Assignment 4 : CS 751

HARISH

May 2, 2015

0.1 Question 1

Using the pages from A3 that boilerpipe successfully processed, download those representations again and reprocess them with boilerpipe.
Document the time difference
Compute the Jaccard Distance x for each Pair of Pages for :Unique terms ,Bigrams,TriGrams
For each of 3 cases build a Cumulative Distribution Function that shows the percentage change on the x-axis and percentage of the population on y-axis
Give 3-4 examples illustrating the range of change that you have measured

Solution to Question1:
I have used jusText library to remove HTML templates in Assignment 3 and 4. It removes headers, footers and navigation links from the page.
jusText was successful for pages that followed HTML standards. Pages which are properly organized .

I have written a code in python named "htmlnonhtml.py" which is located in "q1-A4" folder. This code removes boilerplate content for the links present in "responseAfterCurlCalls" located in "q1-A4" folder.

Only 6100 URIs were unique , out of this only for 1496 URIs boilerpipe sucessfully processed . Most of them are non-HTML skipped because of non content , Images on links , 404, 302 errors etc.

I have used the same pages from A3 for which boilerpipe was successfully processed. I have reprocessed those pages with boilerpipe and extracted the text files again for A4

The time difference between A4 and A3 is

Time(A4)-Time(A3) : Time difference is 28 days March 30 to April 28
I have created two folders "A3-afterboilerplateTextFiles", "A4-afterboilerplatetextfiles" which is located inside "q1-A4" folder. These folders contains text files that were extracted after processing boilerpipe successfully for A3 and A4.

1

I have calculated Unique terms ,Bigrams , Trigrams for the text files located in the following folders :

1)"A3-afterboilerplateTextFiles" located inside "q1-A4" folder
2)"A4-afterboilerplatetextfiles" located inside "q1-A4" folder

I have written a java code named "UniGram.java","BiGram.java", "TriGram.java" located in "q1-A4" folder to extract unique terms, bigrams and trigrams respectively from the text files .

I wrote a "jacarrdDistanceCalculator.py" code located inside "q1-A4" folder . This code reads the textual data generated for each pair of pages from A3 and A4 and calculates the jaccard distance for unique terms , bigrams and trigrams and saves them to an output file named "unijackardsorted.txt" , "bijacardsorted", "trijacardsorted" respectively. These files are located inside the folder "q1-A4"

For each of the 3 cases (i.e., 1-,2-,3-grams) I built a cumulative distribution function that shows percentage change on the x-axis and the percentage of population on the y-axis.

Figure 1 ,Figure 2 , Figure 3 represents the Cumulative Distribution Function for Unigram , BiGram , TriGram respectively . I have plotted the graph using R script . BiJacard.R , UniJacard.R , TriJacard.R are the scripts for generating Cumulative Distribution Function graphs for Bigrams , Uni-Grams, Trigrams respectively . These are located inside "q1-A4" folder.

Following are the three examples illustrating the range of change measured

1) For this link "http://pinterest.com/pin/300896818829344973/" after successfully processing boilerpipe, there is a lot of change in textual content. Interesting part here is that, the unigram jaccard distance for this link is "0.958333333333" where as for Bigram and TriGram jaccard distance is "1"

2) For this link "http://suba.me" after successfully processing boiler-pipe , there is a little change in textual content. But, the jaccard distance for unigram , bigram and trigram are 0.0173611111111, 0.0138248847926, 0.0134831460674 respectively . Jaccard Distances are not same for unigram , bigram and trigram

3) For this link "http://twi.xsrv.jp/eot4j" after successfully processing boilerpipe, there is no change in the textual content. The Jaccard distances(1-gram, 2-gram, 3-gram) for these pair of pages from A3 and A4 is 0.

0.2 Question 2

from Q1 (A4), download all TimeMaps(including TimeMaps with 404 responses, i.e. empty or null TimeMaps)
Upload all TimeMaps to github
Build a CDF for number of mementos for each original URI i.e., x-axis = number of mementos , y-axis = number of links

Solution to Question2:
I have written a code "getURLData.py" which is located inside the "q2-A4" folder. This code reads "urlStatusAfterCurlCalls" file which contains the URIs. This is located in "q2-A4" .
I have used "http://labs.mementoweb.org/timemap/link/" to get the mementos for each URI. This codes saves all the TimeMaps for all the 10000 links. These are saved inside "q2-A4/timemaps/" folder .This list includes the null time maps as well.

After collecting the mementos for each URI , I have written a python script named "mementocount.py" which extracts the count of mementos per each URI and saves it to "timemapsCount.txt" file which is located inside "q2-A4" folder.

Figure 4 shows the cumulative distribution function for all the count of mementos. I have written a R script named "cdfformemntos.R" located inside "q2-A4". Most of the URIs have zero mementos.

0.3   Question 3

Using 20 links that have TimeMaps -
-With¿=20 mementos
-Have existed greater than equal to 2 years i.e., Memento-Datetime of "first memento" is April XX, 2013 or older
-Note : select from Q1/Q2 links , else choose them by hand
For each link , create a graph that shows Jaccard Distance, relative to the first memnto, through time
-x-axis : continuous time ,y-axis: Jaccard Distance relative to the first memento

Solution to Question3:
I have chosen 20 links by hand that satisfies the question criteria. I have saved these files into "final20links.txt" which is located inside "q3-A4" folder. I have written a code "getURLData.py" located inside "q3-A4" folder . This code reads "final20links.txt" and extracts the mementos links for each URI in the list in json format with the help of "http://labs.mementoweb.org/timemap/json/" .

I have saved all of these files with '.json' extension. I have written a python script with name "html-non-html.py" located in "q3-A4" folder which reads each json file in the directory ("q3-A4" folder) and iterates through each memento present in the list (json Array) for each URI and applies boilerpipe to each link. Most of the links were not successfully processed after applying boilerpipe. For the URIs for which boilerpipe successfully processed, I have calculated the Jaccard distance relative to first memento.

I have written a Rscript named "JDRelativetofirstMemento.R" which is located in "q3-A4" folder

Figure 5- Figure 18 shows Jaccard Distance relative to first memento.

0.4 Question 4

Choose a news-related event
UseTwarc.py to collect 1000 tweets , every day for 5 different days
For each day:
- Create a wall
- Build a tag/word cloud for each day
- Create a map using GeoJSON and Github
Discuss in detail lessons learned , experiences , etc

Solution to Question4: (Figure 19 to Figure 33)

Firstly, I have generated a consumer key, consumer secret key, access token and access-secret token.
I have collected 1000 tweets, every day for 5 different days.
I have written a script(using above mentioned keys) named "twarc1.py" located inside folder "q4-A4". I have extracted the tweets on "AppleWatch" tag. I have saved these 1000 tweets daily with ".json" extension .
The files are located inside the folder "q4-A4".Following are the files I have extracted on a daily basis for 5 different days

tweets25th.json
tweets26th.json
tweets27th.json
tweets28th.json
tweets30th.json

I have copied all of these inside "twarc-master/utils/". Then, I have given

above json files as an input to wall.py to create Wall on daily basis. It gave me output wall with '.html' extension. All of these output files are stored inside "q4-A4/wall/" folder. The wall basically gives the information of all the tweets that are related to "Applewatch" for each day.

I have given above files as an input to word.py to create word cloud on daily basis. It gave me output wordcloud files with '.html' extension. The files are located inside the folder "q4-A4/wordcloud/". Word Cloud mainly enlarges mostly used word from the tweets . The frequency of the words is directly proportional to the Size of the word visible in the output wordcloud file.

I have also given above json files as an input to tweets.geojson to get geo coordinates from the tweets. I have uploaded the geojson files located inside 'q4-A4/geojson/' .After uploading these geojson files , github shows the map with markers located using "geojson coordinates".

