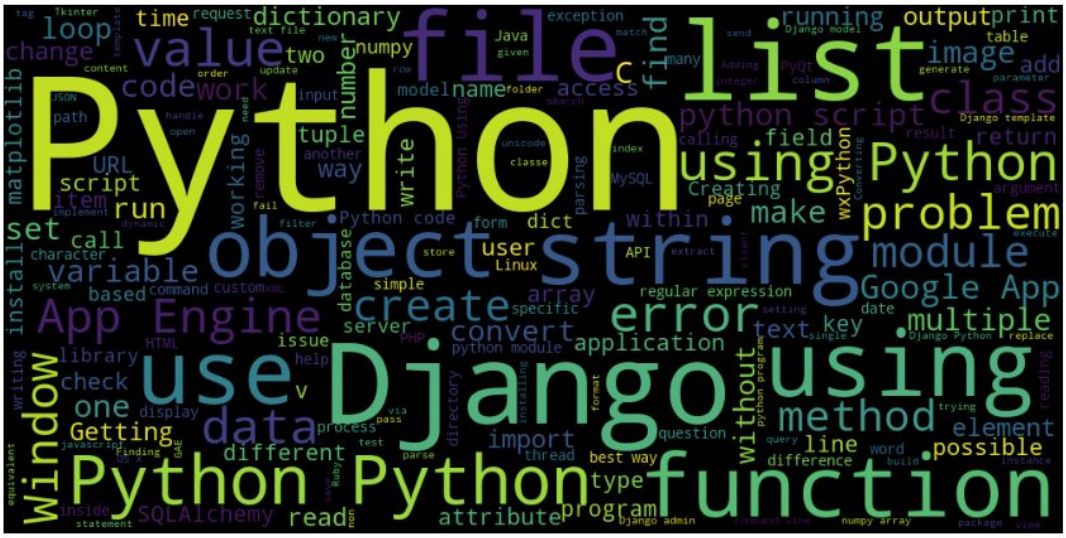
Precog Task 3b Report

Daksh Thapar (2018137)

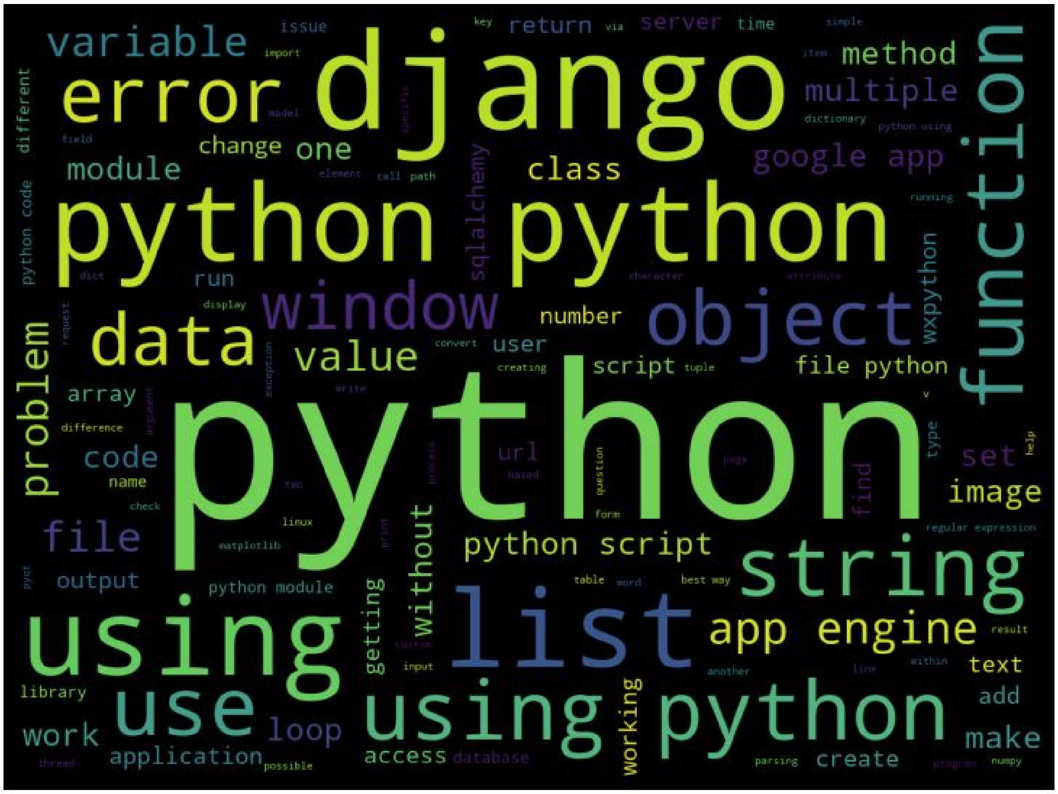
- All visualizations can be found in ipynb and pdf files.
- The field "**Tags**" Posts.xml is used for **Common Subsampling** (common factor among the dataset)

