



Project Outline

Prepared for: COS 720, Department of Computer Science, University of Pretoria

Prepared by: 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

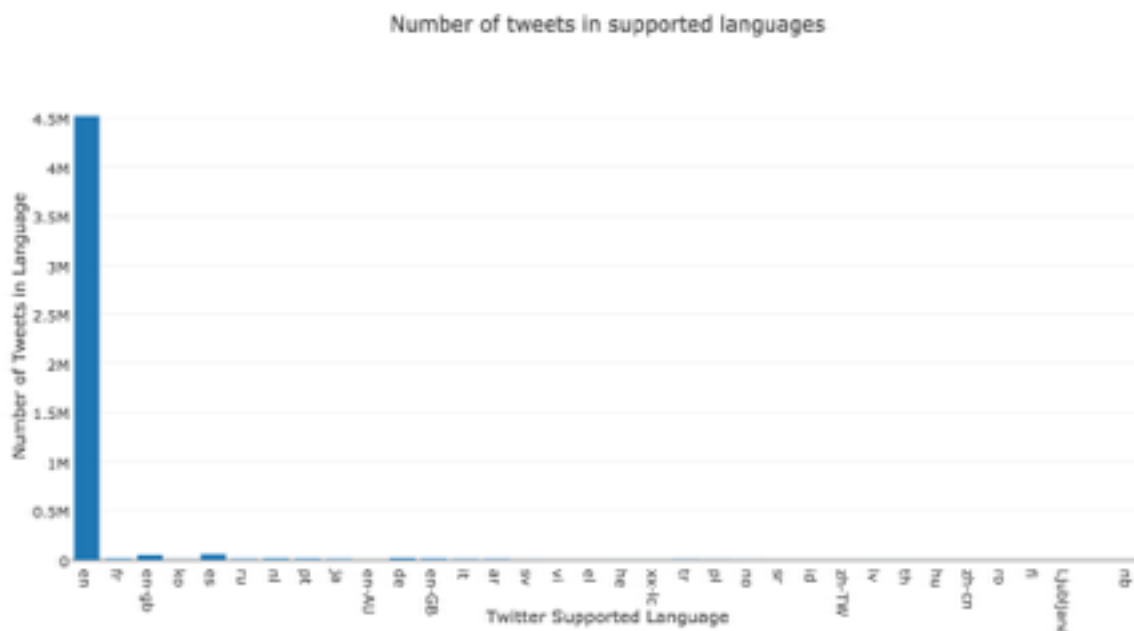
- HTML Character Escaping
 - Removal of non-printable characters
 - Lowercase everything
 - Removal of links
 - Removal of @ Mentions
 - Removal of Stopping words
 - Removal of punctuation
-

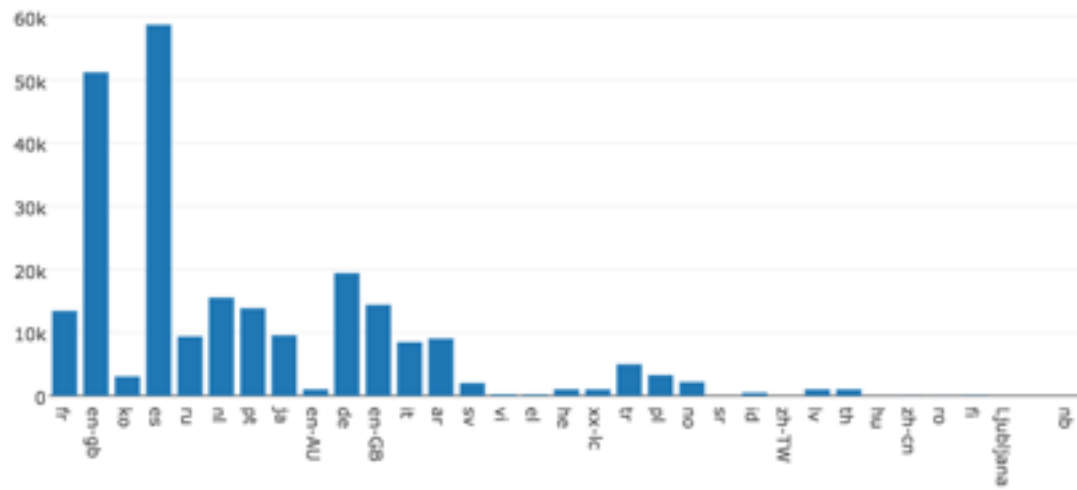
Word Clouds Generated

- User submitted content (Tweet Content)



