Software Design Mini Project #1
Anderson Ang Wei Jian
Project Writeup

Project Overview

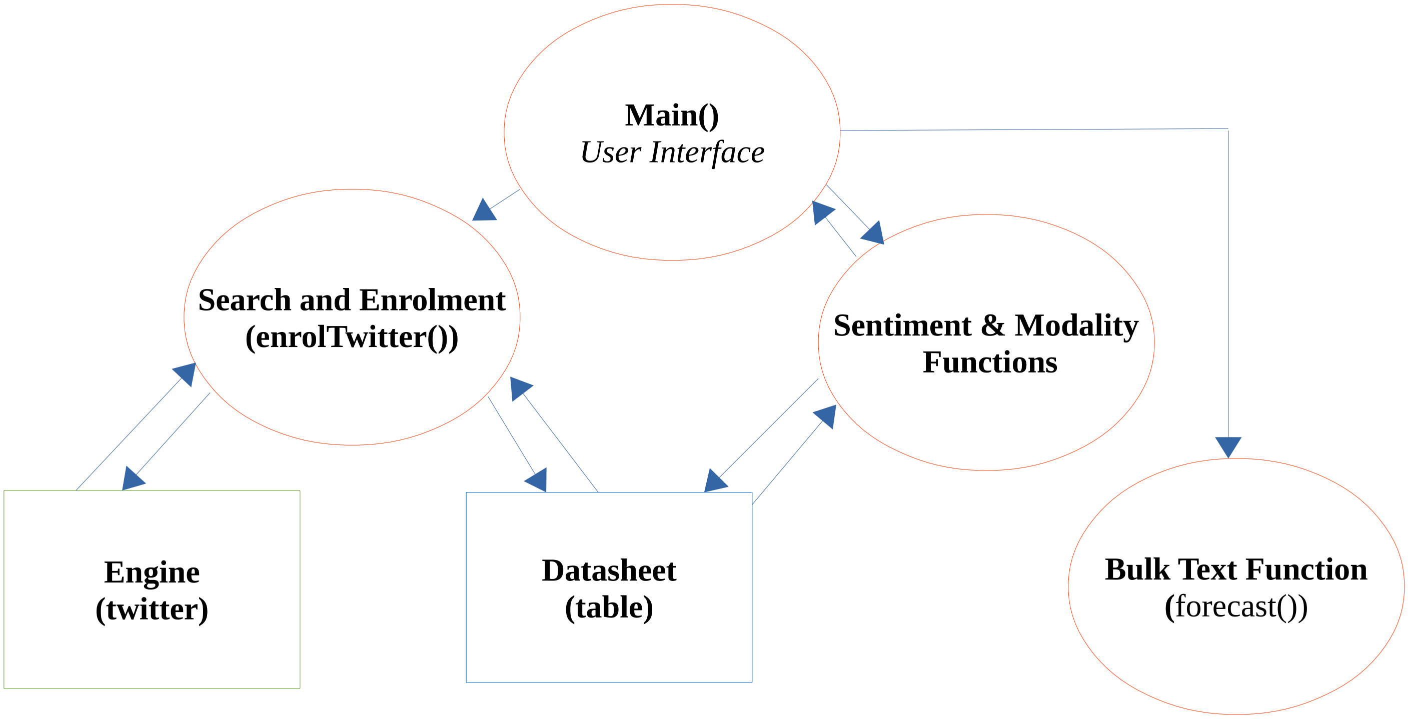
For this project, I used Twitter as my primary data source – Pattern.web supports multiple Twitter related functionalities – such as *streams(), trends(), and search().*

Fundamentally, I wanted to work on comparing opinions between two geolocations in world on any given subject – *whether they view it positively, more so than the other, if* the feed had an opinion tilt, and the level of facuality in the region's tweets on the said subject.

This involves compiling a certain amount (variable and fully controllable, save for Twitter's internal query limits) of input feed based on a given *search term, enrolling them in a database, then lifting them for analysis and post-mortem work.*

Implementation

The main() function serves as the hierarchical master of the program – all critical function calls are made from it, and it takes minimal data, only involving the values which we want to acquire (sentiment, modality, and subjectivity data). It also doubles as a UI preface, since the program is text-based in nature.



