Predicting the Number of Views of StackOverflow Questions

•••

Student: Gevorg Atanesyan
Instructor: Hrant Davtyan

Table of Contents

- Scraping the data
- Data Preprocessing/Cleaning
- Feature Engineering
- Model Selection and Evaluation

Scraping the Data

We used **ScraPy Framework** to scrape most frequent **StackOverflow** questions, with randomized delays between each request.

Features Scraped:

- Title
- Description
- Answers
- Tags
- Views
- Urls

~5000

Questions Scraped

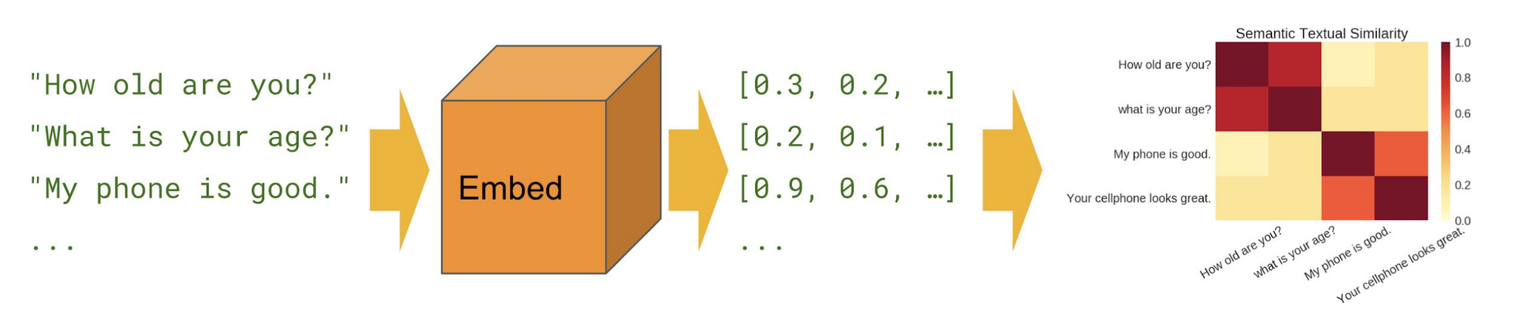
stack**overflow**

Data Preprocessing/Cleaning

- Converting string representation of views to integer (i.e. 2.8m -> 2800000, 1.2k -> 1200 etc.)
- Replace tag of programming language with the corresponding tiobe index (scraped from their website, using **pandas**)
- Converting descriptions from list of strings to a single multiline string
- Dropping NAs
- Using power function on the “views” feature to get rid of the skewness

Feature Engineering

Using **Google's Universal Sentence Encoder** to get dense vector representations (embeddings) for question descriptions



Model Selection and Evaluation

Models Used:

Linear Regression:

- R^2 = 0.75 (without target scaling)
- R^2 = 0.98 (with target scaling)[**overfit**]

Potential Improvements:

- Use more data
- Use regularization

WINNER

LightGBM Regression:

- R^2 = 0.79 (with target scaling)
- MAE ~ 200000
- MAPE ~ 68%

Potential Improvements:

- Use more data
- More aggressive hyperparameter tuning

Q&A

Thank You!