Cumulative Distribution Function
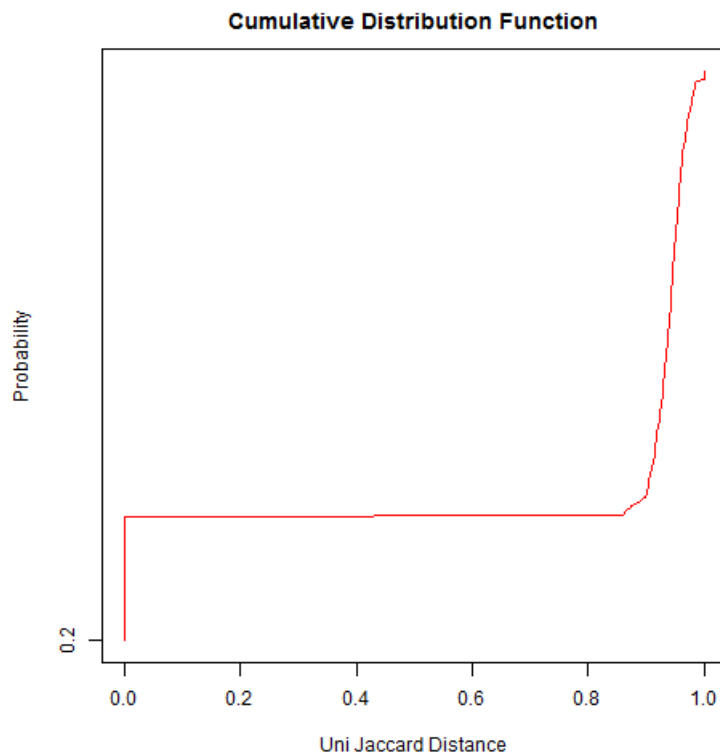
Figure 1: CDF for Unigram Jaccard Distance