One of the key functions is the enrolment function – enrolTwitter() - it is responsible for accessing the search engine (pattern.search), using two range values to control the amount of feedback it requires, trimming the content of the tweets, and then enrolling them into a CSV (comma-separated values) file.

The function also implements four key features to enhance to usability of the data:

- Two range parameters to control the (I) stream count and (II) depth of stream history respectively
- Disabling local caching, to ensure new feeds come in at every refresh
- Removing Twitter-related handles to improve sentiment analysis performance (removed @, RT, #, and http/s related insertions)
- **Registering a unique tweet.id for each unique tweet, to prevent multiple registrations of the same tweet.**

One of the **critical design decisions** was whether to include retweets as part of the statistical analysis – since the content was essentially the same, the inclusion of such tweets from the return stream would skew the data in the same orientation. However, I decided to leave them included, as I realized that a RT represents agreement by a unique individual, and thus it should be considered as a veritable opinion/fact.

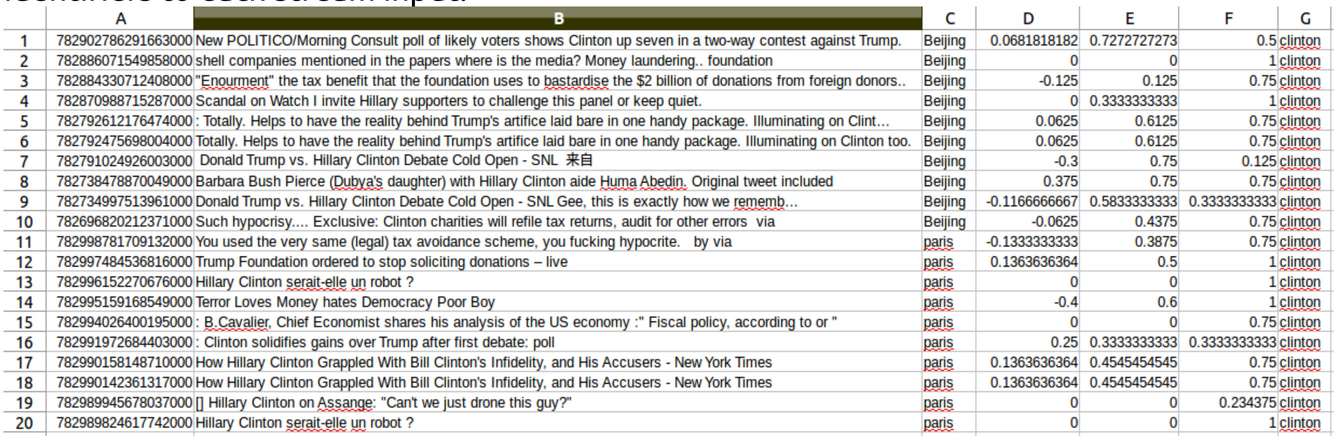
The sentiment and modality functions separately access the CSV archive, and due to way it is built, it is able to utilize content from past and present searches involving the **same term, to generate a more holistic content.**

Additionally, due to the strict requirements for UID (unique identifier) and term matching, the CSV dumping is able to omit streams from unrelated topics or countries.

Finally, the bulk text function explains the significance of the resultant data from the two locales to the user in a communicable manner (*an attempt to reproduce an UIUX orientation here)*

