Recommending Tags for Stack Overflow

Sarah Floris

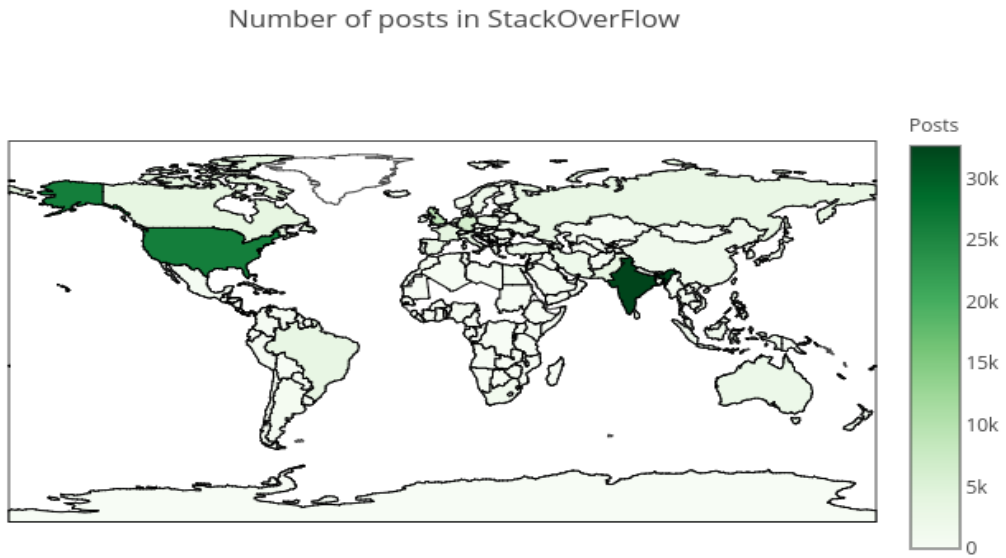
Querying

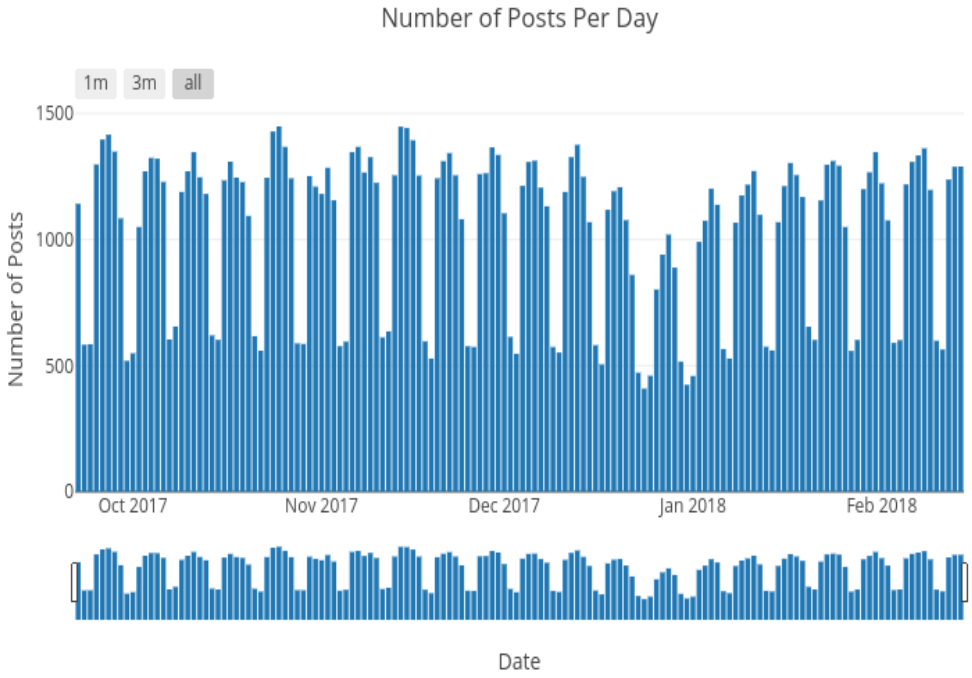
```
SELECT
    Users.Id AS user_id,
    Users.Location AS user_location,
    Users.Age AS user_age,
    Users.Reputation AS user_reputation,
    Posts.Id AS post_id,
    Posts.Score AS post_score,
    Posts.ViewCount AS post_views,
    Posts.CreationDate AS post_creation_date,
    Posts.Tags AS tags,
    Posts.Title AS title,
    Posts.Body AS body
FROM Users,Posts WHERE Users.Location IS NOT NULL AND Users.Age IS NOT NULL
AND Posts.ViewCount IS NOT NULL
AND Users.Id = Posts.OwnerUserId
AND Posts.CreationDate < '2018-02-15'
ORDER BY Posts.CreationDate DESC
```

