stackoverflow

Scraping Stackoverflow and Analysis

Student:Gevorg Atanesyan

Instructor:Hrant Davtyan

Introduction

Stackoverflow is one of the main websites which is used mostly by everyone who has relations with the IT sector. It is a way that developers ask questions based on their problem and get the answers.So it is an open community interface for people who code.The main purpose of this project is to scrap most frequently asked questions from stackoverflow and do analysis based on views.

Project Description

With the help of scraping techniques scrap stackoverflow (https://stackoverflow.com/questions?tab=frequent) content such as:

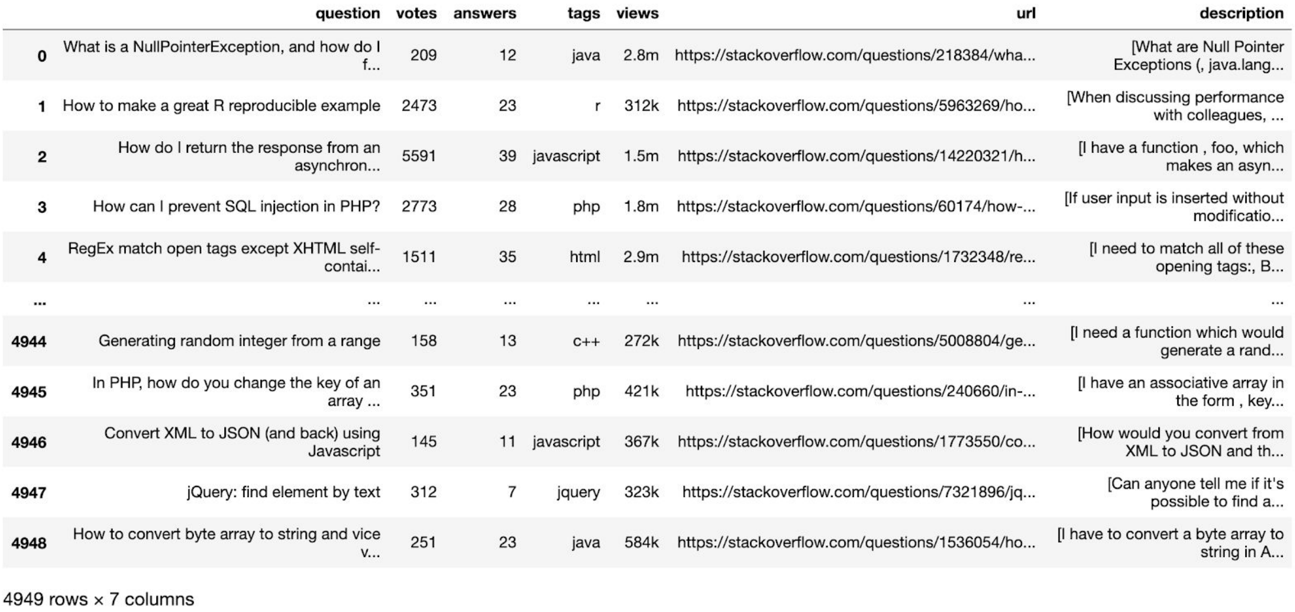
- Title
- Number of answers
- Number of Votes
- Tags
- Description
- Views
- Url

Data Scraping

There are a lot of packages which will help us to scrape data but in this project we use sracpy and scrapy framework. Almost 5000 rows data is scraped from stackoverflow. The main content is scraped with scrapy only the main description is scraped with the help of the scrapy framework.

With the scrapy scraped Titles,Number of answers,Number of Votes,Tags and urls because in one page we can get all this data so I requested only 100 requests to scrap the main data. The description part was a little bit difficult because I could not send too many requests . So in this case I used a scrapy framework which gives us an opportunity to use custom settings such as Download Delay,Randomize Downlad Delay. We used this custom setting because sending too many requests from one ip address system will block us after e.g 300 request.So with the help of custom settings requests have delays .