6
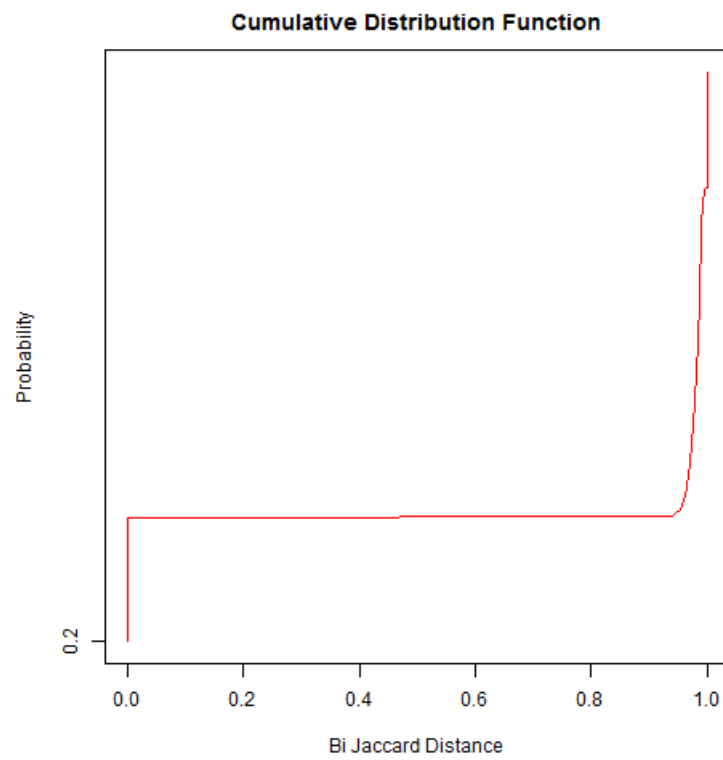
Figure 2: CDF for Bigram Jaccard Distance
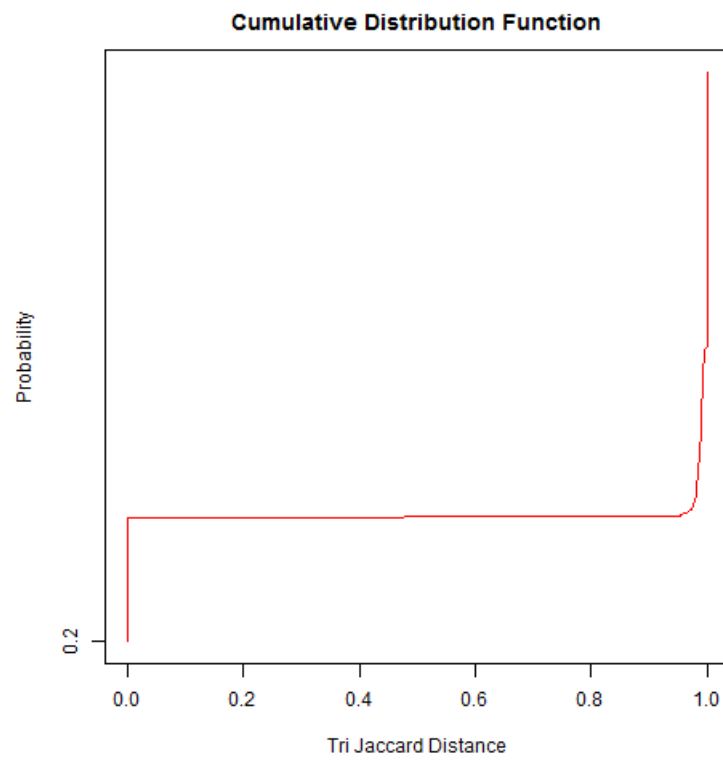
Figure 3: CDF for Trigram Jaccard Distance
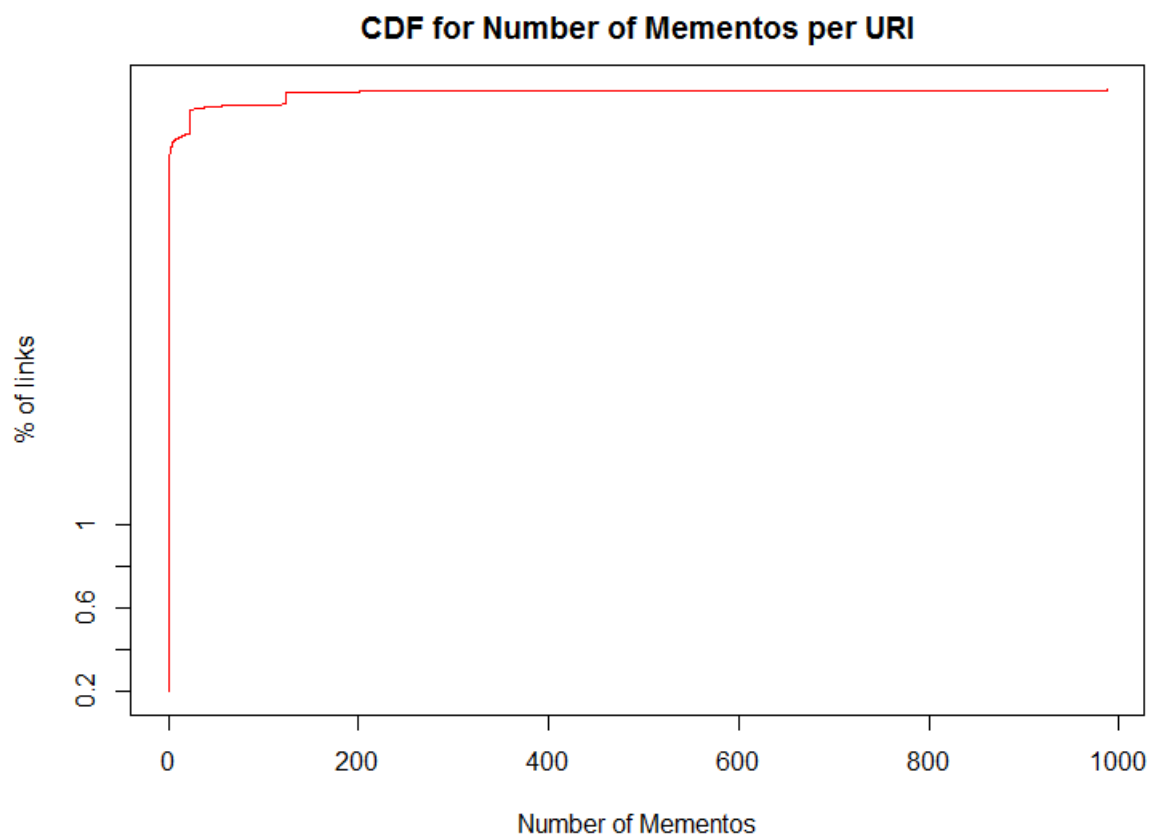
Figure 4: CDF for number of mementos for each original URI
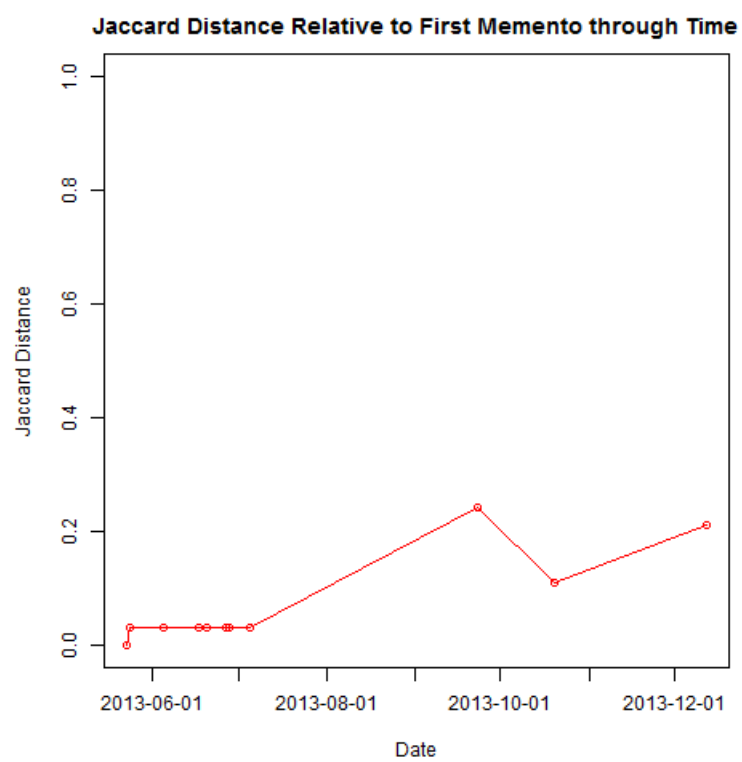
Figure 5: Jaccard Distance Relative to First Memento
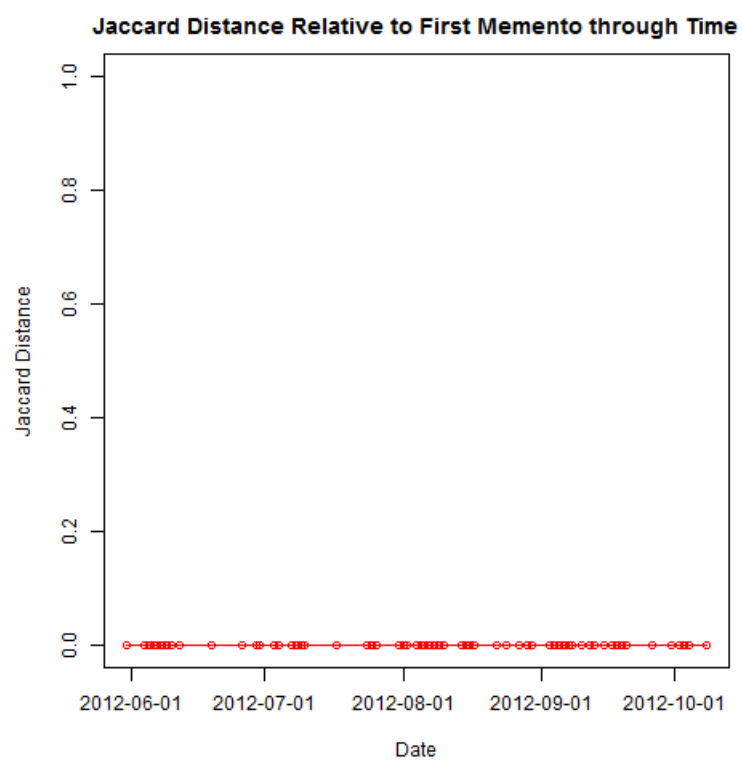