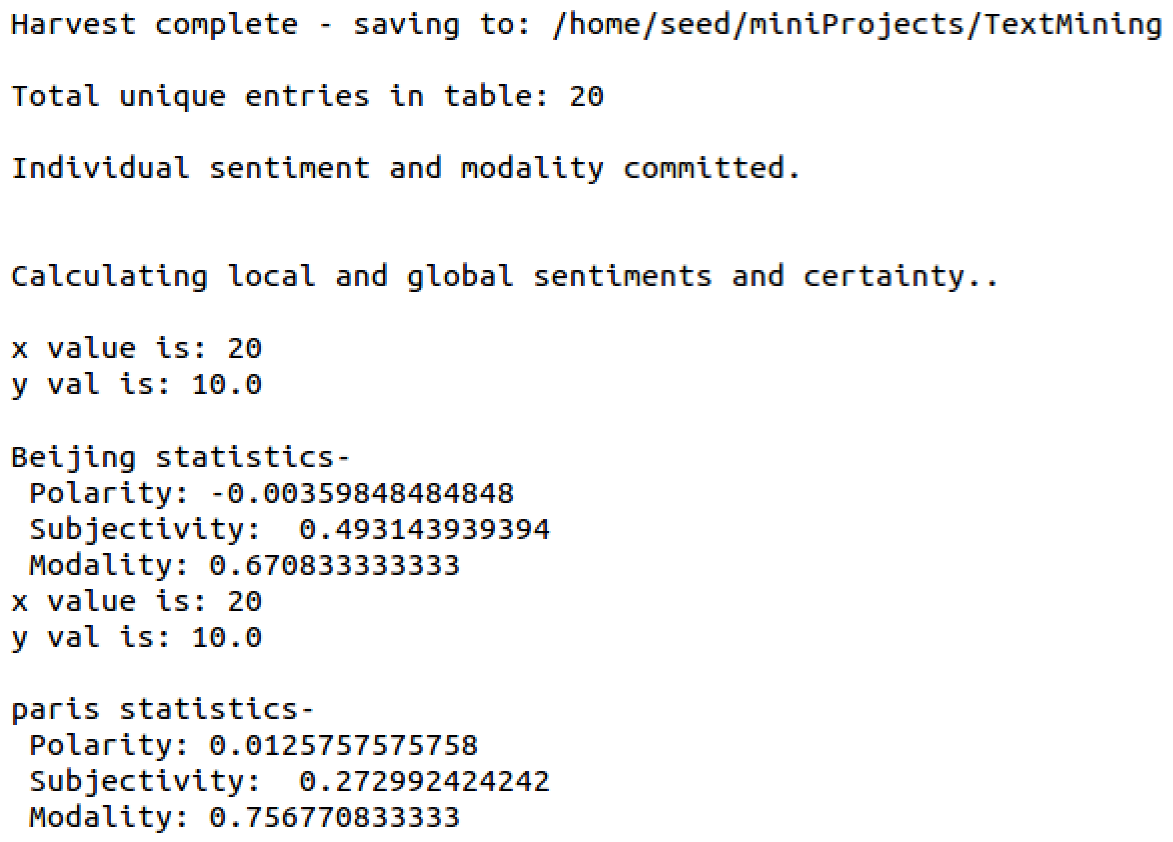
Results

As seen below, the program provides a CSV dump with a few categorical and unique identifiers to each stream input:



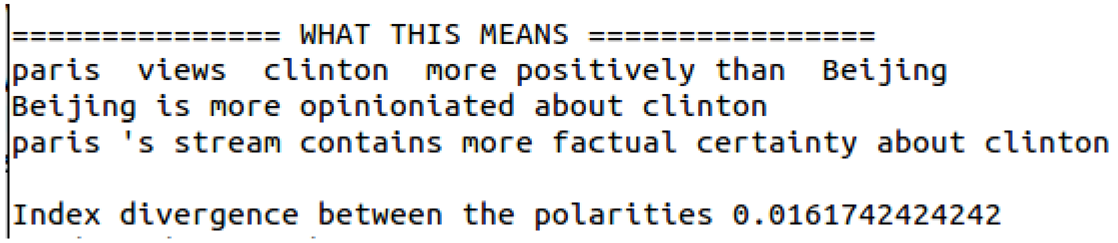
	A	B	C	D	E	F	G
1	782902786291663000	New POLITICO/Morning Consult poll of likely voters shows Clinton up seven in a two-way contest against Trump.	Beijing	0.0681818182	0.7272727273	0.5	clinton
2	782886071549858000	shell companies mentioned in the papers where is the media? Money laundering.. foundation	Beijing	0	0	1	clinton
3	782884330712408000	"Enourment" the tax benefit that the foundation uses to bastardise the $2 billion of donations from foreign donors..	Beijing	-0.125	0.125	0.75	clinton
4	782870988715287000	Scandal on Watch I invite Hillary supporters to challenge this panel or keep quiet.	Beijing	0	0.3333333333	1	clinton
5	782792612176474000	: Totally. Helps to have the reality behind Trump's artifice laid bare in one handy package. Illuminating on Clint…	Beijing	0.0625	0.6125	0.75	clinton
6	782792475698004000	Totally. Helps to have the reality behind Trump's artifice laid bare in one handy package. Illuminating on Clinton too.	Beijing	0.0625	0.6125	0.75	clinton
7	782791024926003000	Donald Trump vs. Hillary Clinton Debate Cold Open - SNL 来自	Beijing	-0.3	0.75	0.125	clinton
8	782738478870049000	Barbara Bush Pierce (Dubya's daughter) with Hillary Clinton aide Huma Abedin. Original tweet included	Beijing	0.375	0.75	0.75	clinton
9	782734997513961000	Donald Trump vs. Hillary Clinton Debate Cold Open - SNL Gee, this is exactly how we rememb…	Beijing	-0.1166666667	0.5833333333	0.3333333333	clinton
10	782696820212371000	Such hypocrisy.... Exclusive: Clinton charities will refile tax returns, audit for other errors via	Beijing	-0.0625	0.4375	0.75	clinton
11	782998781709132000	You used the very same (legal) tax avoidance scheme, you fucking hypocrite. by via	paris	-0.1333333333	0.3875	0.75	clinton
12	782997484536816000	Trump Foundation ordered to stop soliciting donations – live	paris	0.1363636364	0.5	1	clinton
13	782996152270676000	Hillary Clinton serait-elle un robot ?	paris	0	0	1	clinton
14	782995159168549000	Terror Loves Money hates Democracy Poor Boy	paris	-0.4	0.6	1	clinton
15	782994026400195000	: B.Cavalier, Chief Economist shares his analysis of the US economy :" Fiscal policy, according to or "	paris	0	0	0.75	clinton
16	782991972684403000	: Clinton solidifies gains over Trump after first debate: poll	paris	0.25	0.3333333333	0.3333333333	clinton
17	782990158148710000	How Hillary Clinton Grappled With Bill Clinton's Infidelity, and His Accusers - New York Times	paris	0.1363636364	0.4545454545	0.75	clinton
18	782990142361317000	How Hillary Clinton Grappled With Bill Clinton's Infidelity, and His Accusers - New York Times	paris	0.1363636364	0.4545454545	0.75	clinton
19	782989945678037000	[] Hillary Clinton on Assange: "Can't we just drone this guy?"	paris	0	0	0.234375	clinton
20	782989824617742000	Hillary Clinton serait-elle un robot ?	paris	0	0	1	clinton

The program calculates each individual stream's sentiment values, and enrolls them alongside the original stream.



```
Harvest complete - saving to: /home/seed/miniProjects/TextMining

Total unique entries in table: 20

Individual sentiment and modality committed.


Calculating local and global sentiments and certainty..

x value is: 20
y val is: 10.0

Beijing statistics-
 Polarity: -0.00359848484848
 Subjectivity:  0.493143939394
 Modality: 0.670833333333
x value is: 20
y val is: 10.0

paris statistics-
 Polarity: 0.0125757575758
 Subjectivity:  0.272992424242
 Modality: 0.756770833333
```

It then provides the user with each locale's statistics concerning the topic searched, and the general polarity of opinion. In calculating the locale's compound statistics, the program assumes a fair/common weight for each individual stream committed – it also identifies the number of entries made regarding the topic in the region, for averaging.