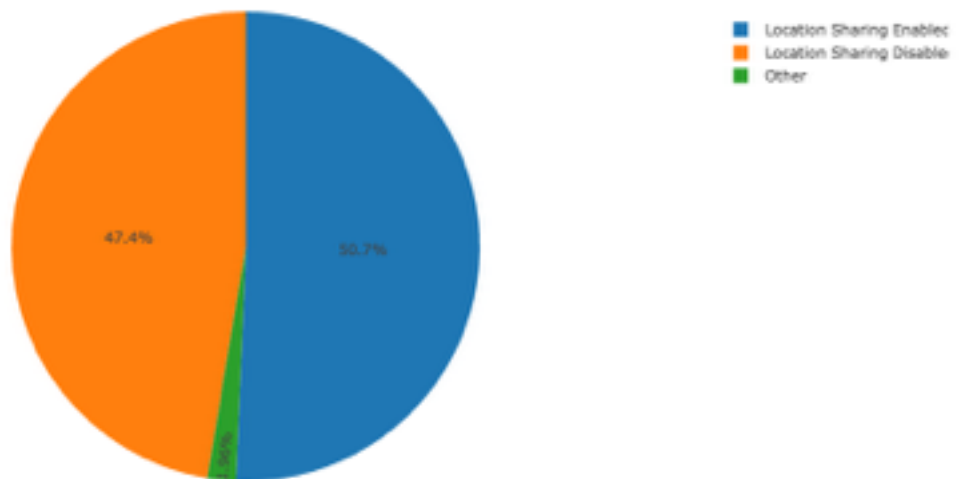
- Languages used to tweet



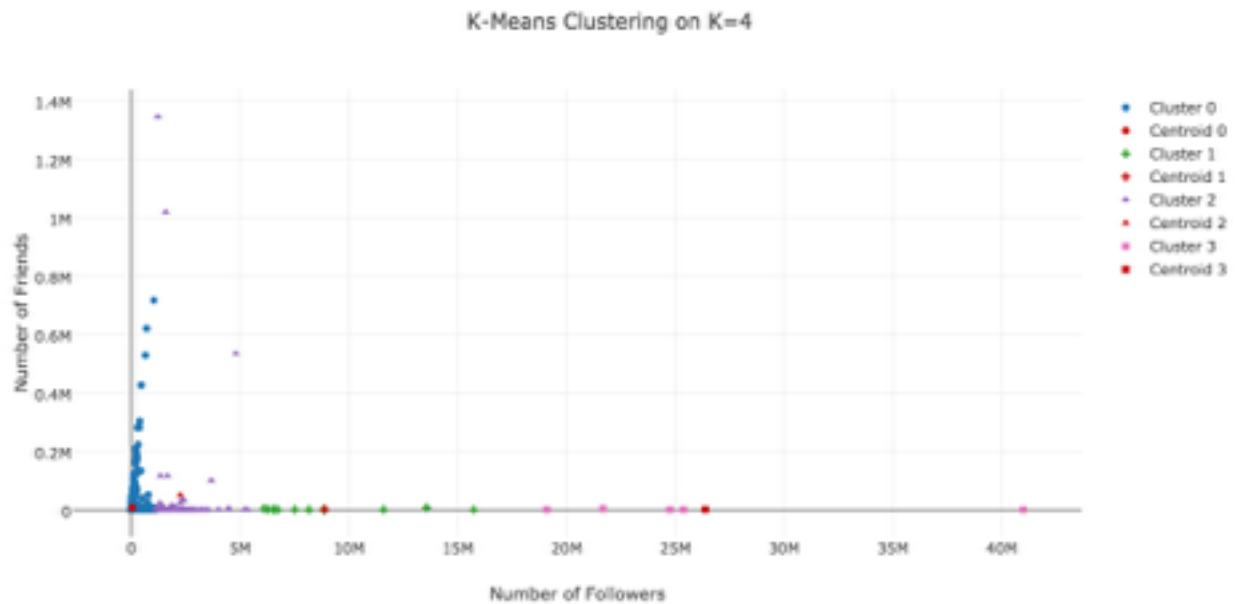
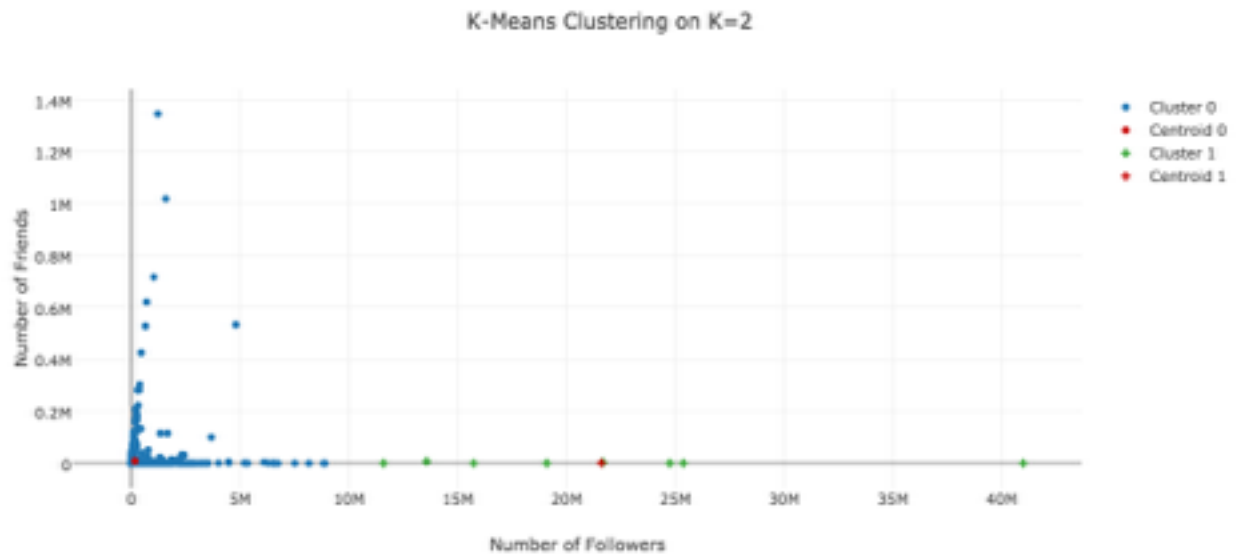