Figure 6: Jaccard Distance Relative to First Memento
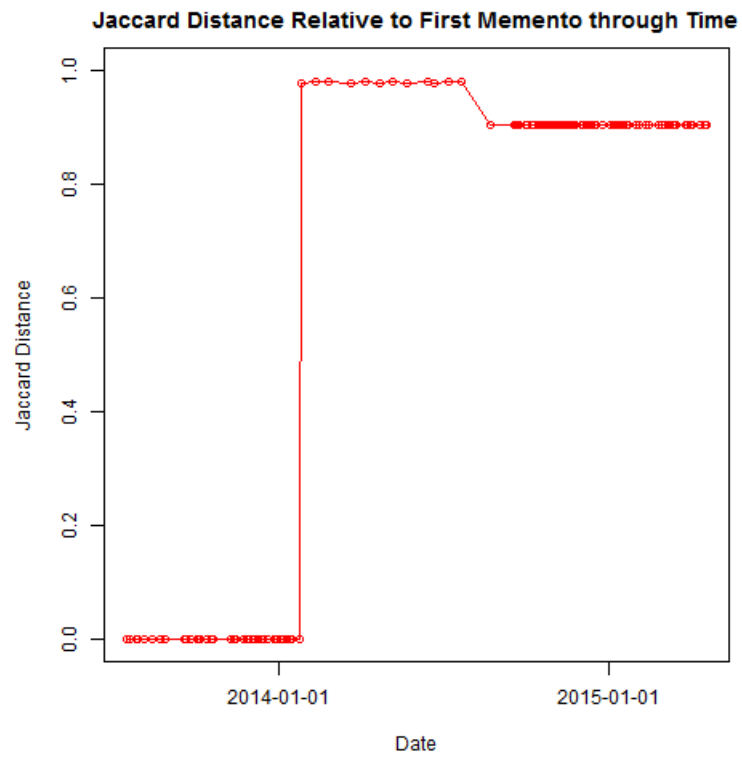
Figure 7: Jaccard Distance Relative to First Memento
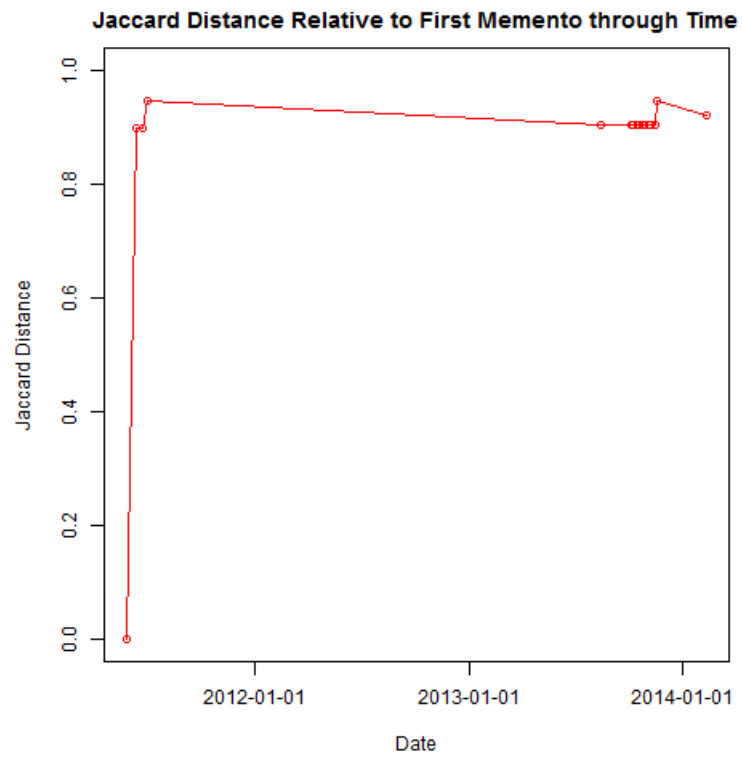
Figure 8: Jaccard Distance Relative to First Memento
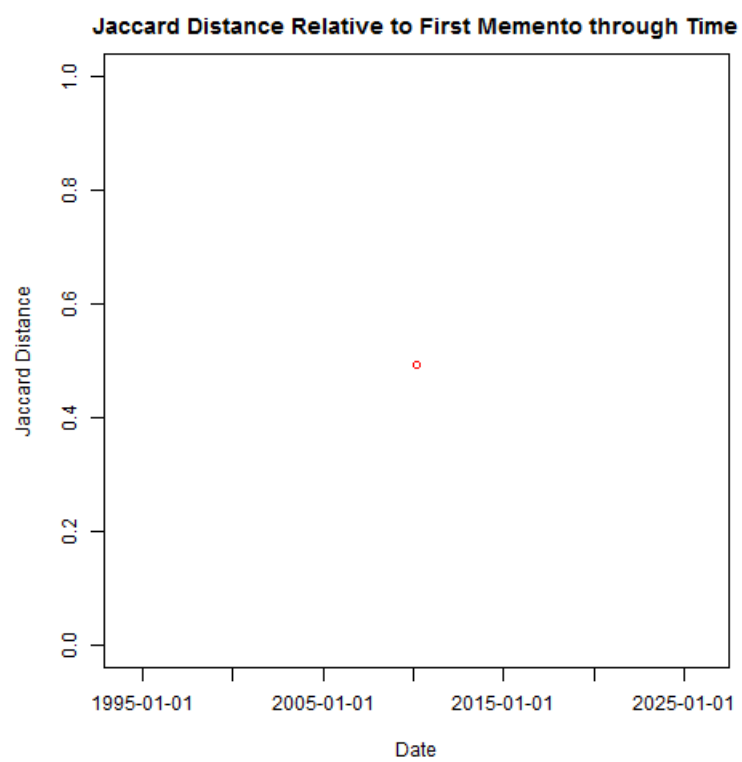
Figure 9: Jaccard Distance Relative to First Memento
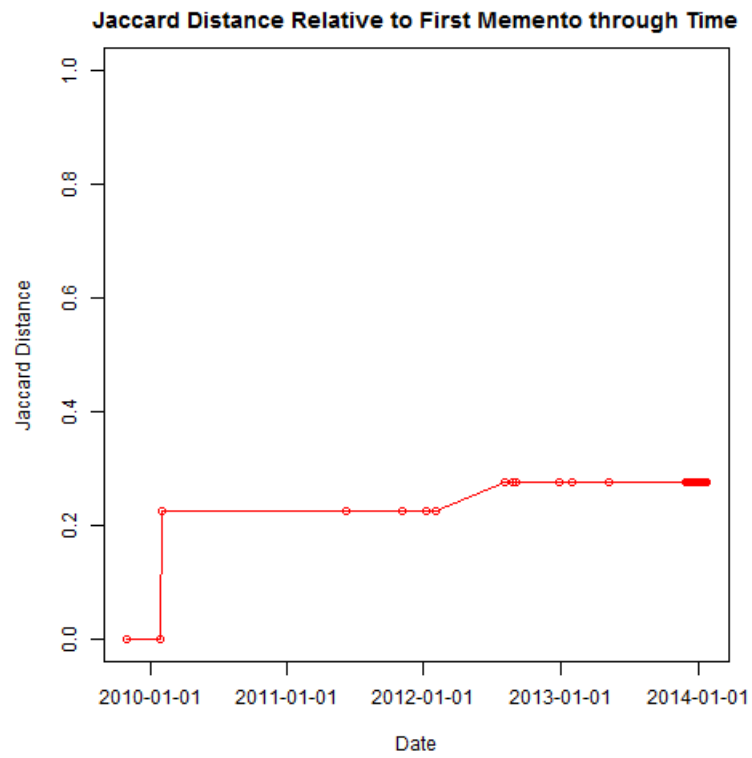