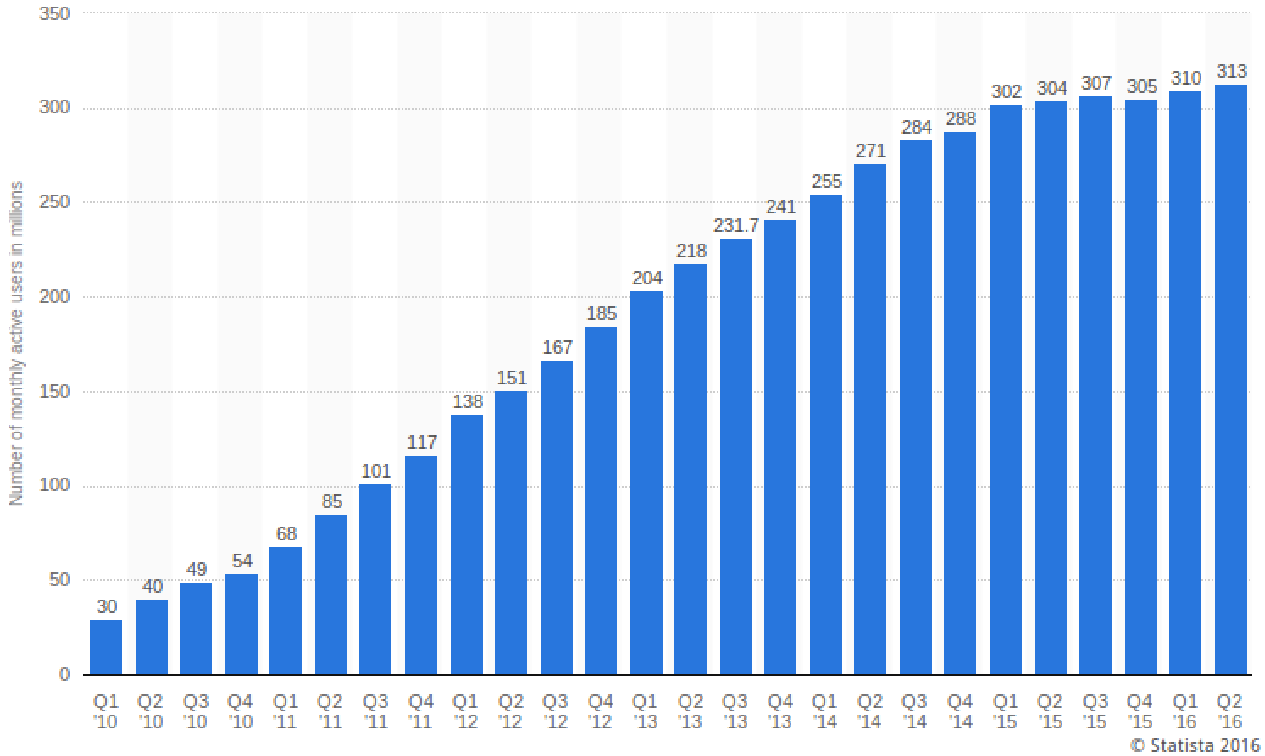


```
=============== WHAT THIS MEANS ================
paris  views  clinton  more positively than  Beijing
Beijing is more opinionated about clinton
paris 's stream contains more factual certainty about clinton

Index divergence between the polarities 0.0161742424242
```

Finally, the bulk text function takes over, and computes the absolute difference between the two locale's opinions of the specified topic. Since the value ranges between a limit of -1 to +1, any difference above 0.25 can be seen as a polarized result. It also interprets the consequences of the localized data, in terms of its polarity, subjectivity, and certainty/factuality.

Reflection
After running a couple of simulations involving geographically dispersed cities, what truly shocked me was the seemingly small stratification of opinion between them on many subjects – individually, opinions may seem deeply stratified, but taking into account the

amount of users on Twitter (313 million monthly active users (MAU)), a good representation of a locale's opinion can be registered on a credible whole.



© Statista 2016

Maybe it is the age of informational globalism after all – *regardless of the media sources one is locally exposed to*, the user is able to obtain information from sources beyond their *local reach, enabling them to arrive at a highly similar opinion to their regionally distant counterparts.*