



Project Outline

Prepared for: Jason Richard Evans (13032608)

Prepared by: Vivian Venter (13238435)

Available At: GitHub - <https://github.com/thepickpocket/COS720-Securities>

25 April 2016

PROJECT OUTLINE

Project Outline

The project is build in Python (Version 2.7.11) using MongoDB (Version 3.2.5) as the database. Our database consists of the following:

DESCRIPTION	AMOUNT
Total Records	4764733
Total Unique Profiles	6848

Dependancies

- plotly - For creating graphs.
- pytagcloud - For creating word clouds.
- HTMLParser - Helps with cleaning the data.
- pytz - Dependency of used library.
- pygame - Dependency of used library.
- pymongo - To create a mongodb client in python.
- simplejson - Dependency of used library.

Preparations

Before any data could be read into the database, the encoding of the file needed to be changed from Latin1 to UTF-8 using the following linux command:

```
iconv -f latin1 -t utf-8 data.txt > utf_data.txt
```

After encoding the data in the right format, we needed to escape the quotation characters correctly as MongoDB needs it, so instead of escaping with a backslash (\") we escaped with a quotation mark (")

Project Implementation

Data Cleaning Techniques Used

- HTML Character Escaping
 - Removal of non-printable characters
 - Lowercase everything
 - Removal of links
 - Removal of @ Mentions
 - Removal of Stopping words
 - Removal of punctuation
-

Word Clouds Generated

- User submitted content (Tweet Content)

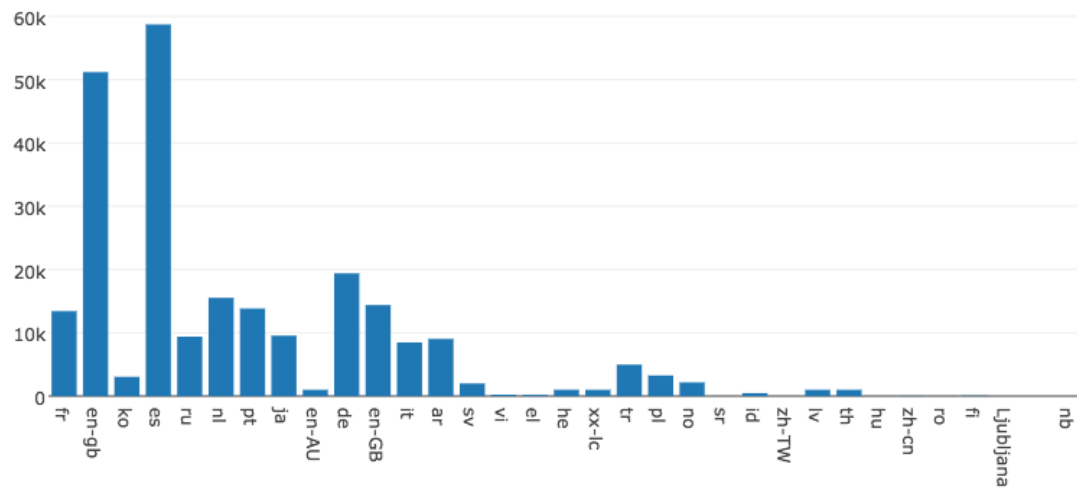


-

- Languages used to tweet

This bar chart displays the number of tweets for various languages supported by Twitter. The y-axis represents the 'Number of Tweets in Language' in millions, ranging from 0 to 4.5M. The x-axis lists the 'Twitter Supported Language' codes. English (en) is the most prevalent language, with over 4.5 million tweets. Other languages like French (fr), German (de), and Spanish (es) follow with significantly lower counts, all below 0.5 million. The majority of other languages have very low tweet counts, near zero.

Twitter Supported Language	Number of Tweets (approx.)
en	4.5M+
fr	~0.1M
en-gb	~0.05M
ko	~0.02M
es	~0.05M
ru	~0.01M
nl	~0.01M
pt	~0.01M
ja	~0.01M
en-AU	~0.01M
de	~0.02M
en-GB	~0.01M
it	~0.01M
ar	~0.01M
sv	~0.01M
vi	~0.01M
el	~0.01M
he	~0.01M
xx-ic	~0.01M
tr	~0.01M
pl	~0.01M
no	~0.01M
sr	~0.01M
id	~0.01M
zh-TW	~0.01M
lv	~0.01M
th	~0.01M
hu	~0.01M
zh-cn	~0.01M
ro	~0.01M
fi	~0.01M
luján	~0.01M
nb	~0.01M



- Whether people enabled location sharing or not

Percentage of Users Sharing/Hiding Location When Tweeting

