Analysis of Stack Overflow Q&A

Semih Barutcu

12/10/2020

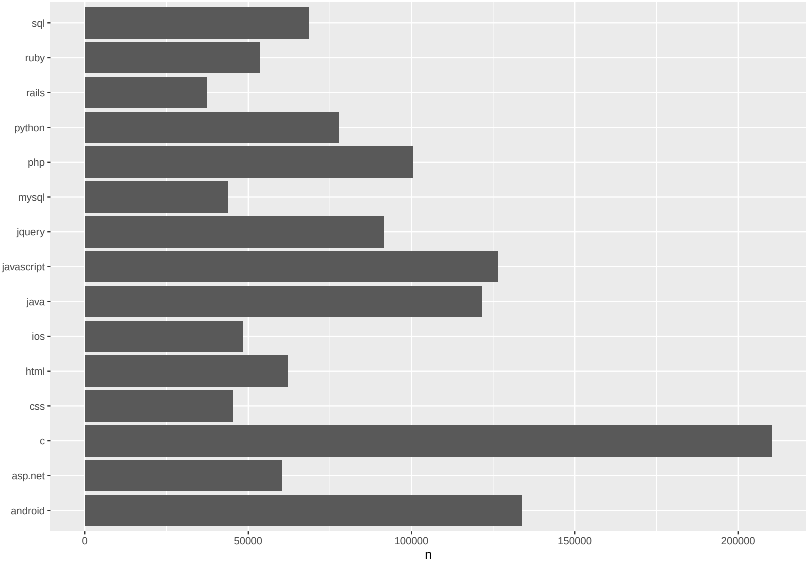
Introduction

- Exploring the relationship between questions and tags.
- Using Text Mining tools
- Predicting tags from the question bodies.
- Resource Book: Text Mining with R by Julia Silge and David Robinson. https://www.tidytextmining.com/

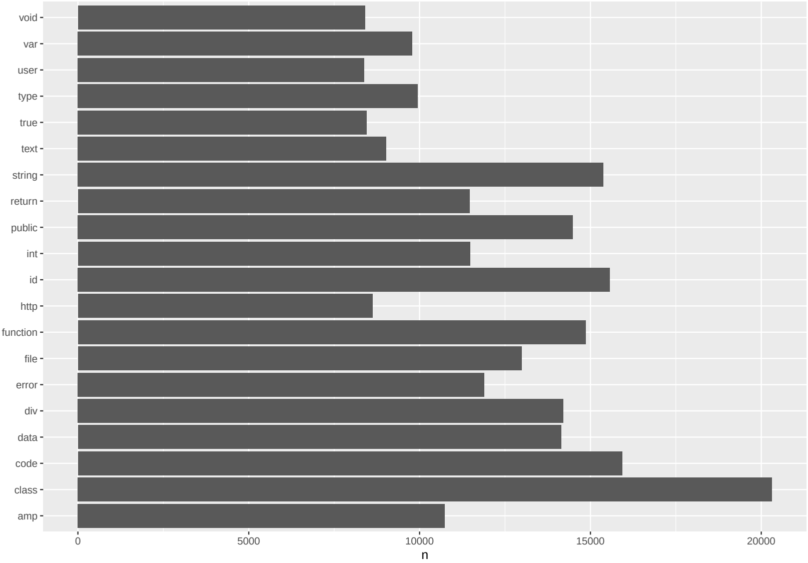
About The Dataset

- A kaggle dataset: StackSample: 10% of Stack Overflow Q&A. https://www.kaggle.com/stackoverflow/stacksample.
- Dataset with the text of 10% of questions and answers from the Stack Overflow.
- Used Questions and Tags datasets.