- Created **Word Clouds for Titles and Bodies** of text using Posts.xml

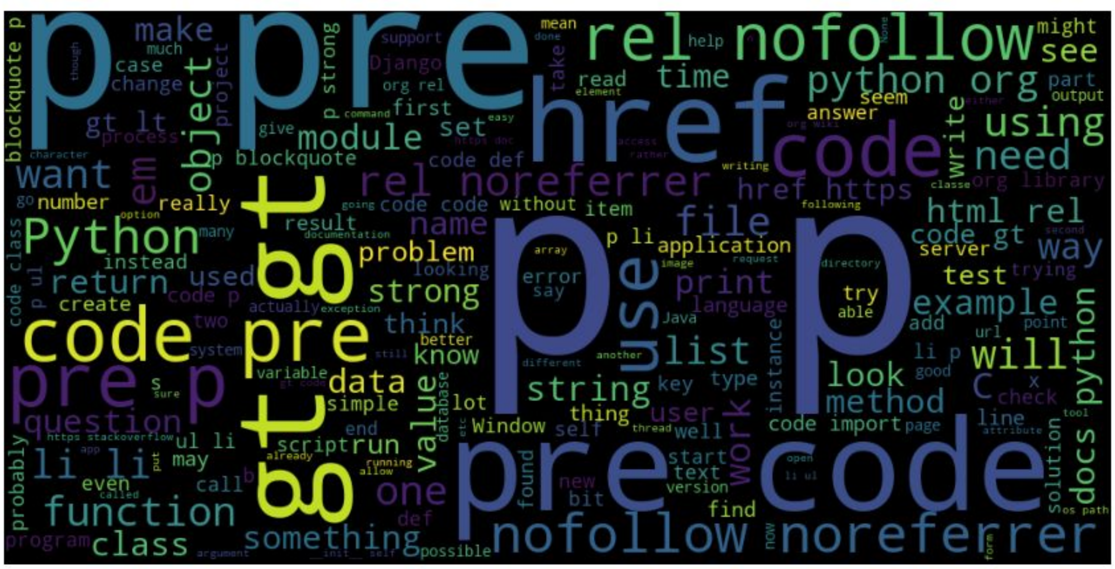
Titles:



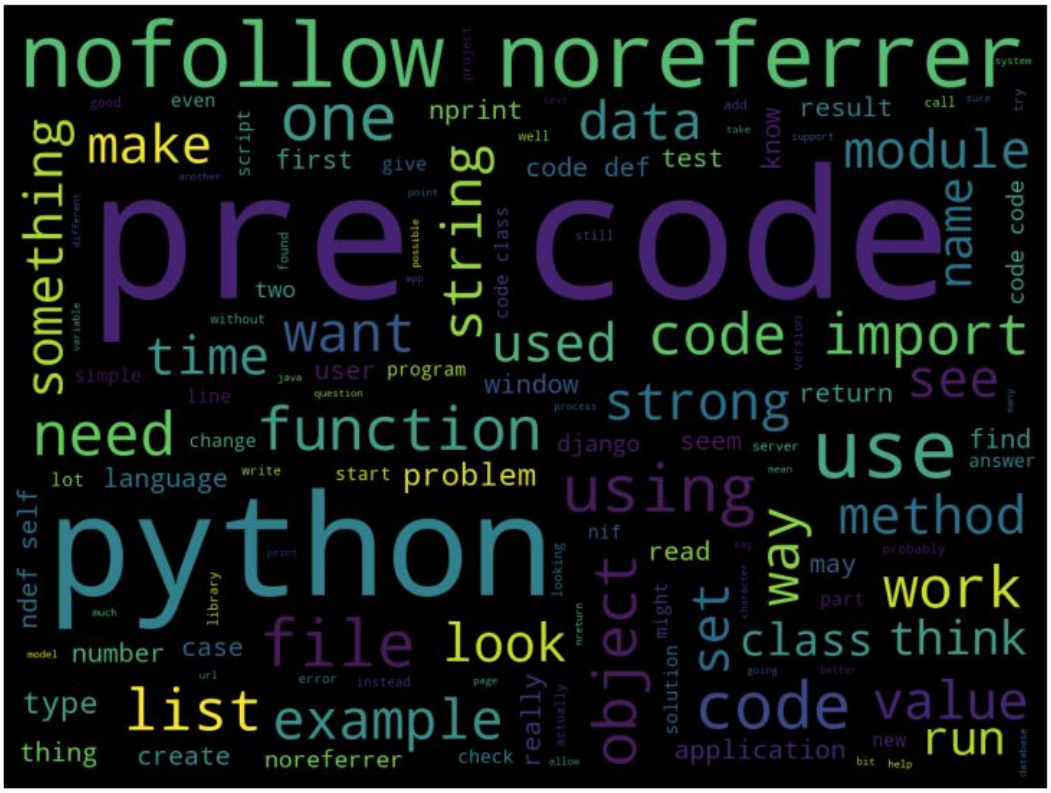
Titles (preprocessed, stopwords removed, tokenized)