Figure 10: Jaccard Distance Relative to First Memento
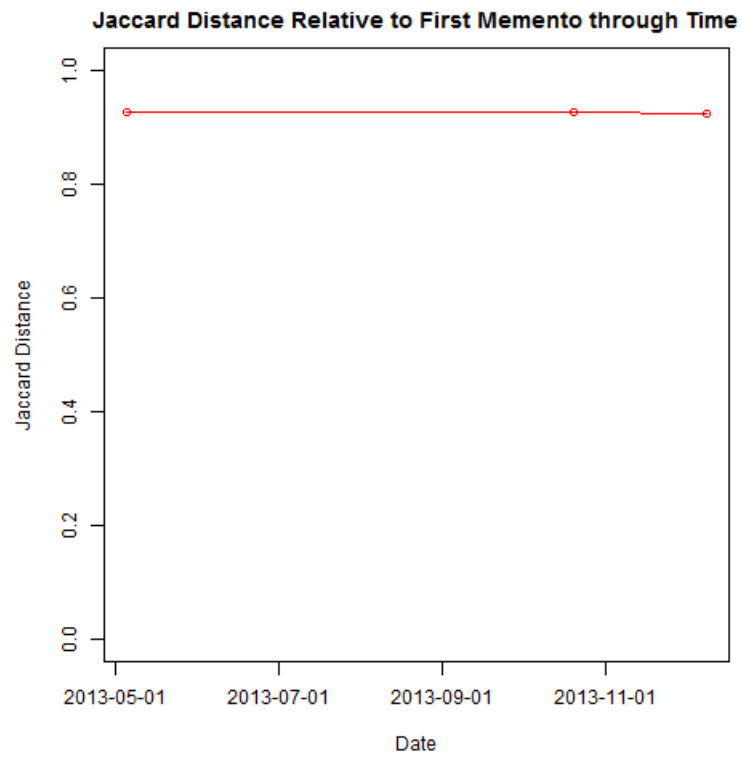
Figure 11: Jaccard Distance Relative to First Memento
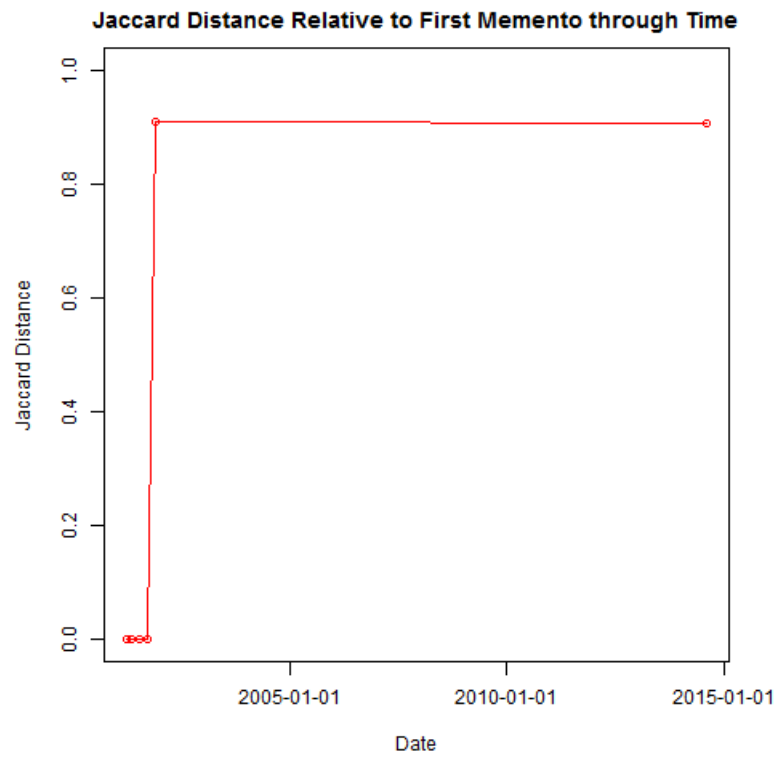
Figure 12: Jaccard Distance Relative to First Memento
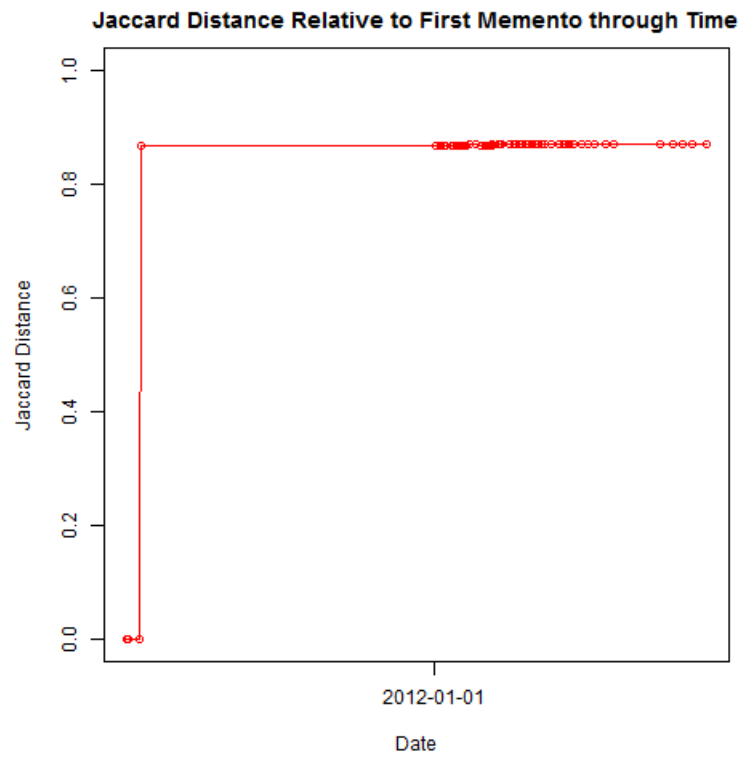
Figure 13: Jaccard Distance Relative to First Memento
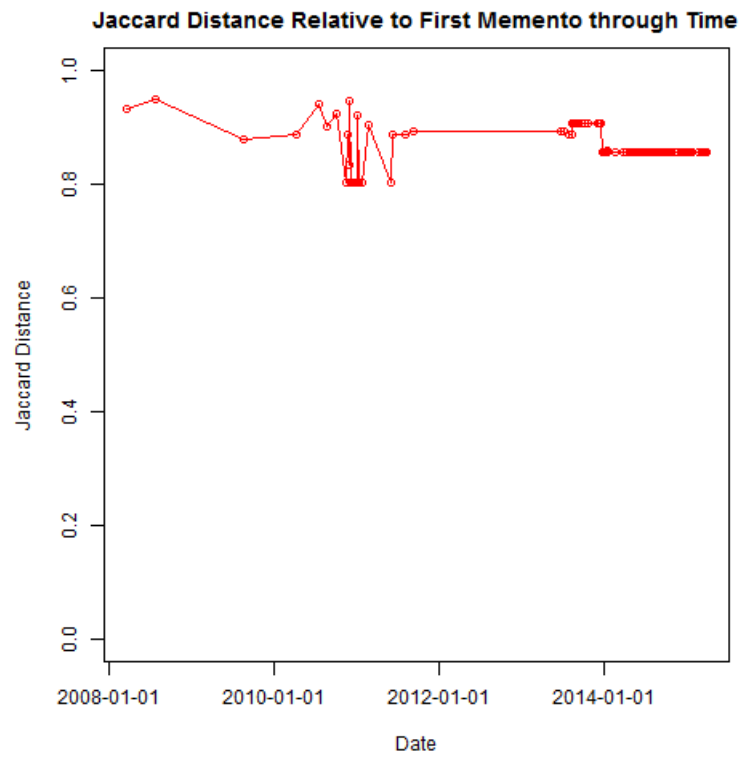
Figure 14: Jaccard Distance Relative to First Memento