Data Cleaning

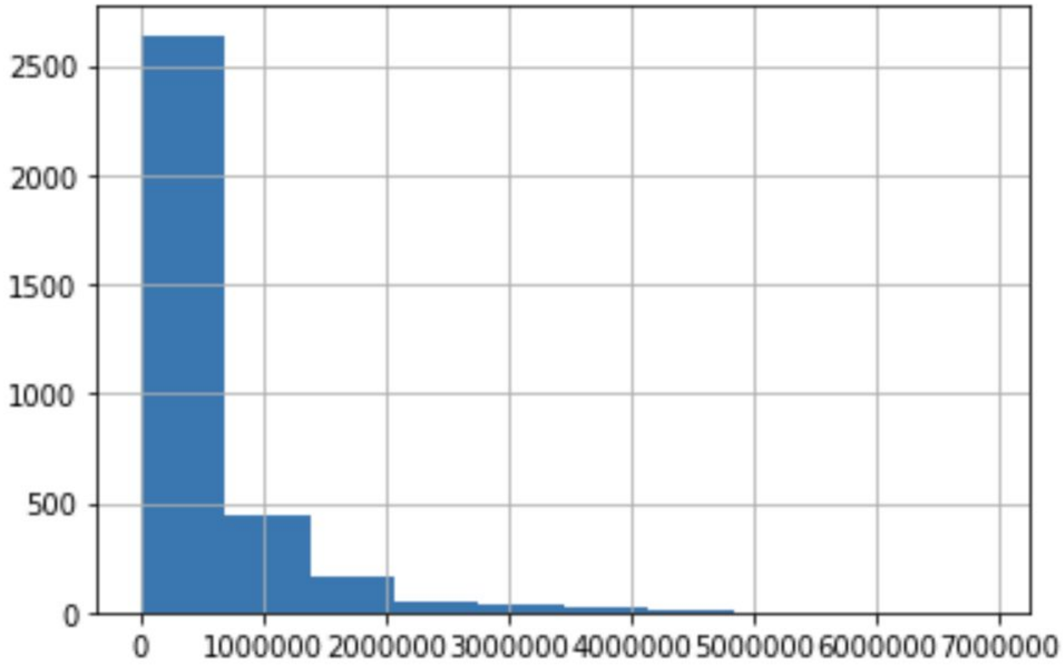


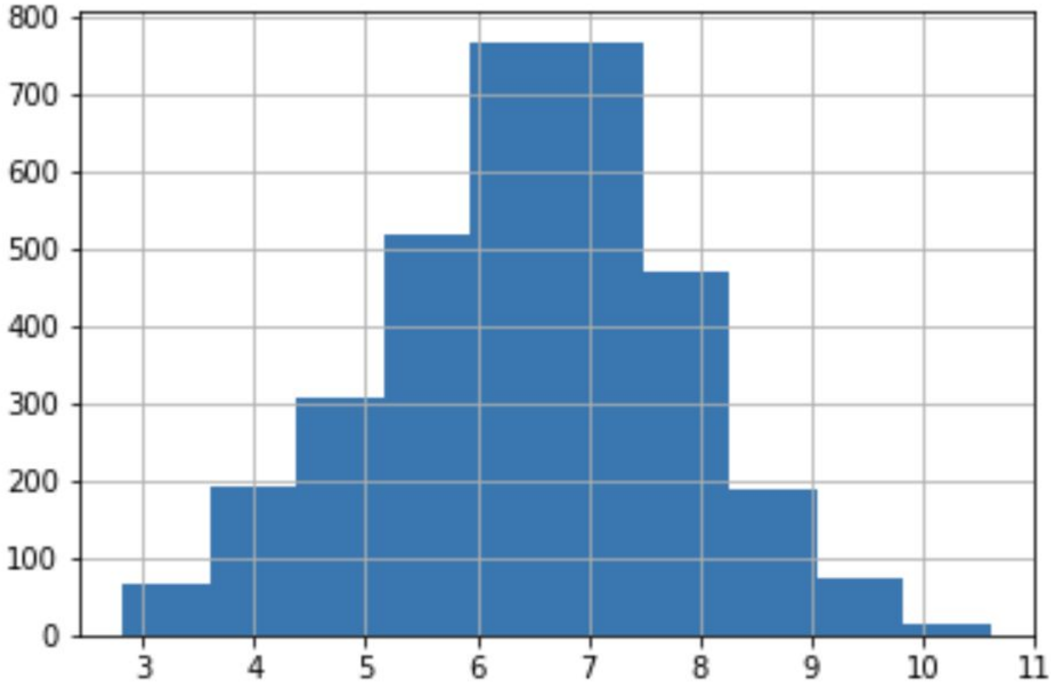
	question	votes	answers	tags	views	url	description
0	What is a NullPointerException, and how do I f...	209	12	java	2.8m	https://stackoverflow.com/questions/218384/wha...	[What are Null Pointer Exceptions (, java.lang...
1	How to make a great R reproducible example	2473	23	r	312k	https://stackoverflow.com/questions/5963269/ho...	[When discussing performance with colleagues, ...
2	How do I return the response from an asynchron...	5591	39	javascript	1.5m	https://stackoverflow.com/questions/14220321/h...	[I have a function , foo, which makes an asyn...
3	How can I prevent SQL injection in PHP?	2773	28	php	1.8m	https://stackoverflow.com/questions/60174/how-...	[If user input is inserted without modificatio...
4	RegEx match open tags except XHTML self-contai...	1511	35	html	2.9m	https://stackoverflow.com/questions/1732348/re...	[I need to match all of these opening tags:, B...
...	...	...	...	...	...	...	...
4944	Generating random integer from a range	158	13	c++	272k	https://stackoverflow.com/questions/5008804/ge...	[I need a function which would generate a rand...
4945	In PHP, how do you change the key of an array ...	351	23	php	421k	https://stackoverflow.com/questions/240660/in-...	[I have an associative array in the form , key...
