181	275
Lookin' Good	107	314
Marriage	168	256
Men/Women	106	233
Miscellaneous	89	259
Money	92	208
Nationality	111	195
News & Politics	138	229
Partying & Bad Behavior	187	251
Pick-Up Lines	0	267
Police & Military	85	147
Pop Culture & Celebrity	130	338
School	86	96
Sports & Athletes	96	144
Technology	63	123
Travel & Car	96	178
Work	97	156
Yo' Mama	1	439

Total of long jokes: 2779
Total of short jokes: 6311

Category	Long	Short
Animal	56	44
Blonde	54	44
Blue Collar	56	44
Dark Humor	56	44
Dirty	56	44
Doctor	56	44
Fat	0	100
Food	56	44
God	56	44
Gross	56	44
Insults	56	44
Kids	56	44
Lookin' Good	56	44
Marriage	56	44
Men/Women	56	44
Miscellaneous	56	44
Money	56	44
Nationality	56	44
News & Politics	56	44
Partying & Bad Behavior	56	44
Pick-Up Lines	0	100
Police & Military	56	44
Pop Culture & Celebrity	56	44
School	56	44
Sports & Athletes	56	44
Technology	56	44
Travel & Car	56	44
Work	56	44
Yo' Mama	0	100

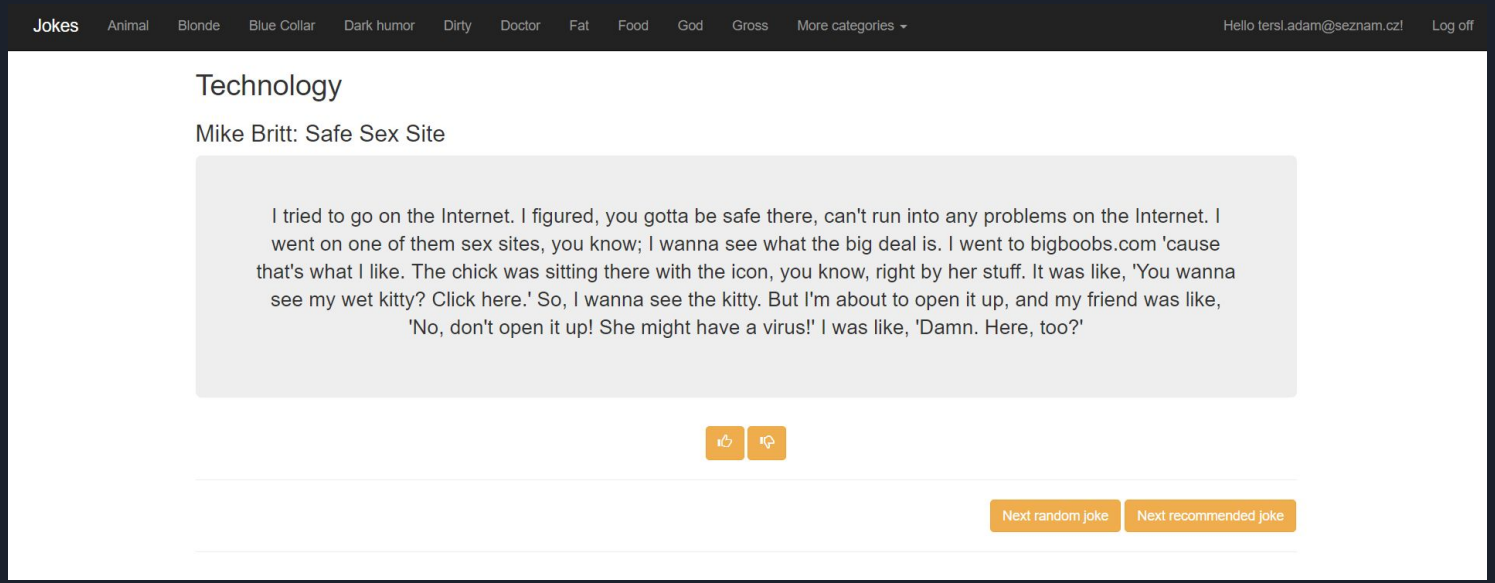
Total of long jokes: 1454
Total of short jokes: 1444