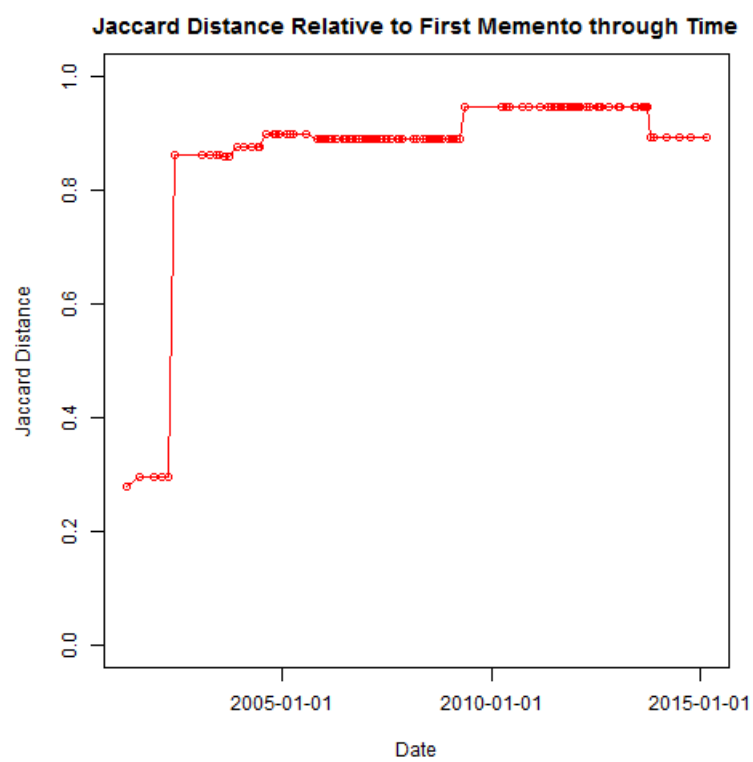
Figure 15: Jaccard Distance Relative to First Memento
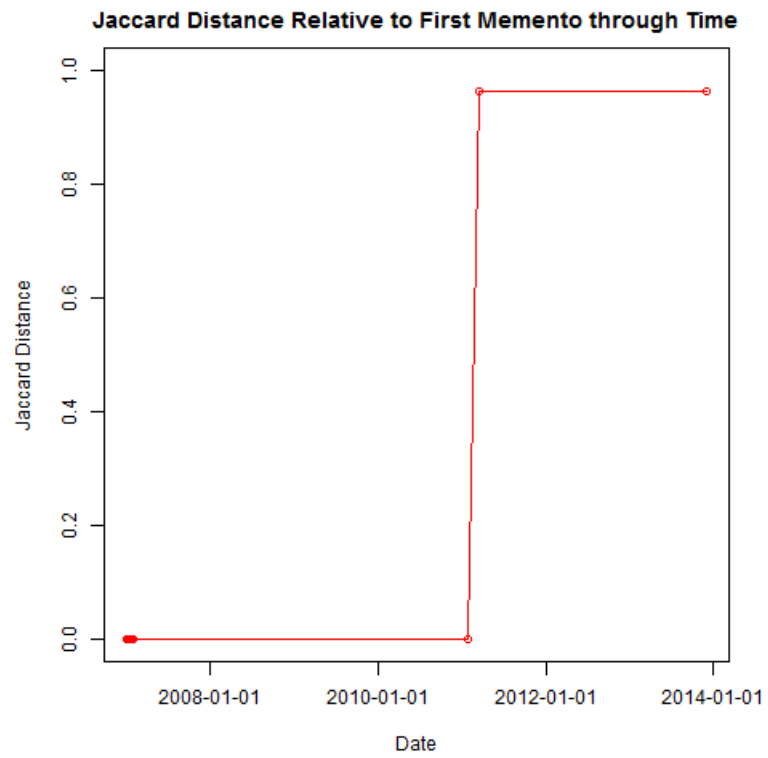
Figure 16: Jaccard Distance Relative to First Memento
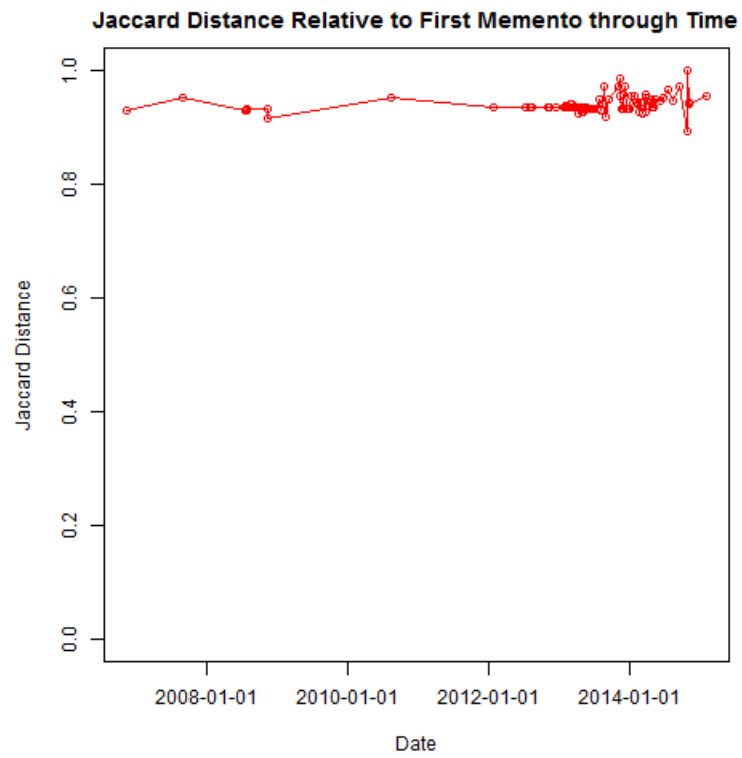
Figure 17: Jaccard Distance Relative to First Memento
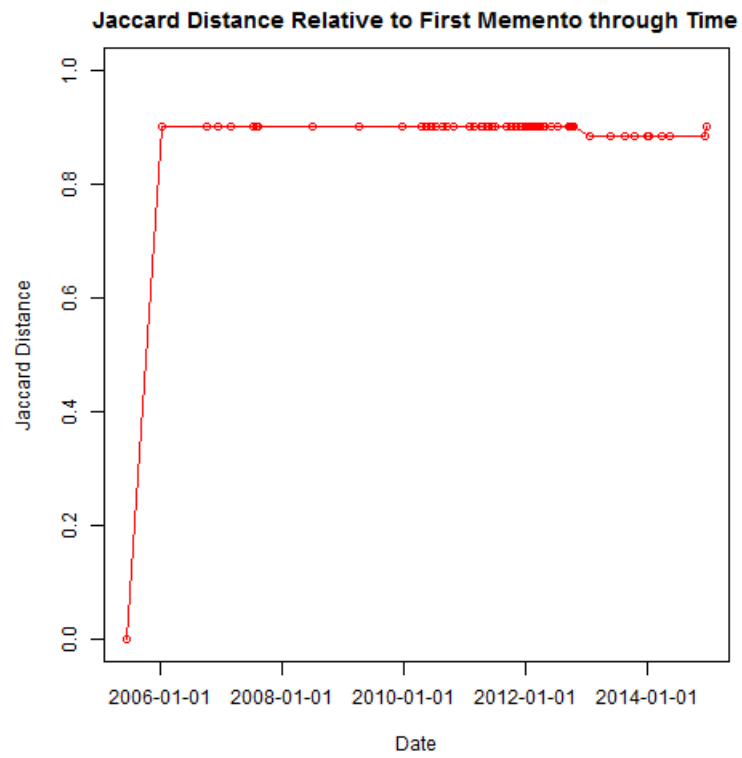
Figure 18: Jaccard Distance Relative to First Memento