Website: data.stackexchange.com/stackoverflow

Post Characteristics

India	32632
United States	26181
United Kingdom	8676
Germany	6454
France	3927
Canada	3618

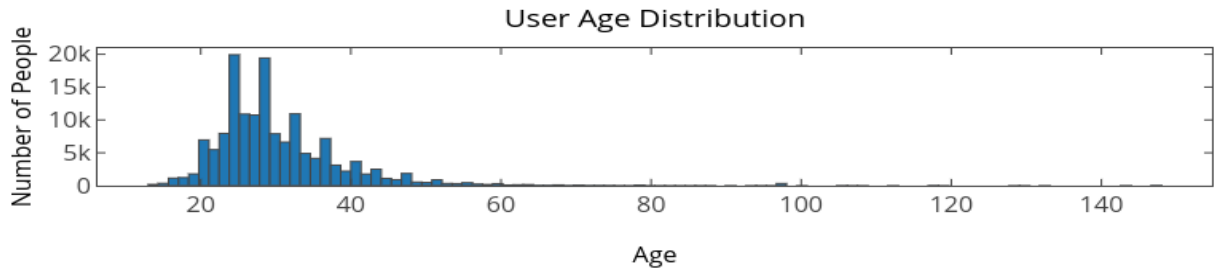


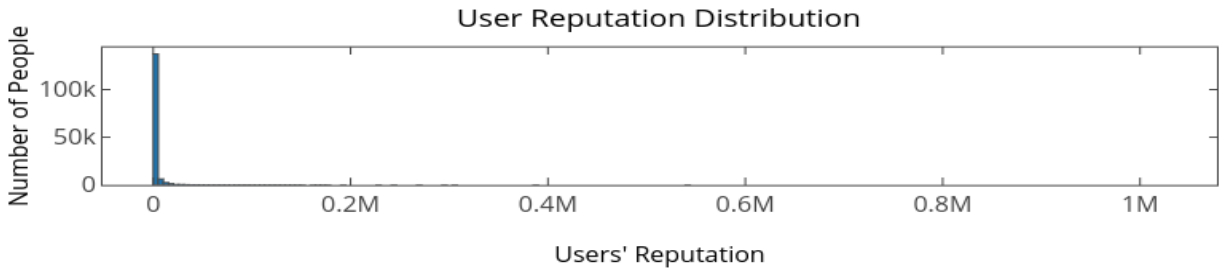


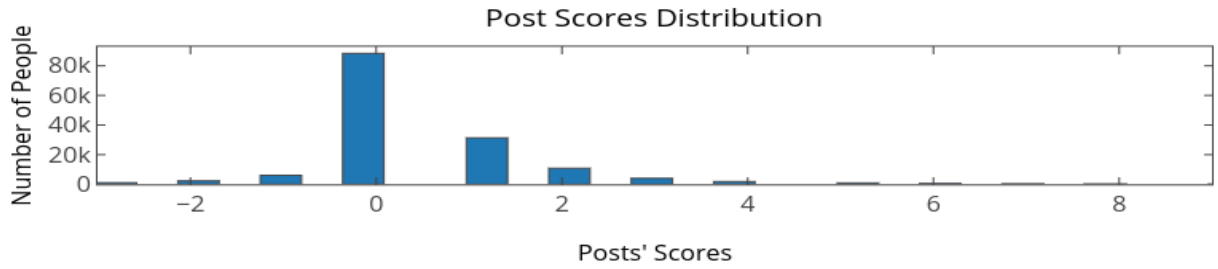
Number of posts decreases during:

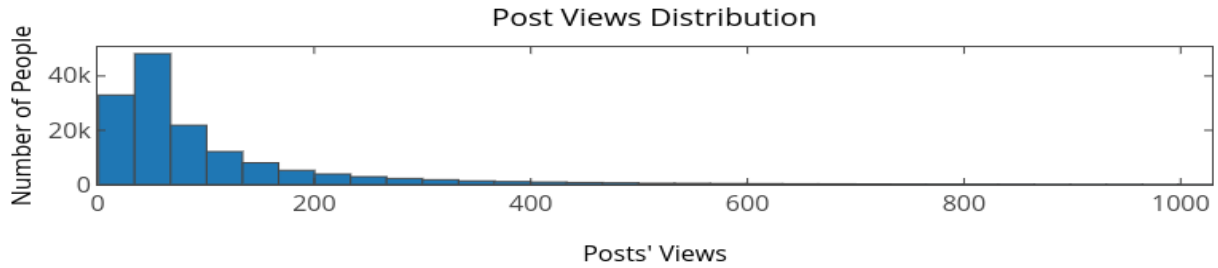
- Weekends
- Christmas time
- New Years

Post Characteristics



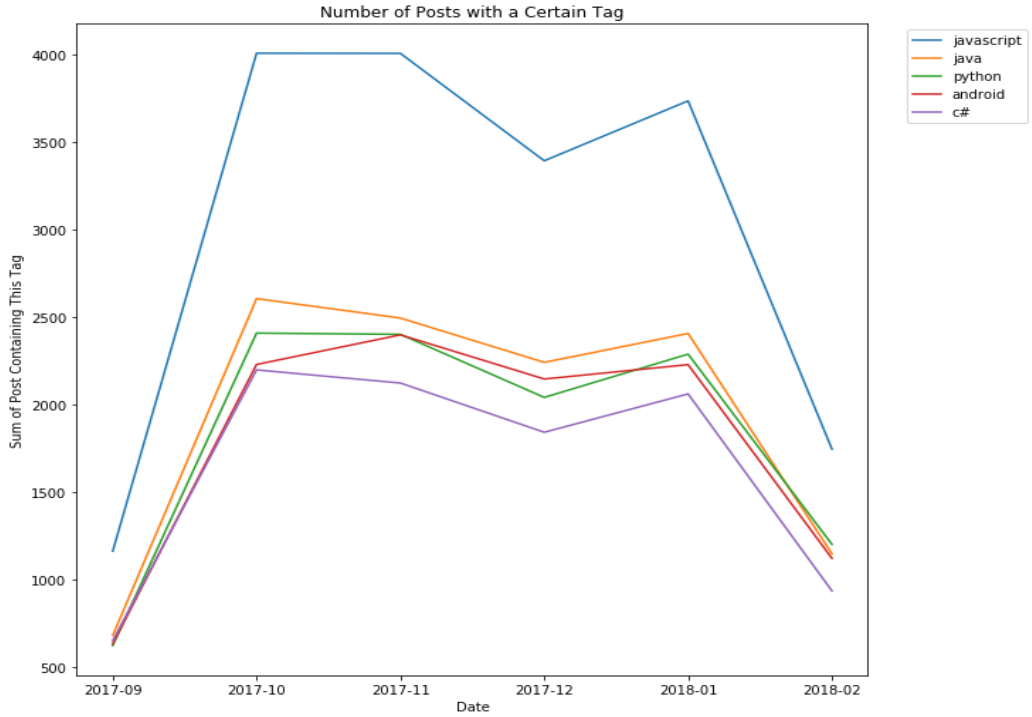


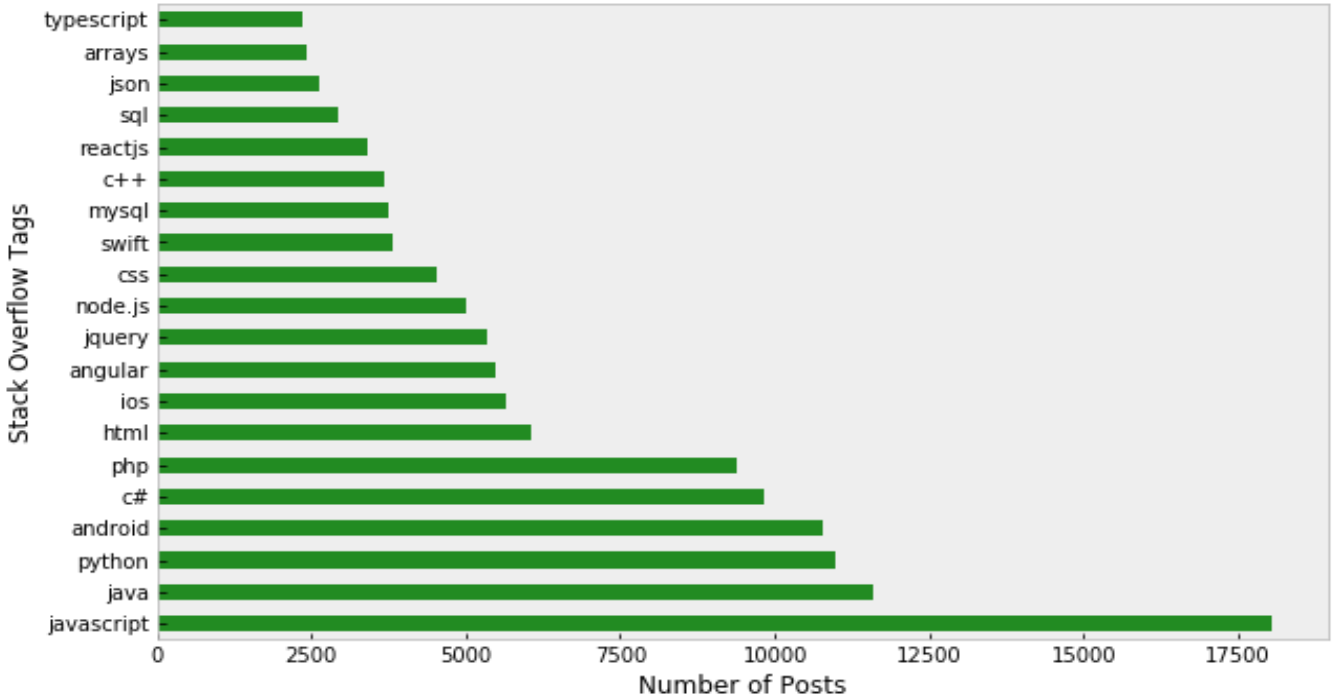


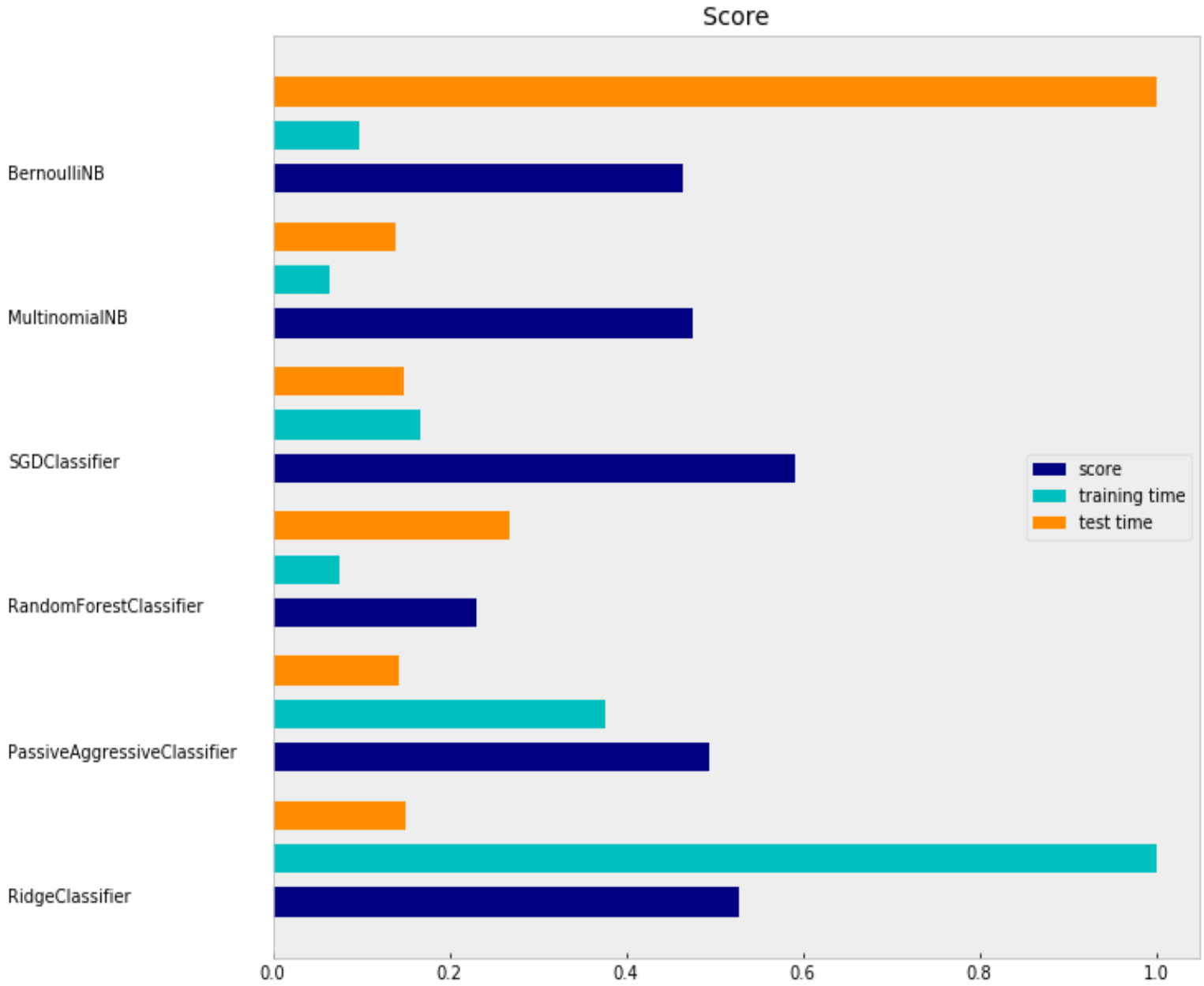


	post_score	post_views	user_age	user_reputation
count	150000.000000	150000.000000	150000.000000	1.500000e+05
mean	0.558020	145.574493	30.252747	2.348137e+03
std	6.006465	1087.404256	8.942015	1.196191e+04
min	-19.000000	1.000000	13.000000	1.000000e+00
25%	0.000000	37.000000	25.000000	5.300000e+01
50%	0.000000	61.000000	28.000000	3.200000e+02
75%	1.000000	124.000000	34.000000	1.241000e+03
max	2200.000000	332291.000000	148.000000	1.026718e+06

Tags


Number of Posts with a Certain Tag
javascript
java