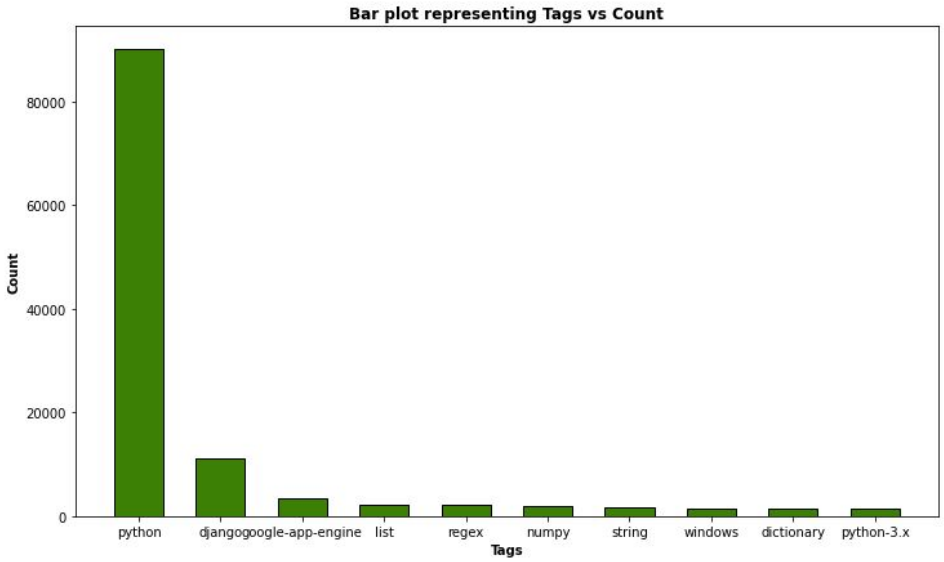
Bodies



Bodies (preprocessed, stopwords removed, tokenized)



