Pravar D Mahajan
mahajan.89@osu.edu

Social Media & Text Analytics – Assignment 1

This is a report on the first assignment for the course of Social Media & Text Analytics. The assignment is based on using Twitter's streaming API to answer some basic questions related to the distribution of languages used by users and the users' locations.

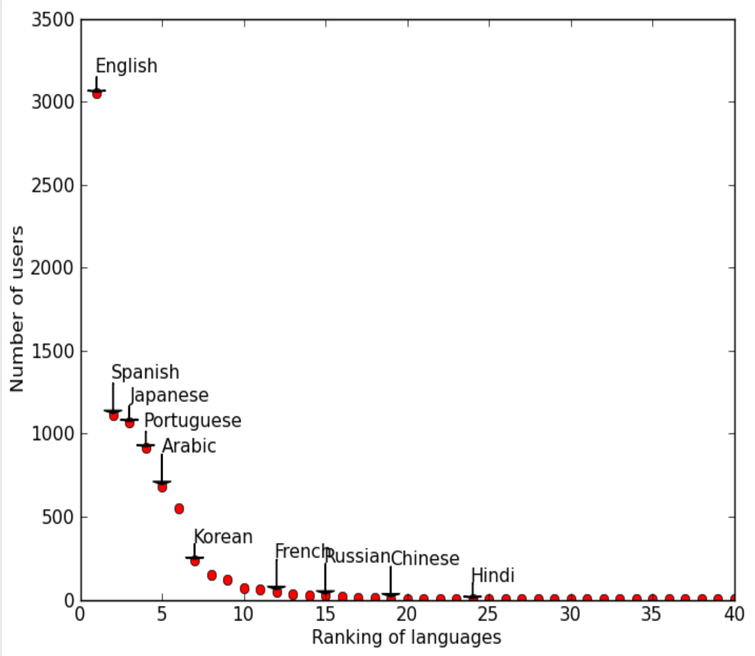
The first step of the assignment was to stream ~10,000 tweets using the twitter's streaming API. The streamed tweets were saved into file `twitter_data_file.dat`

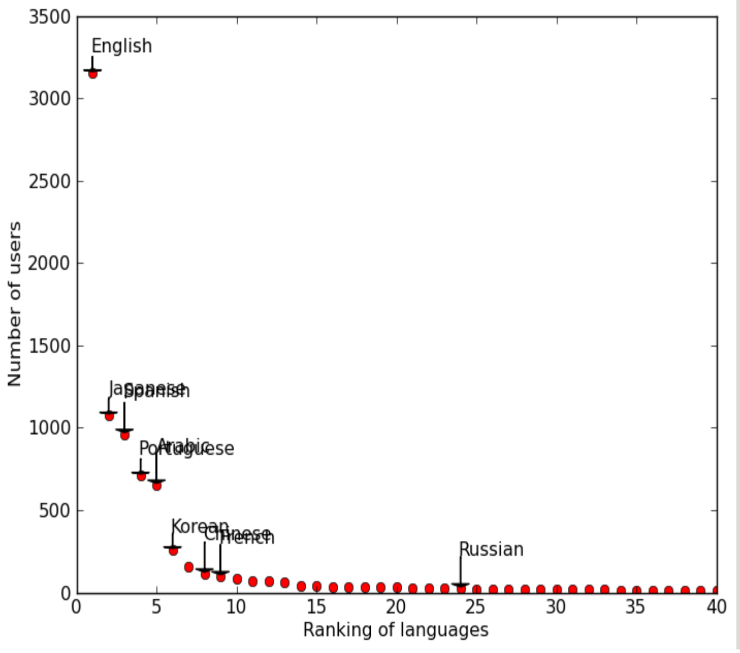
The next step in the assignment was to perform analysis on the `lang` tag provided by the twitter API. The tag is used to get the language of the tweet as identified by twitter. Not all tweets get `lang`-tagged. Tweets corresponding to user actions (such as deletion of a tweet), do not get `lang`-tagged. However, all the text tweets (identified as having `text` field in the tweet data) do have lang have lang tags. The results of the distribution of languages have been provided in the table at the end.

A comparison was made between twitter's language tag and the language labels identified by an off the shelf library for tagging language for a text – langid. This library requires a text as input and gives corresponding language label as output. The input was given from the text field present in the tweet's data structure, with some pre-processing. The processing involved removing some noisy data – the text "RT" was dropped if it was in the beginning of the tweet, and hashtag references (identified as "#" followed by one or more alphanumeric characters or underscores) and references to the users (identified as "@" followed by alphanumeric characters and underscores) were removed. The result of the language tagging via this library has been summarized in the table at the end.

One major difference between langid's tagging and twitter's tagging is that langid identifies a lot of minority languages as well, like Oriya, Assamese, Welsh, Czech etc, whereas twitter tags them as "unidentified" (und). For the major languages, like English, Spanish, French, both the methods work equally well.

Included below is a plot of the distribution of number of users for each language, as identified by twitter (on the left) and langid (on the right).