python
android
c#
Sum of Post Containing This Tag
4000
3500
3000
2500
2000
1500
1000
500
2017-09
2017-10
2017-11
2017-12
2018-01
2018-02
Date


typescript
arrays
json
sql
reactjs
c++
mysql
swift
css
node.js
jquery
angular
ios
html
php
c#
android
python
java
javascript
Stack Overflow Tags
0
2500
5000
7500
10000
12500
15000
17500
Number of Posts


Score
BernoulliNB
MultinomialNB
SGDClassifier
RandomForestClassifier
PassiveAggressiveClassifier
RidgeClassifier
score
training time
test time
0.0
0.2
0.4
0.6
0.8
1.0

	precision	recall	f1-score	support
mysql	0.66	0.87	0.75	2146
ios	0.54	0.82	0.65	1070
css	0.29	0.06	0.10	488
html	0.66	0.86	0.74	1960
arrays	0.76	0.78	0.77	736
reactjs	0.47	0.34	0.40	972
json	0.31	0.14	0.19	1212
javascript	0.57	0.66	0.61	1121
typescript	0.71	0.73	0.72	2340
android	0.49	0.41	0.44	3595
php	0.34	0.08	0.12	1023
java	0.18	0.13	0.15	496
jquery	0.59	0.32	0.42	736
python	0.53	0.66	0.59	986
c#	0.61	0.83	0.70	1956
swift	0.72	0.94	0.81	2212
angular	0.51	0.77	0.62	661
sql	0.60	0.44	0.51	566
c++	0.53	0.24	0.33	751
node.js	0.47	0.11	0.18	499
avg / total	0.56	0.60	0.56	25526

Limitations

Computing power was a precious commodity

User's id matched the posts owner's id

Users wrote their own age and location

Further Research

Introduce the StackOverflow API to pull and query data

Identify a software tag (i.e. python is a programming language)

Implement big data tools such as dask and docker

Provide a dashboard to verify tag as correct and incorrect