- **Top 10 occurring tags using Posts.xml**



21]:
```
print(tags_)
print(counts_)
```

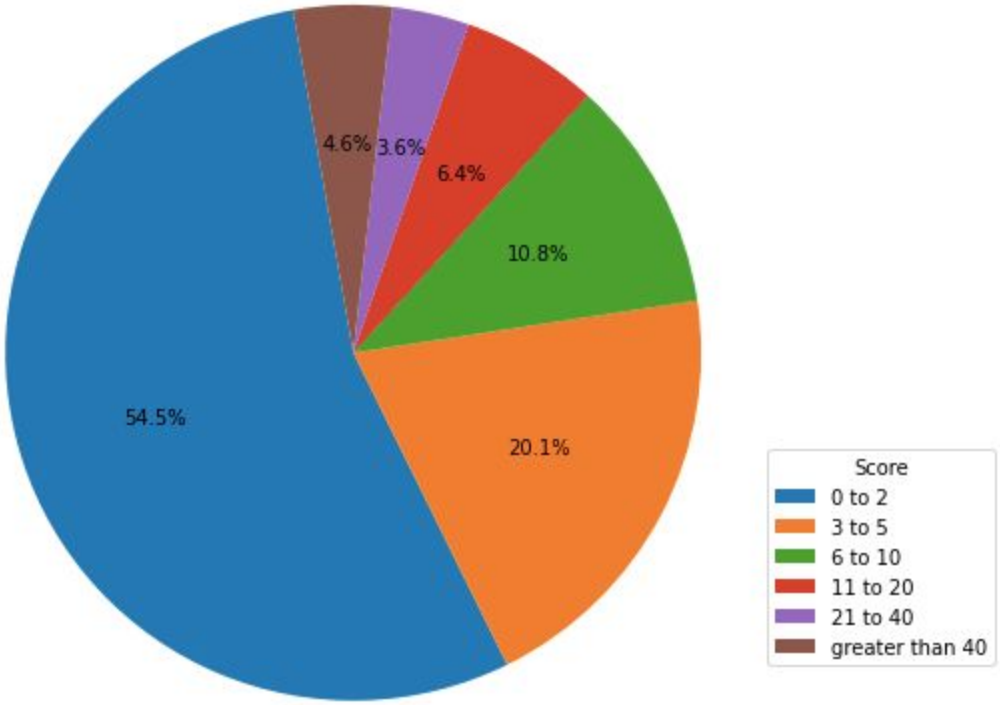
```
['python', 'django', 'google-app-engine', 'list', 'regex', 'numpy', 'string', 'windows', 'dictionary', 'python-3.x']
[90051, 11046, 3427, 2283, 2146, 1961, 1801, 1462, 1452, 1353]
```

- **Tags in Posts.xml**

```
df["Tags"]
```

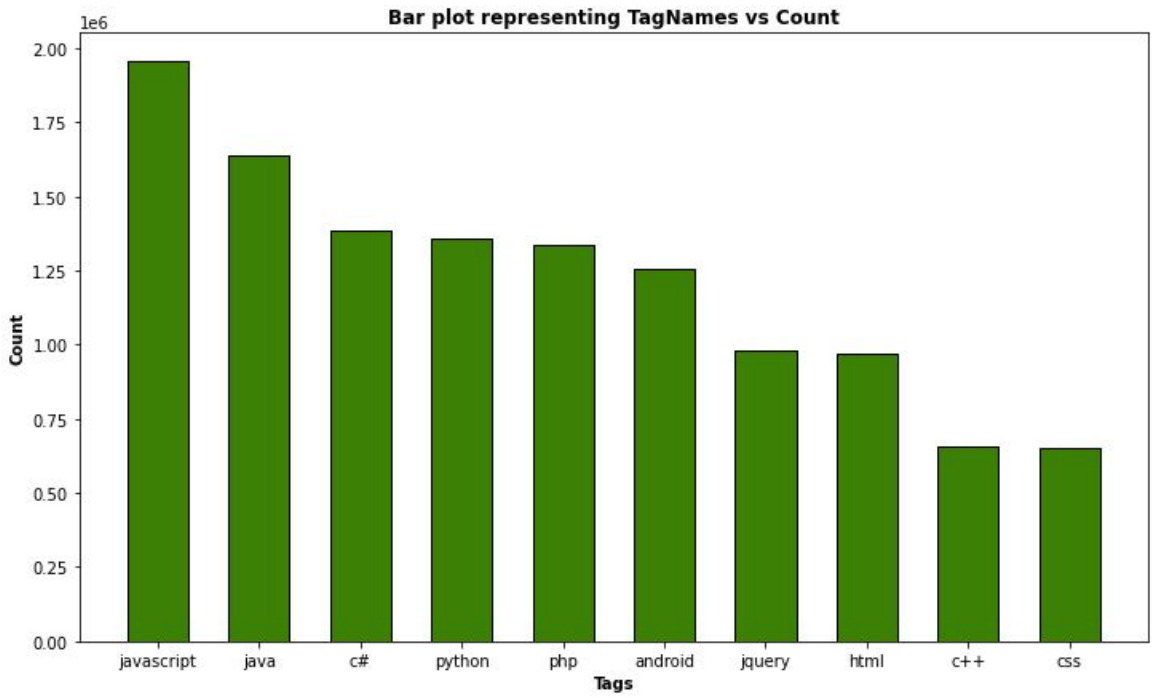
```
0                                                <python><xml>
1                                                          NaN
2                          <python><macos><fonts><photoshop>
3                                                          NaN
4                                                          NaN
                                  ...
299189                                    <python><eclipse>
299190                          <python><unit-testing><nose>
299191                                                     NaN
299192    <python><serialization><mongodb><flask><jinja2>
299193                                                 <python>
Name: Tags, Length: 299194, dtype: object
```