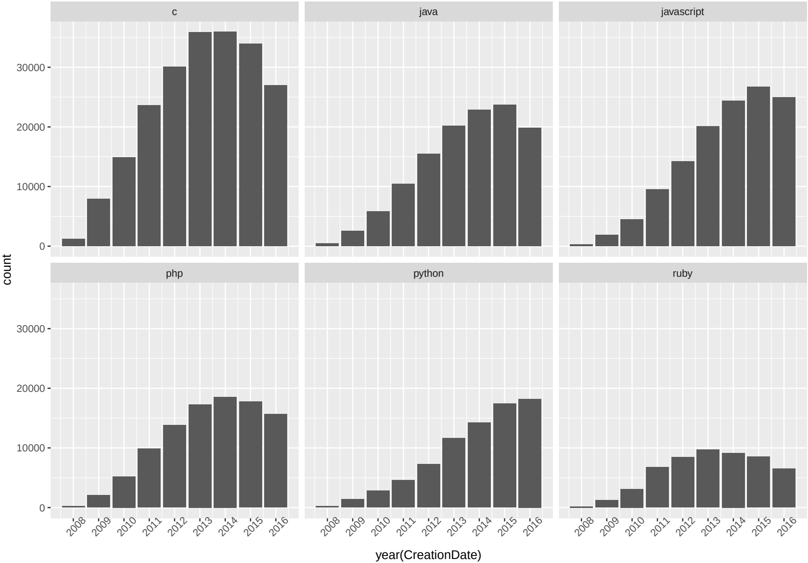
The Most Popular Tags



The Most Popular Words



Popularity of Programming Languages by Year



Average Scores of the Selected Languages

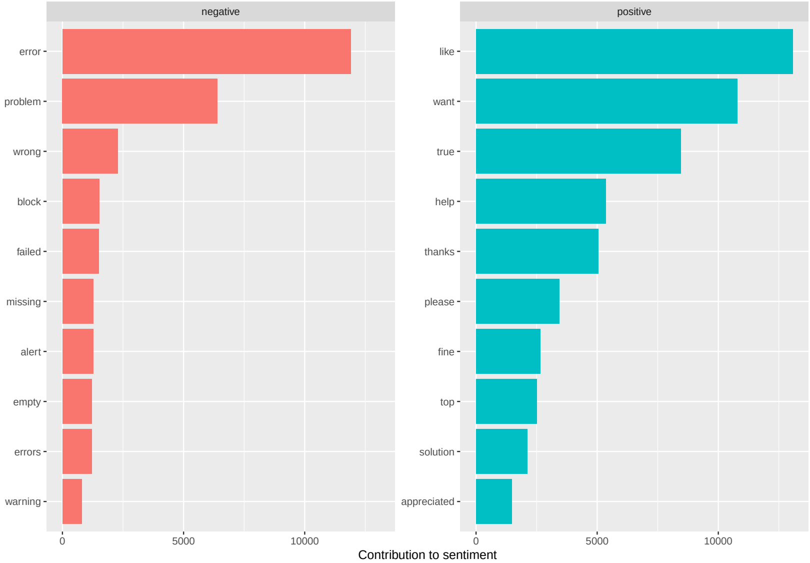
```
## # A tibble: 6 x 6
##   tag          Avg   Min   Max Q_025 Q_975
##   <chr>      <dbl> <int> <int> <dbl> <dbl>
## 1 c           2.13   -20  1473    -2    13
## 2 python      1.98   -23   824    -2    12
## 3 ruby        1.94   -11   648    -1    12
## 4 java        1.78   -45  3613    -2    11
## 5 javascript  1.67   -18  2363    -2     9
## 6 php         0.99   -13  1760    -2     6
```

Sentiment Analysis

- Used AFINN lexicon.

```
## # A tibble: 6 x 2
##        Id sentiment
##     <int>     <dbl>
## 1  1180         6
## 2  8050         2
## 3 12890        -1
## 4 20850        -5
## 5 22570        -3
## 6 32780        -1
```

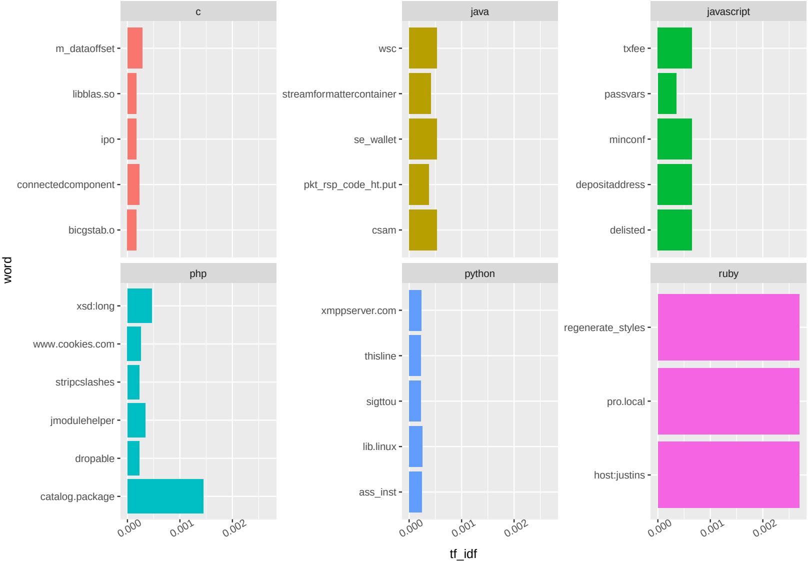
Top 10 Positive and Negative Words



Term Frequency

```
## # A tibble: 968,691 x 8
## # Groups:   tag [6]
##          Id word                   n total tag          tf
##       <int> <chr>              <int> <int> <chr>     <dbl>
##  1 27710710 catalog.package      201  1019 php    0.000804
##  2 25799910 host:justins         158  1692 ruby   0.00123
##  3 25799910 host:justins         158  1692 ruby   0.00123
##  4 25799910 pro.local            158  1692 ruby   0.00123
##  5 25799910 pro.local            158  1692 ruby   0.00123
##  6 25799910 regenerate_styles    158  1692 ruby   0.00123
##  7 25799910 regenerate_styles    158  1692 ruby   0.00123
##  8 16463710 emai                  61   812 ruby   0.000474
##  9 34561050 capin                 35   933 ruby   0.000272
## 10 18322730 xsd:long              65   594 php    0.000260
## # ... with 968,681 more rows
```

The Words with Highest tf_idf Scores of Selected Languages



Conclusion