GROUP 14

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

GROUP 14

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 - Dependancy of used library.
- pygame - Dependancy of used library.
- pymongo - To create a mongodb client in python.
- simplejson - Dependancy 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

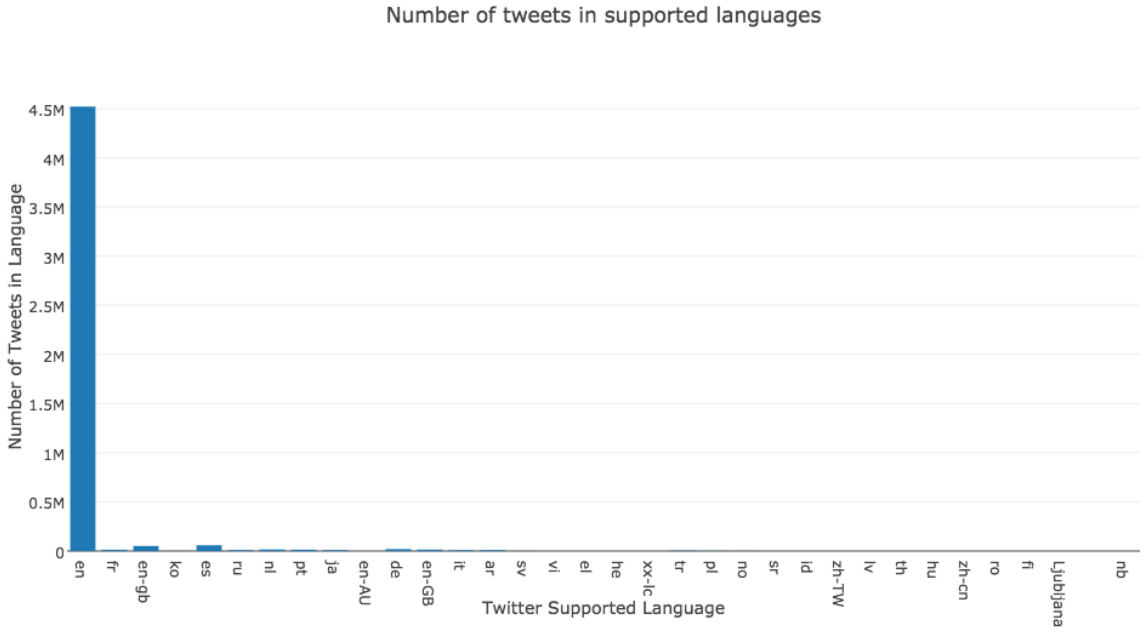
Word Clouds Generated

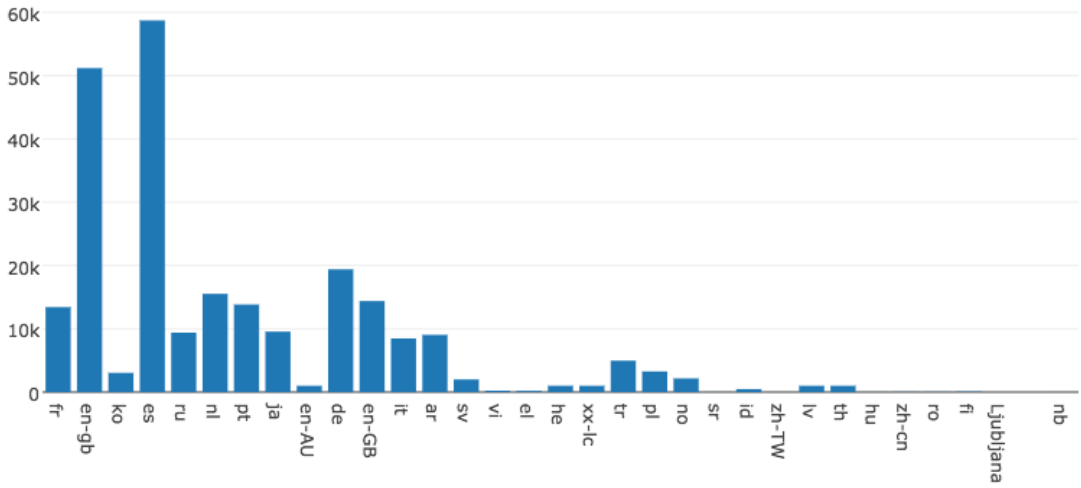
- User submitted content (Tweet Content)

- Locations

Statistics Generated

- Languages used to tweet





- Whether people enabled location sharing or not