- **Distribution of Scores field for posts in Posts.xml**

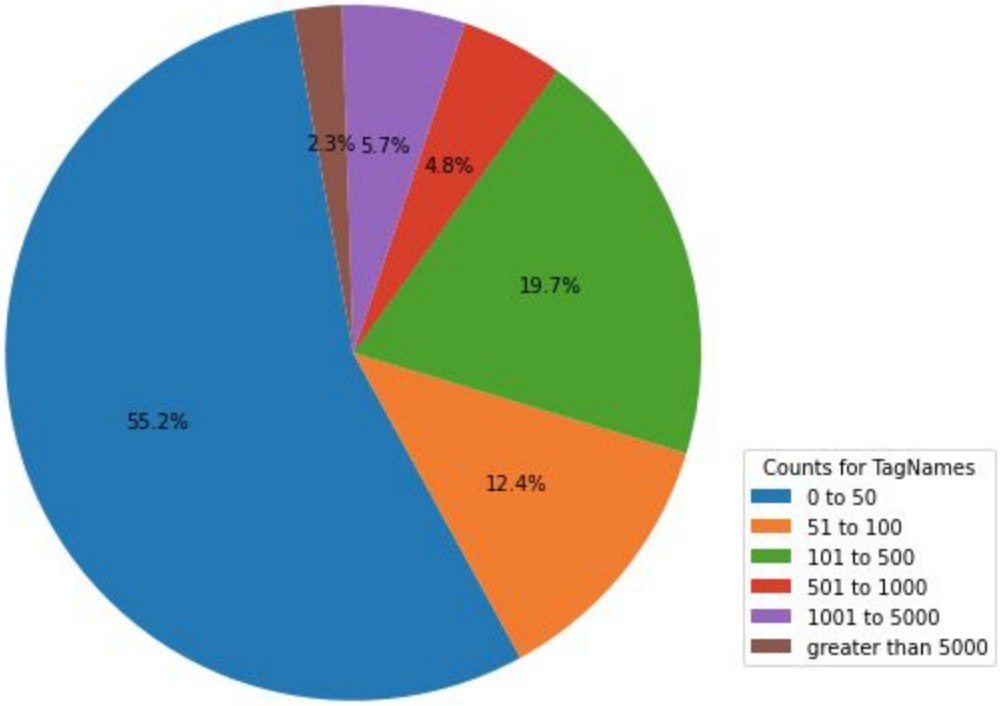


- **Top 10 TagNames in Tags.xml**

```
[[1955557, 'javascript'],
 [1641102, 'java'],
 [1385220, 'c#'],
 [1359126, 'python'],
 [1335050, 'php'],
 [1254482, 'android'],
 [978412, 'jquery'],
 [970699, 'html'],
 [656969, 'c++'],
 [649436, 'css']]
```



