Project: Q&A from Stack Overflow

Objective:

1. R & Python question difference
2. What types of questions and answers are more popular
3. Question tag recommendation system

Data source:

R Q&A from Stack overflow: <https://www.kaggle.com/stackoverflow/rquestions>

Python Q&A from Stack overflow: <https://www.kaggle.com/datasets?sortBy=hottest&group=featured&search=stack>

Dataset description: take R Q&A datasets for example, Python Q&A datasets are the same.

R Q&A datasets are organized as three tables, including full text of questions and answers from Stack Overflow that are tagged with the r tag.

Table 1: Questions. It contains the title, body, creation date, score, and owner ID for each R question.

Table 2: Answers. It contains the body, creation date, score, and owner ID for each of the answers to these questions.

Table 3: Tags contains the tags on each question besides the R tag.

Time range for the Q&A data is from 2008-09-16 to 2016-10-19.

Statistical and machine learning methods we use:

Text Mining

Two sample test

Linear regression

Logistic regression

LDA

Topic Modelling

Part 1:

For this part, we try to analyze the difference between R and Python.

1. How do they change in terms of the popularity degree?
2. What’s the hottest topics for the two languages respectively?
3. How long does it take to have an answer posted after asking a question on this website?
4. Scores you can get as being active on the website?

Question 1: How do they change in terms of the popularity degree?

This is the question’s amount change from 2008 to 2016. It shows that Python and R keep growing in a high speed. Under the assumption that Stack Overflow is a relatively mature Q&A website that it’s users number doesn’t change extremely.

Question 2: What’s the hottest topics for the two languages respectively?

From the word cloud, we can see the focus attentions for R are ggplot2, dataframe and shiny, concentrating on data visualization and report delivery. Also, statistics, regression are common topics for R.

For Python, django, numpy, pandas, matplotlib attract our attention.

Question 3: How long does it take to have an answer posted after asking a question on this website?

For R, the mean time is 47.96628 days, median time is 0.032 days, that is over 45 minutes. While, for Python, the mean time is 70.01714 days, the median is 0.02 days, that is less than half an hour. Since the time data is not normal, we use rank sum test and find that the time for R and Python are significantly different.

You may think these as, that if you ask a question about Python on this website, you may receive an answer sooner, and keep receiving answers in a longer period.

Question 4: Scores you can get as being active on the website?

On a Q&A website, if you are an expert in programming languages, there might be some tricks for you to get higher scores.

By exploring the data, we find there’s obvious linear relationship between the question’s score and its corresponding highest answer’s score.

For R, answer’s score= 1.01+ 0.93\* question’s score

For Python, answer’s score= 0.80+1.11\* question’s score

It indicates that you can get higher score by answering popular questions. Also, get relatively higher score by answering a question about Python. Since the coefficient for Python is 1.11, higher than that for R.