4946	Convert XML to JSON (and back) using Javascript	145	11	javascript	367k	https://stackoverflow.com/questions/1773550/co...	[How would you convert from XML to JSON and th...
4947	jQuery: find element by text	312	7	jquery	323k	https://stackoverflow.com/questions/7321896/jq...	[Can anyone tell me if it's possible to find a...
4948	How to convert byte array to string and vice v...	251	23	java	584k	https://stackoverflow.com/questions/1536054/ho...	[I have to convert a byte array to string in A...

4949 rows × 7 columns

We can see how our data looks when all scraping parts are finished. The next important thing was to do data cleaning. As we can see views are in string format so I converted them to floats with the help of lambda function.(e.g 2.8m became 2.800.000, 312k became 312.000).The next step which I have done to change tags to numbers. Tiobe is a website(https://www.tiobe.com/tiobe-index/) where we can see programming language rankings. So I scraped tiobe rankings and with the help of map function I changed all tag names with the tiobe's rankings (e.g Tags C became 1, Java became 2, Python became 3 etc). After mapping tiobe ranking I got about 3300 row tags with tiobe rankings and 1700 rows NA's which I dropped. The next thing was converting descriptions from lists of strings to a single multiline string.

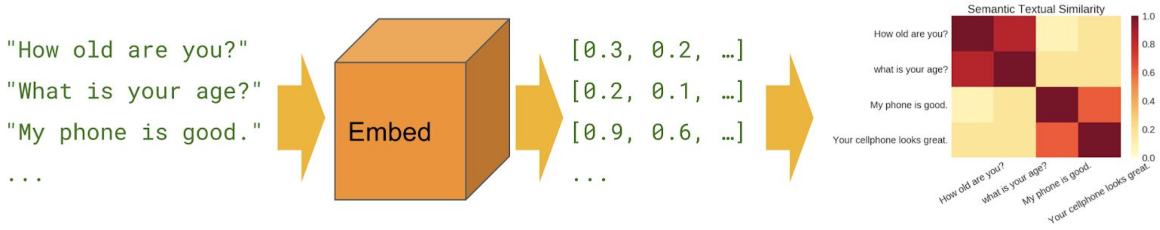
Using power function on the “views” feature to get rid of the skewness after multiplying views with 0.15