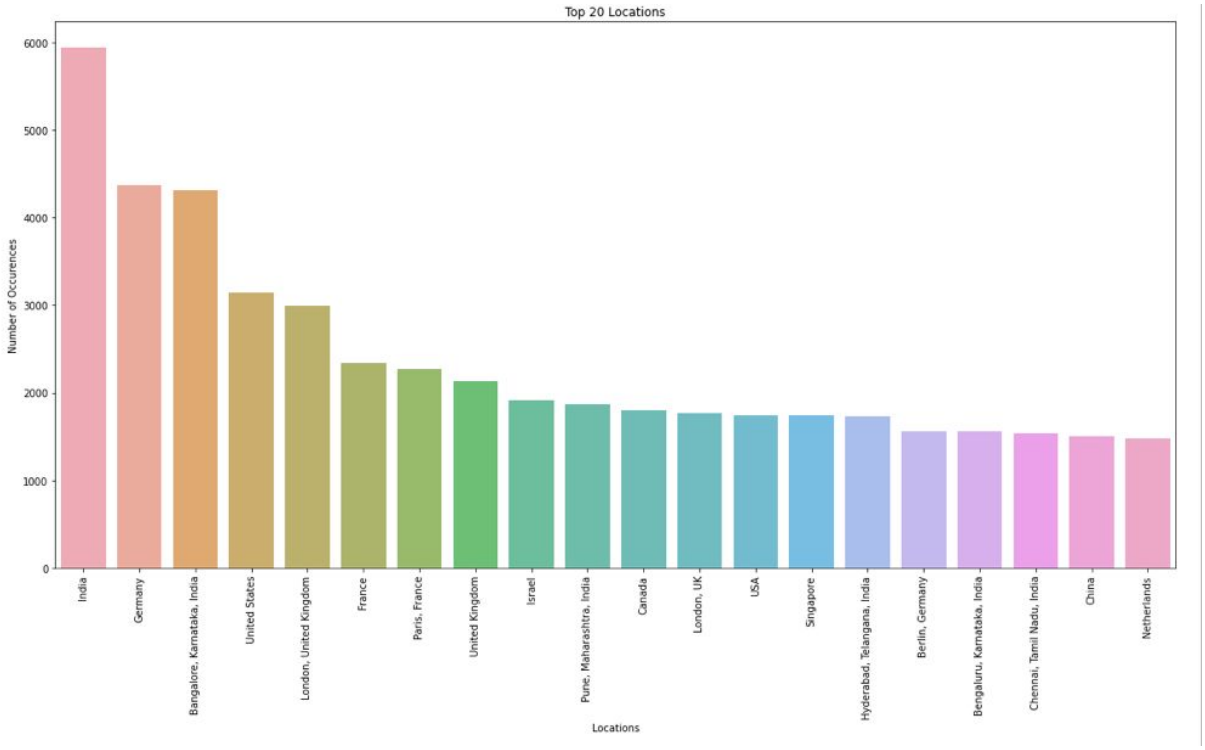
- **Counts for TagNames in Tags.xml**



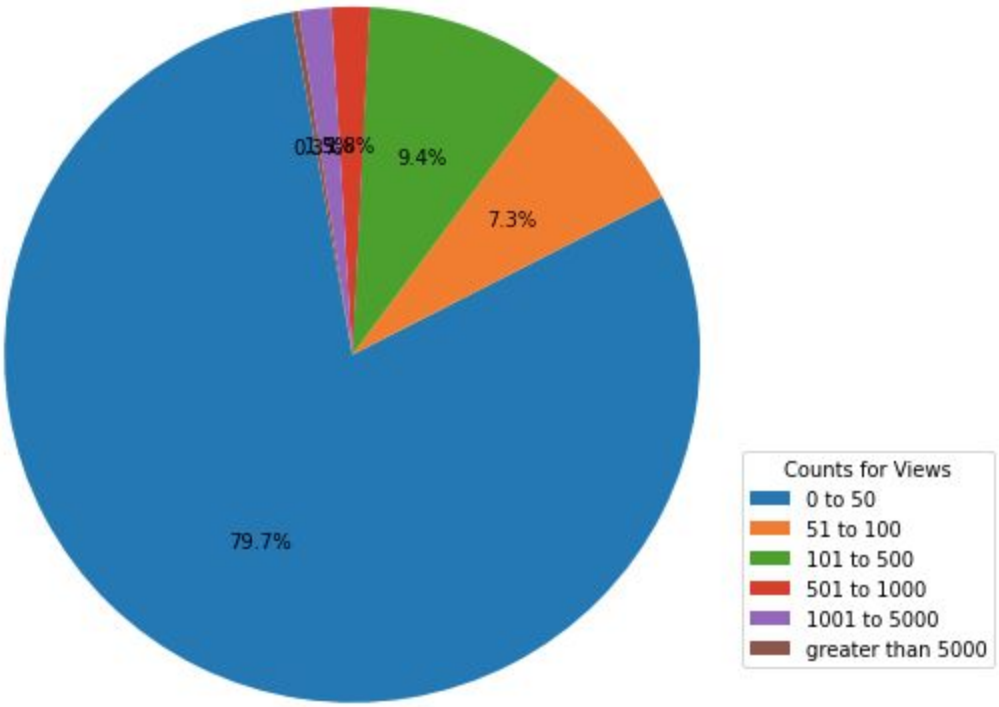
- **Locations for Users in Users.xml**

```
df["Location"].value_counts()
```