Figure 19: WordCloud for Event "Apple Watch" on 25th March

Figure 20: WordCloud for Event "Apple Watch" on 26th March

Figure 21: WordCloud for Event "Apple Watch" on 27th March

Figure 22: WordCloud for Event "Apple Watch" on 28th March

Figure 23: WordCloud for Event "Apple Watch" on 30th March
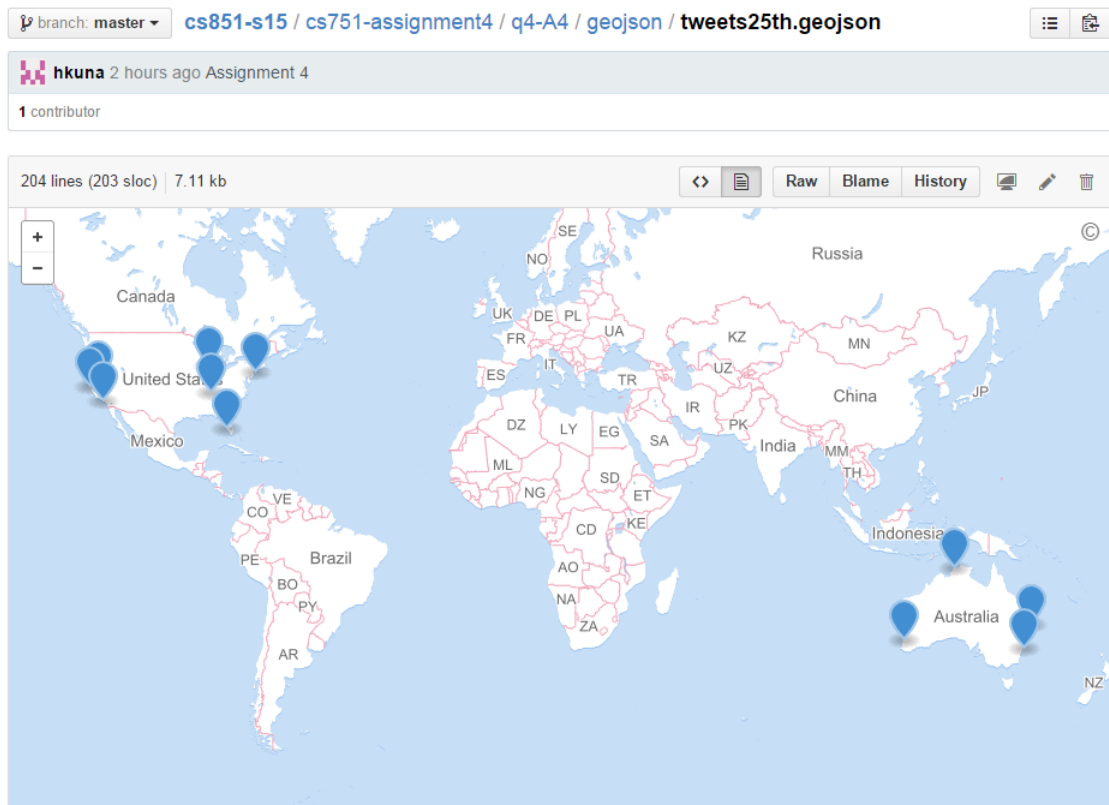
Figure 24: Geo-Coordinates for Event "Apple Watch" on 25th March
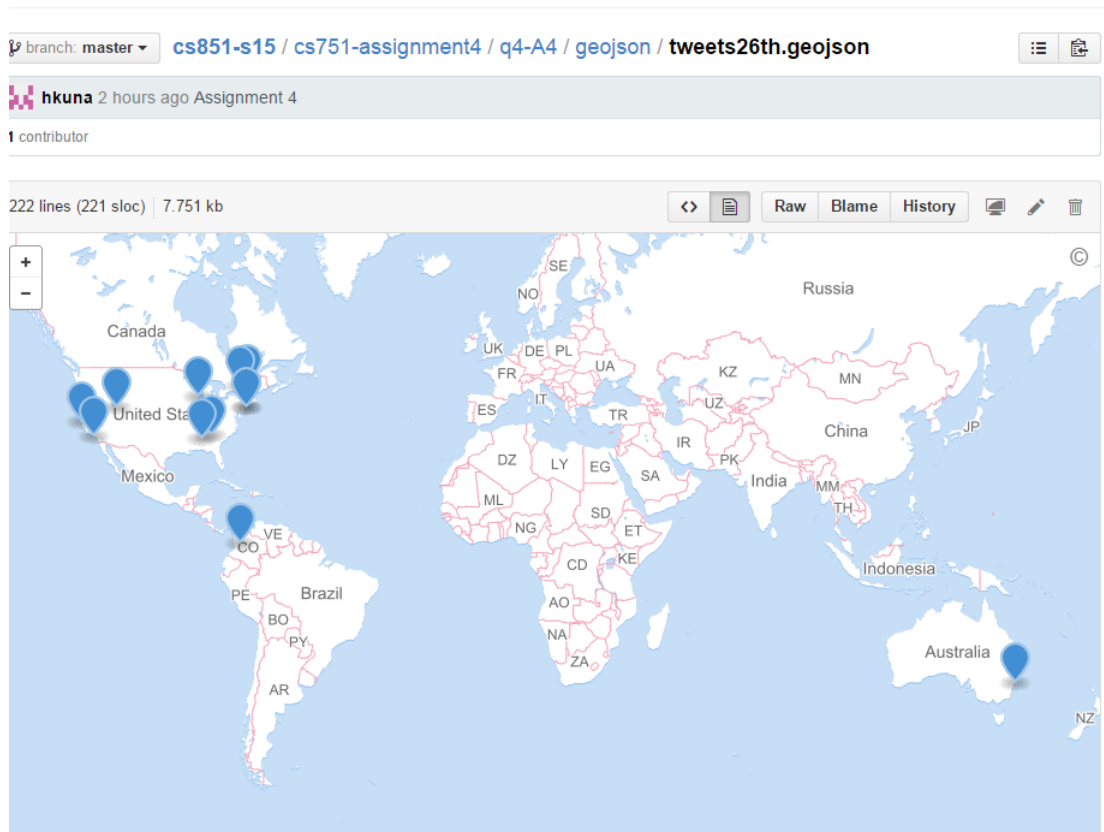
Figure 25: Geo-Coordinates for Event "Apple Watch" on 26th March
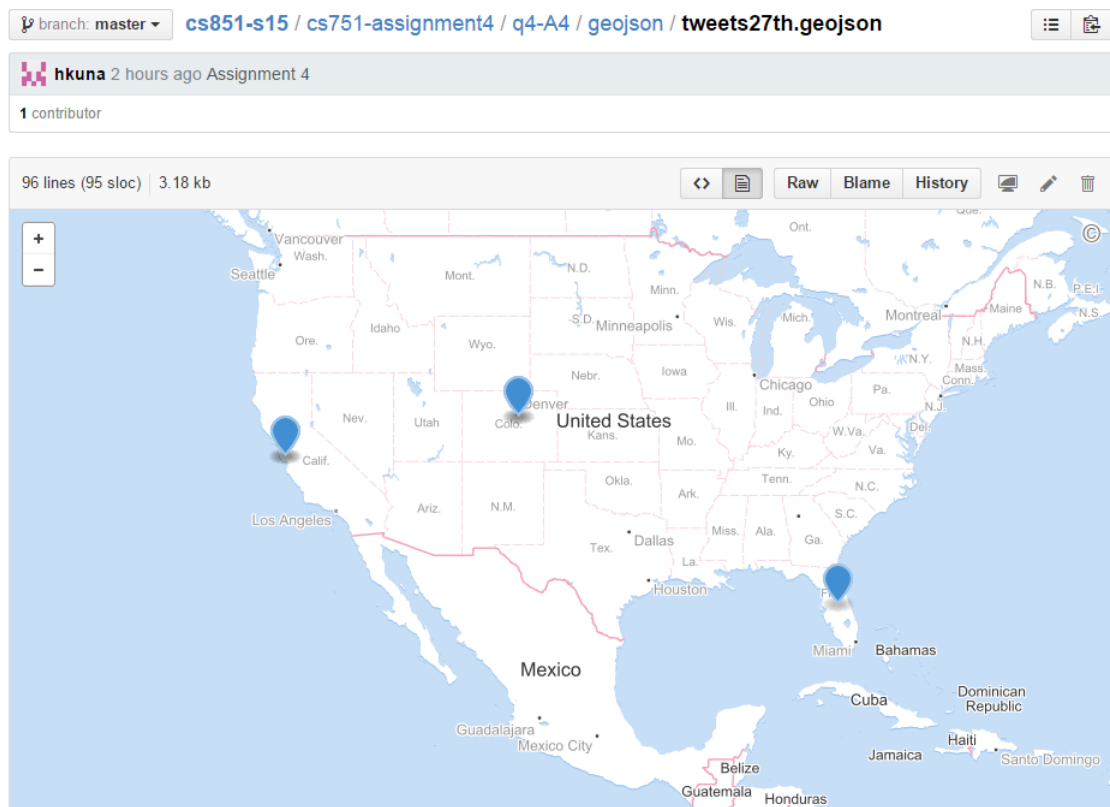
Figure 26: Geo-Coordinates for Event "Apple Watch" on 27th March
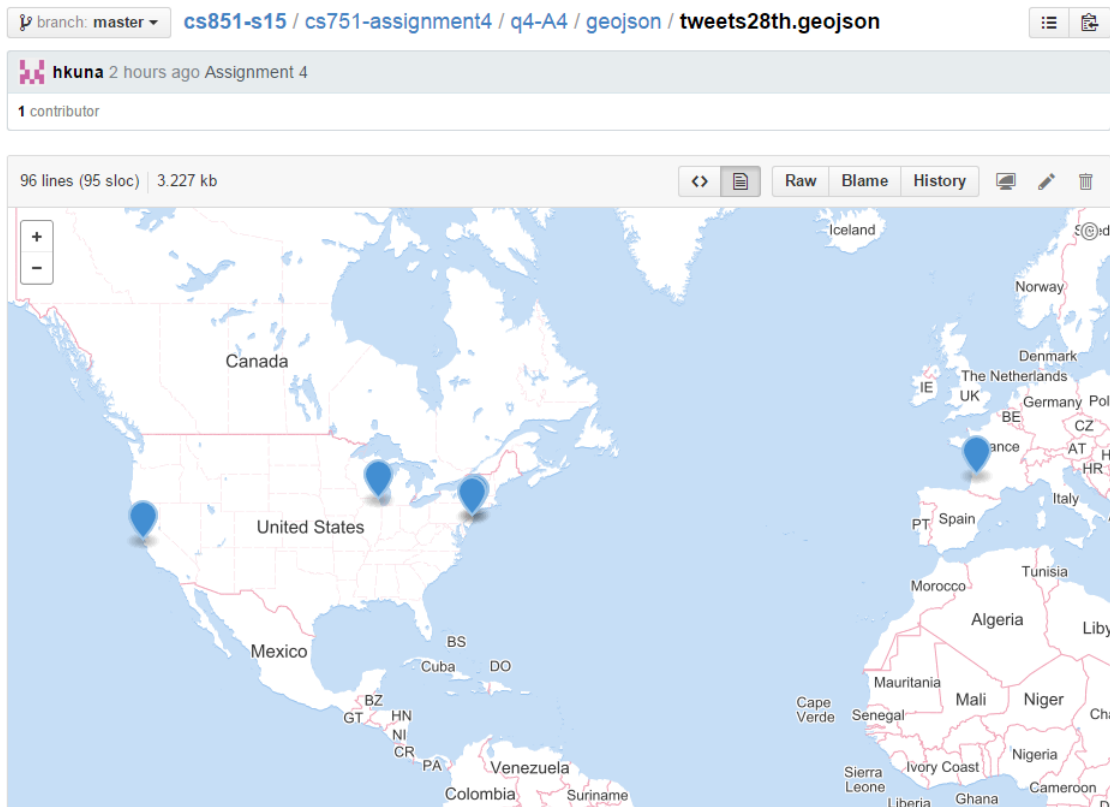
Figure 27: Geo-Coordinates for Event "Apple Watch" on 28th March
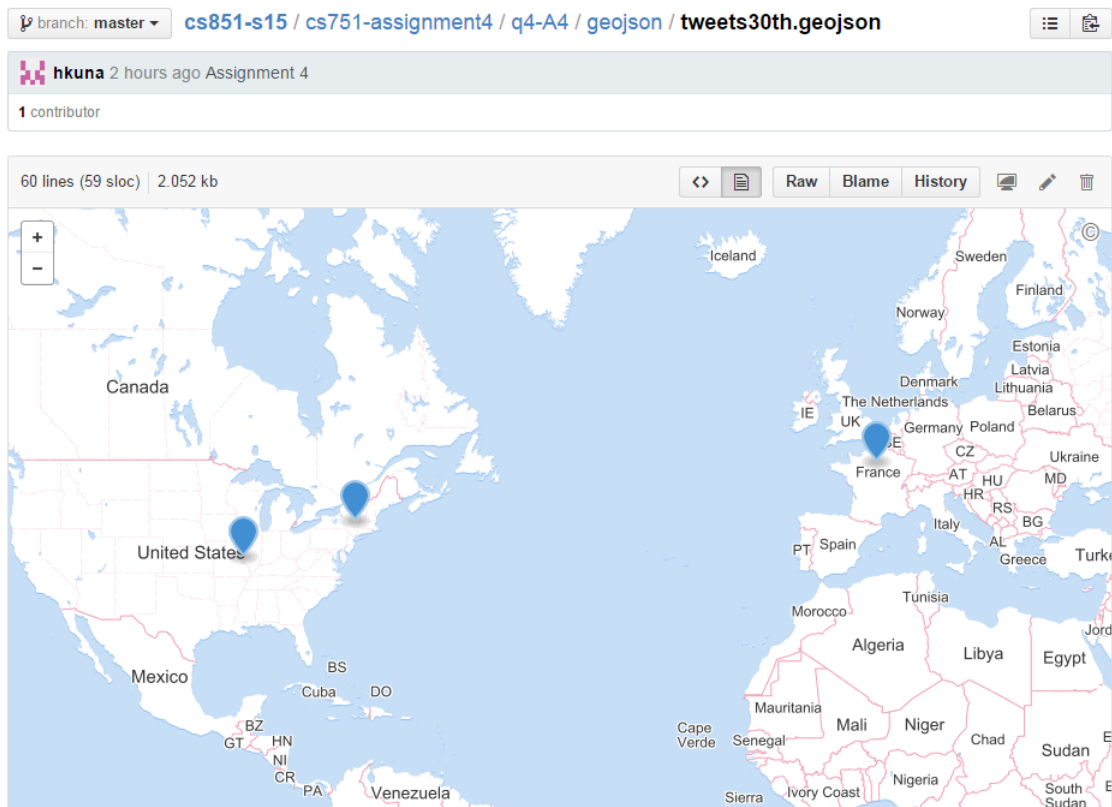
Figure 28: Geo-Coordinates for Event "Apple Watch" on 30th March