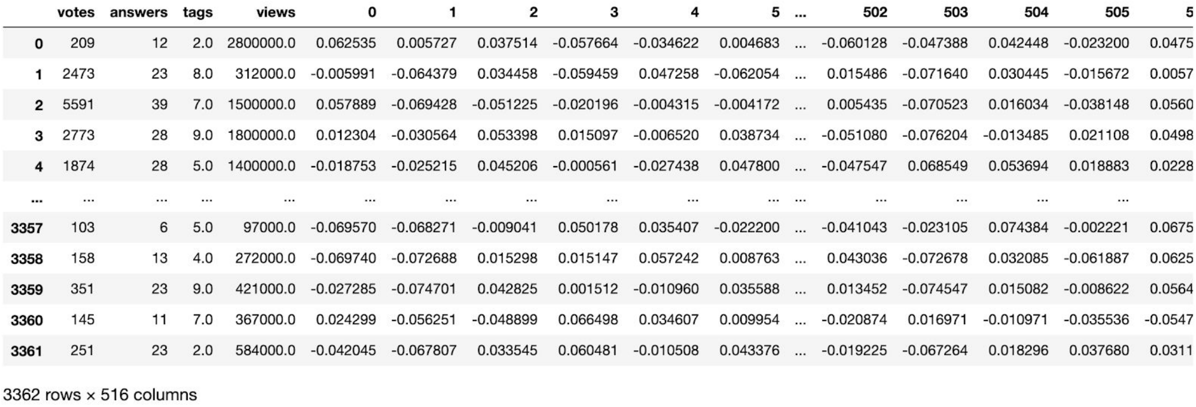
Feature Engineering

For feature engineering I used Google’s Universal Sentence Encoder to get dense vector representations (embeddings) for question descriptions(https://tfhub.dev/google/universal-sentence-encoder/4).



It is already trained on data which takes an input of an English text; the output is a 512 dimensional vector.

After doing Feature engineering and data cleaning out final data was



	votes	answers	tags	views	0	1	2	3	4	5	...	502	503	504	505	5
0	209	12	2.0	2800000.0	0.062535	0.005727	0.037514	-0.057664	-0.034622	0.004683	...	-0.060128	-0.047388	0.042448	-0.023200	0.0475
1	2473	23	8.0	312000.0	-0.005991	-0.064379	0.034458	-0.059459	0.047258	-0.062054	...	0.015486	-0.071640	0.030445	-0.015672	0.0057
2	5591	39	7.0	1500000.0	0.057889	-0.069428	-0.051225	-0.020196	-0.004315	-0.004172	...	0.005435	-0.070523	0.016034	-0.038148	0.0560
3	2773	28	9.0	1800000.0	0.012304	-0.030564	0.053398	0.015097	-0.006520	0.038734	...	-0.051080	-0.076204	-0.013485	0.021108	0.0498
4	1874	28	5.0	1400000.0	-0.018753	-0.025215	0.045206	-0.000561	-0.027438	0.047800	...	-0.047547	0.068549	0.053694	0.018883	0.0228
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3357	103	6	5.0	97000.0	-0.069570	-0.068271	-0.009041	0.050178	0.035407	-0.022200	...	-0.041043	-0.023105	0.074384	-0.002221	0.0675
3358	158	13	4.0	272000.0	-0.069740	-0.072688	0.015298	0.015147	0.057242	0.008763	...	0.043036	-0.072678	0.032085	-0.061887	0.0625
3359	351	23	9.0	421000.0	-0.027285	-0.074701	0.042825	0.001512	-0.010960	0.035588	...	0.013452	-0.074547	0.015082	-0.008622	0.0564
3360	145	11	7.0	367000.0	0.024299	-0.056251	-0.048899	0.066498	0.034607	0.009954	...	-0.020874	0.016971	-0.010971	-0.035536	-0.0547
3361	251	23	2.0	584000.0	-0.042045	-0.067807	0.033545	0.060481	-0.010508	0.043376	...	-0.019225	-0.067264	0.018296	0.037680	0.0311

3362 rows × 516 columns

Model Selection

The data is divided into training and testing sets (test size 30%).

With Linear regression results were:

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

The way to improve this result is to scrap more data or use regularization

LightGBM Regression:

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

The way to improve this results to do tuning based on parameters or scrap more data.So to sum up LightGBM was better Mean absolute error is 200000 but with the more data it could be improved.