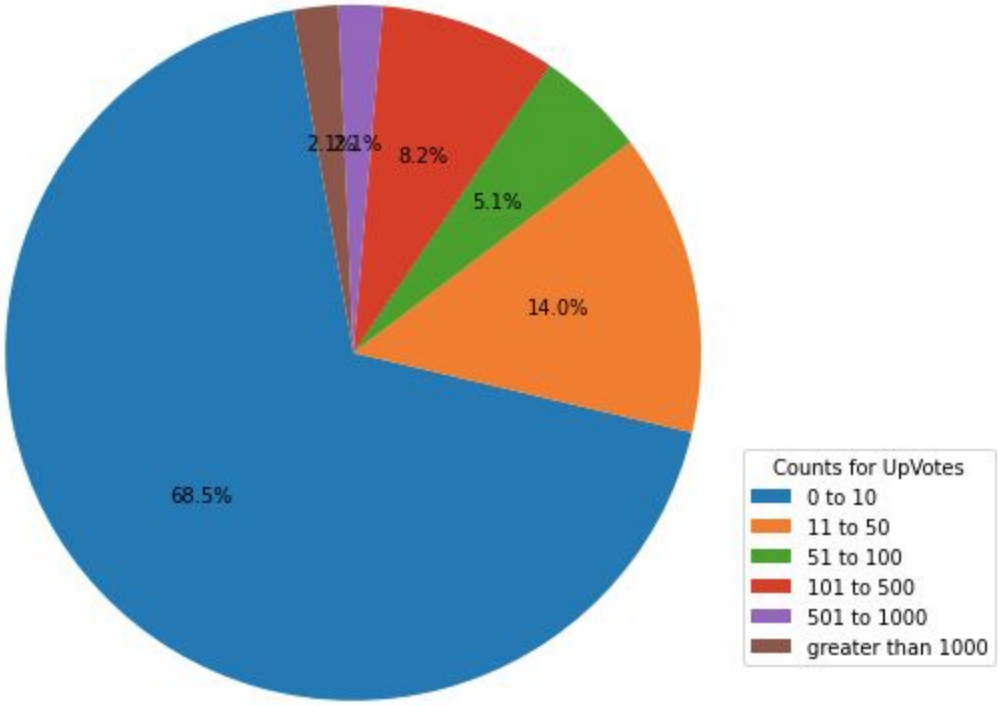
```
                              7051
India                         5937
Germany                       4365
Bangalore, Karnataka, India   4309
United States                 3145
                              ...
Udupi                            1
Isle of Wight, Newport, UK       1
@m c0w                           1
Argenbühl, Germany               1
Jasper, AL                       1
Name: Location, Length: 30275, dtype: int64
```



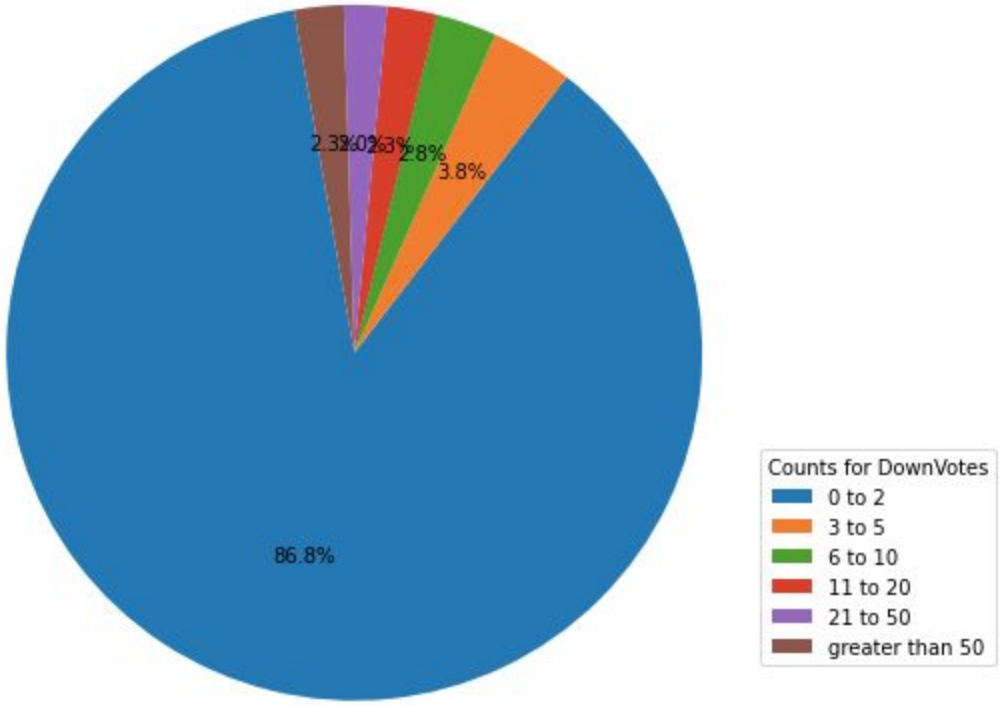
- **Counts for Views for Users in Users.xml**