- Whether people enabled location sharing or not

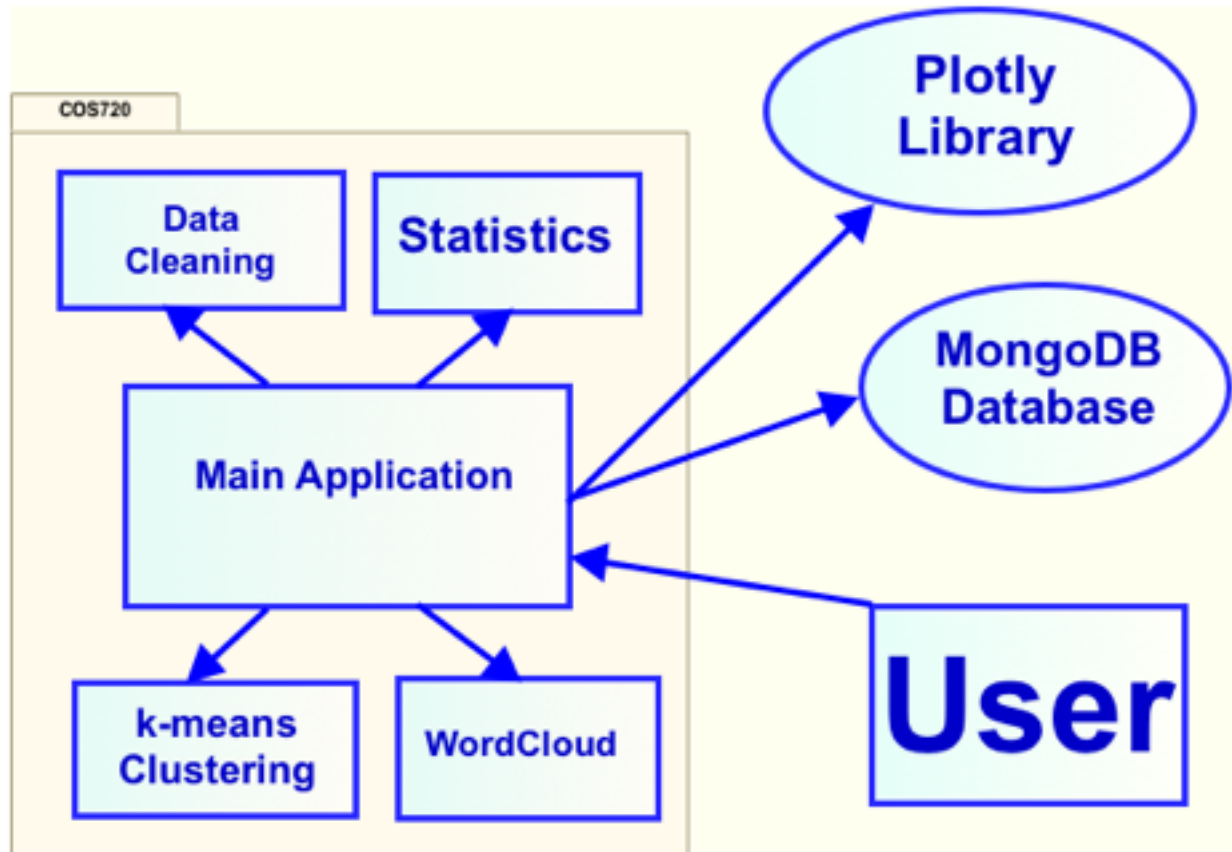
Percentage of Users Sharing/Hiding Location When Tweeting