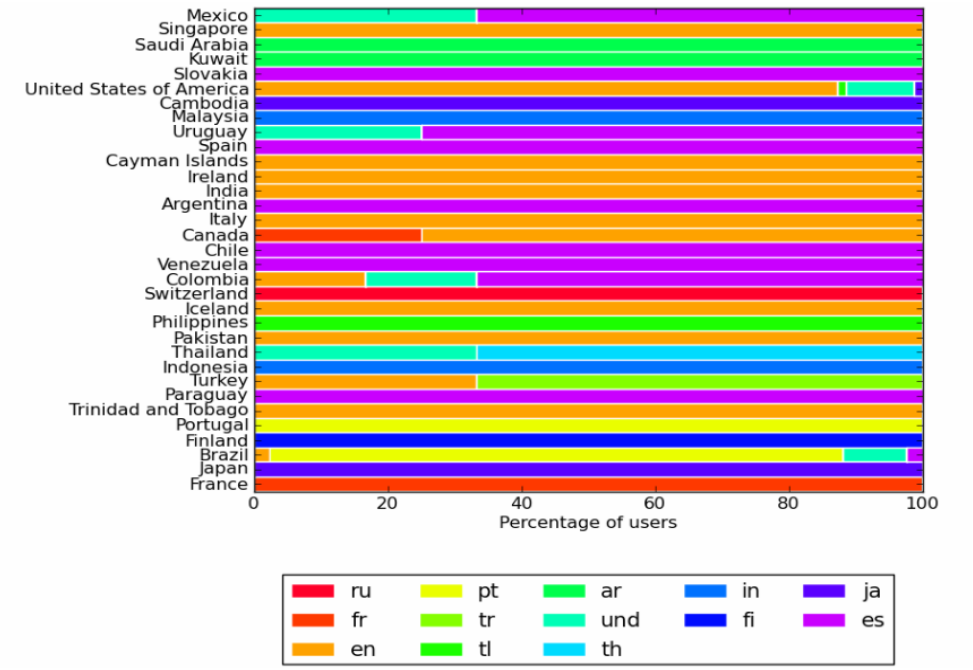




Tweets originating in the US were identified via the `country_code` tag in the `place` field in the tweet's data structure. Here's the distribution of languages of tweets in US. Distribution of languages across different countries of the world can be visualized in the form of bar graph on the right.

Language distribution in US

Language	**% of Tweets**
English	87.34
Undefined	10.12
Tagalog	1.27
Japanese	1.27



Since the `geo` field has been deprecated, we found very few tweets to be "geo-tagged". For US based tweets, only 6% of the tweets were geo tagged. The field which is used now instead, `coordinates`, was present in all the tweets having `place` information.

Instructions to run the code:

(1) If twitter data is already present in the json format, rename the file to twitter_data_file.dat and execute:

```
$> python main.py
```

(2) If twitter data is not present in the required format or if the data needs to be refreshed, then the script `main.py` can be edited for this purpose. Setting `refresh = True` in main.py (line 8) will stream new tweets.

(3) Plots are saved in the `plots` directory.

Table on the next page

Tag	# Twitter	# Langid
Afrikaans	0	7
Amharic	0	19
Aragonese	0	15
Arabic	683	653
Assamese	0	1
Azerbaijani	0	2
Byelorussian	0	1
Bulgarian	0	2
Bengali (Bangla	0	16
Breton	0	4
Bosnian	0	1
Catalan	0	29
Czech	3	32
Welsh	1	7
Danish	3	25
German	19	68
Greek	2	7
English	3051	3156
Esperanto	0	16
Spanish	1111	961
Estonian	8	10
Basque	6	4
Farsi	1	32
Finnish	2	36
Faeroese	0	1
French	50	98
Irish	0	7
Galician	0	40
Hebrew	0	13
Hindi	4	7
Croatian	0	6
Haitian Creole	26	1
Hungarian	0	6
Armenian	0	1
Indonesian	0	81
Indonesian	123	0
Icelandic	4	4
Italian	34	71
Japanese	1069	1079
Javanese	0	8
Georgian	0	6
Cambodian	0	18
Korean	239	262
Kurdish	0	4
Latin	0	27
Luxembourgish	0	5
Laothian	1	3
Lithuanian	1	15
Latvian (Lettish)	1	7
Malagasy	0	2
Macedonian	0	2
Malayalam	1	7
Marathi	0	2
Malay	0	21
Maltese	0	12
Norwegian Bokr	0	1
Nepali	3	1
Dutch	8	33
nn(unknown iso	0	1
Norwegian	4	9
Occitan	0	4
Oriya	0	1
Punjabi	0	2
Polish	2	39
Pashto (Pushto	0	7
Portuguese	916	714
Quechua	0	4
Romanian	2	8
Russian	24	24
Kinyarwanda (Ru	0	4
Northern Sami	0	3
Sinhalese	0	6
Slovak	0	12
Slovenian	1	13
Albanian	0	3
Swedish	7	19
Swahili (Kiswahili	0	5
Tamil	2	6
Telugu	0	1
Thai	148	154
Tagalog	67	35
Turkish	61	60
Uighur	0	5
Ukrainian	0	1
Undefined	552	0
Urdu	1	21
Vietnamese	3	13
Volapuk	0	2
Wallon	0	2
Xhosa	0	1
Chinese	7	114
Zulu	0	3