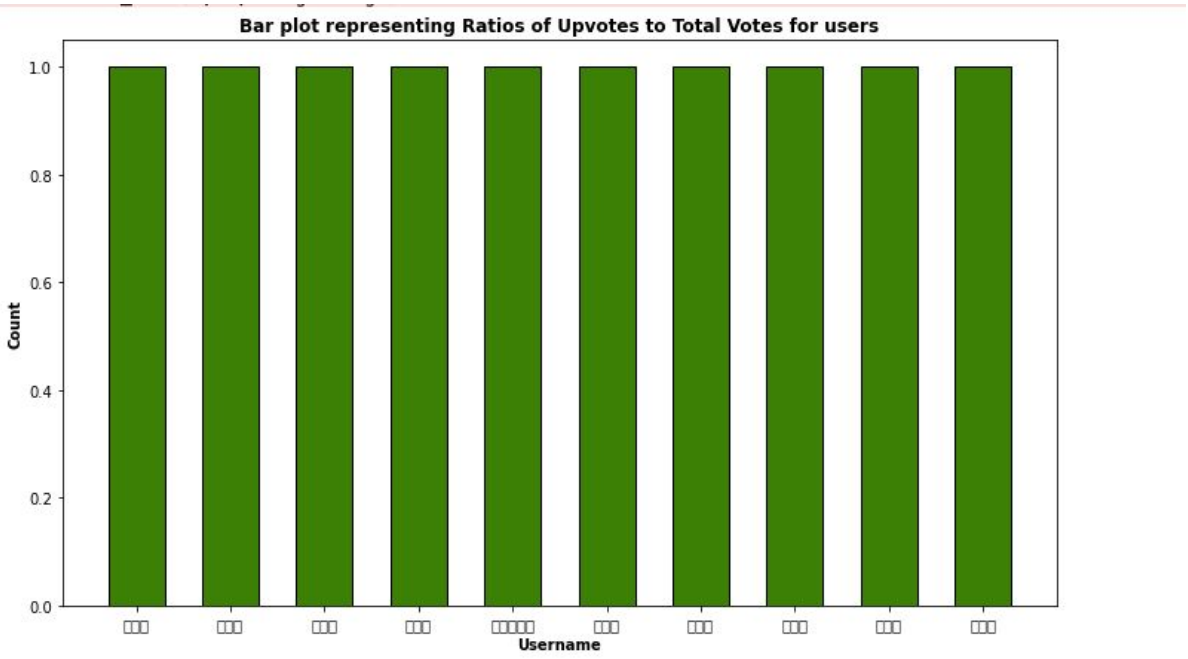
- **Counts for UpVotes for Users in Users.xml**



- **Counts for DownVotes for Users in Users.xml**



- **UpVotes to Total votes Ratios for Users in Users.xml**



Top and Worst 10 ratios for Upvotes:Total votes

```
l[:10]
```

```
[[1.0, '황준필'],
 [1.0, '황인규'],
 [1.0, '황민욱'],
 [1.0, '한여진'],
 [1.0, '파워뿡뿡이'],
 [1.0, '최희원'],
 [1.0, '최혜선'],
 [1.0, '최연석'],
 [1.0, '진주형'],
 [1.0, '주은혜']]
```

```
l[-11:-1]
```

```
[[0.0, 'Alex'],
 [0.0, 'Alec'],
 [0.0, 'Alban A. - SonarSource Team'],
 [0.0, 'Al-waleed Shihadeh'],
 [0.0, 'Ajjo'],
 [0.0, 'Ahmedou'],
 [0.0, 'Abhijit'],
 [0.0, 'Aaron Cyrman'],
 [0.0, 'ALGOholic'],
 [0.0, '24x7servermanagement']]
```