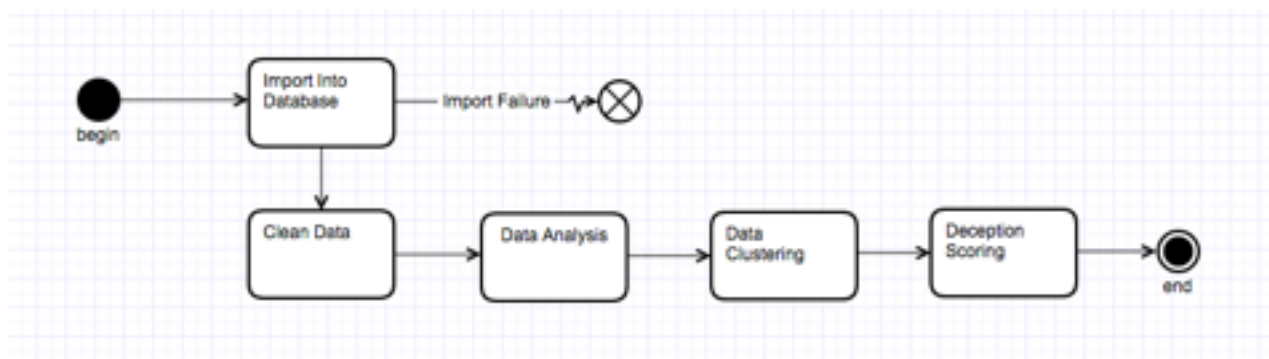
K-Means Clustering of the twitter data



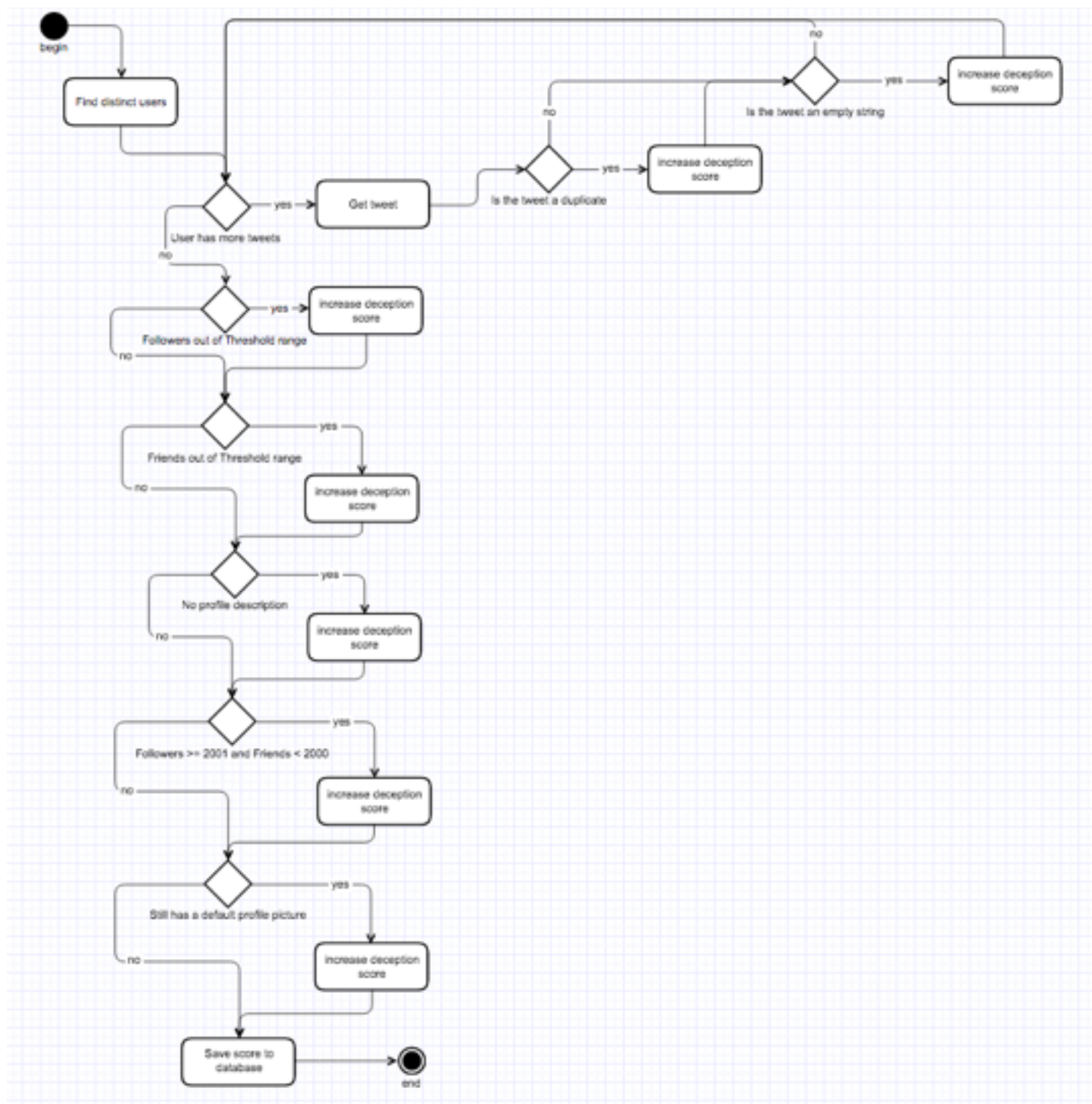
High Level System Design



UML Activity Diagrams




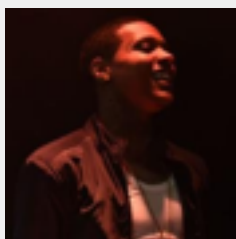
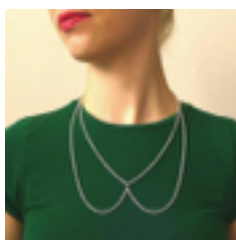
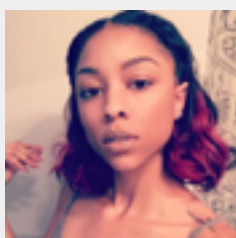




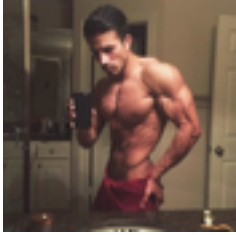



Overall



Deception Scoring

Output of deception scoring

Profile	Username	Following	Followers	DeceptionScore
	wolfpupy	204	164832	80%
	stumptowncoffee	1313	61008	80%
	ju	211	29179	87%
	LILDURK2x	0	311	85%
	madamegodot	864	2185	82%
	Jade	640	2077	64%

Profile	Username	Following	Followers	DeceptionScore
	14luisml	675	5635	73%
	ubuntudev	73	24764	80%
	Guzmanfitness	29	50968	76%
	beastdw	76	1201605	90%
	adidasUprising	243	13376	80%
	Jahborne	1221	32963	64%