Web application

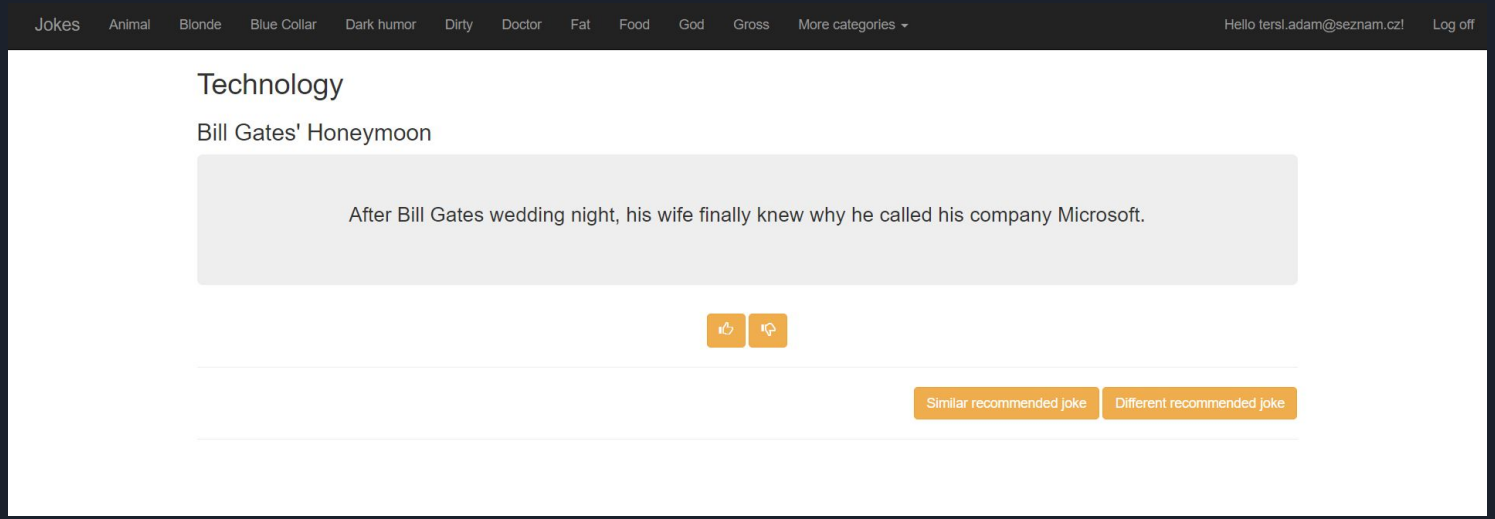
Software used:

- ASP.NET MVC: https://www.asp.net/mvc
- Entity Framework: https://docs.microsoft.com/en-us/aspnet/entity-framework
- Bootstrap 4: https://getbootstrap.com/

Index


Jokes
Animal
Blonde
Blue Collar
Dark humor
Dirty
Doctor
Fat
Food
God
Gross
More categories
Hello tersl.adam@seznam.cz!
Log off
Technology
Mike Britt: Safe Sex Site
I tried to go on the Internet. I figured, you gotta be safe there, can't run into any problems on the Internet. I went on one of them sex sites, you know; I wanna see what the big deal is. I went to bigboobs.com 'cause that's what I like. The chick was sitting there with the icon, you know, right by her stuff. It was like, 'You wanna see my wet kitty? Click here.' So, I wanna see the kitty. But I'm about to open it up, and my friend was like, 'No, don't open it up! She might have a virus!' I was like, 'Damn. Here, too?'
Next random joke
Next recommended joke

Chosen joke by user



TF/IDF recommending technique

- Every joke has 4 keywords
- Term frequency
  - calculated of every keyword of given joke
  - how many times keyword appears in category
  - result = appearance of keyword in category * all keywords in category
- Inverse document frequency
  - calculated of every keyword of given joke
  - how many times keyword appears in whole database of jokes
  - result = log(all keywords in whole database / appearance of keyword in whole database)
- term frequency * inverse document frequency
- recommends first not rated joke with max(tf * idf) from all shuffled jokes

Collaborative filtering by category

- category recommending
- takes all prefered categories of user
- selects each category of user and finds users which have same category in prefered ones
- counts categories which appears in other users
- recommends not rated joke from category which appears the most

Recommend similar/different joke

- Content + Collaborative based filtering
- Recommending:
    1. From the prefered category
    2. Not rated by the user
    3. Length category: Same X Other
    4. Jaccard index value: Closest X Farthest
- Jaccard index → set of keywords → |intersection| / |union|

Proposed evaluation

- RMSE (root mean square error)
- log-likelihood
- ratio between positively rated jokes and all seen jokes per user?
- statistics of how often each category was recommended?

Thank you