file:///C:/cs751/cs751-assignment4/q4-A4/wall/tweets25th.html

Title Here

created on the command line with _twarc_

태현♡なおべいびー
nok1230

昨日、Applewatch見てきたけど やっぱ
いらん□□ http://t.co/K9OhPuOxwL

0 Retweets
Sun Apr 26 02:14:47 +0000 2015

Rene Limberger
renelimberger

Has anyone been able to get
PUICCrownInputSequencer to work?
#WatchKit #AppleWatch #PUICore

0 Retweets
Sun Apr 26 02:14:39 +0000 2015

J.R. Wright
jrwrightchd

RT @AlexFayman: GO INSIDE
INTERNET & EMERGE W/ #CIA
#DATABASE http://t.co/wk2bwyUdeb
#iOS #iOS83 #AppleWatch #ios82
#ipad #BookBoost http:/...
1 Retweets
Sun Apr 26 02:14:36 +0000 2015

Alex Fayman
AlexFayman

GO INSIDE INTE
W/ #CIA #DATAB
http://t.co/wk2bwy
#AppleWatch #ios
#BookBoost http:
1 Retweets
Sun Apr 26 02:14:21

Don Blair
d__blair

iPhone Life is giving away an Apple
Watch! Enter today so we can BOTH
WIN: http://t.co/RWPEjEu1pV

0 Retweets
Sun Apr 26 02:14:09 +0000 2015

kathy whitney
Risingstar2

Win an #AppleWatch from
http://t.co/78HBjZEIQu #contest
http://t.co/llYEQgpfVl

0 Retweets
Sun Apr 26 02:13:45 +0000 2015

Musics Store
Jessie_Courser

#AppleWatch Apple Watch Sport
38mm Silver Aluminium Blue Band
NEWUNOPENED**SEALED IN
BOX http://t.co/cY5JxAKaSQ #Apple
#iWatch
0 Retweets
Sun Apr 26 02:13:35 +0000 2015

たけ
take_kagoMU

@Badboys2Men
すかー 私的にはA
から盛り上がってく
ます。分かりにくい
の良かった体験が
[...]ります。
Sun Apr 26 02:13:31

InspiredBluSolutions
NspiredBlueNews

RT @steube: Early Look At The
Apple Watch App Ecosystem
http://t.co/ECFbasD2Bp #AppleWatch
#wearables #WearableTech
http://t.co/eDEYrc4IoQ
3 Retweets
Sun Apr 26 02:13:20 +0000 2015

Basic_Tech
xXThe_MLGXx

Enter here for a chance to Win a
Brand New Apple Watch! Follow
@PapiGfunk http://t.co/kA6xW7KYh9
#AppleWatch #AppleWatchSport
#Apple
0 Retweets
Sun Apr 26 02:13:07 +0000 2015

Game Dev Bot
GameDevBot97

RT @editorseditor:
#GameDevelopers on #AppleWatch:
smartwatches are all about context
http://t.co/dCQlLgK4Yv

1 Retweets
Sun Apr 26 02:13:02 +0000 2015

opal8877
opal8877

ニコニコ超会議20
AppleWatchをつ
でも評判悪い！ ht
http://t.co/XWCk5

0 Retweets
Sun Apr 26 02:12:50

ばっきゃ
singlekill_17

RT @bakusy0u_gazou: AppleWatch
の闇 http://t.co/YKyoxsvFl8

8132 Retweets
Sun Apr 26 02:12:33 +0000 2015

Lucas Ibarra
lucas_ibarra1

#AppleWatch VS. Mini Cooper -
¡FORCETOUCH TEST!
https://t.co/Mi4FAJtPf0

0 Retweets
Sun Apr 26 02:12:13 +0000 2015

Taeho Kim
NuclearNoJam

[지금 애플워치를 사면 안되는 이유]
현재 카메라, 모션센서 등 기능을 서드
파티에서 접근할 수 없어 '애플워치'에
특화된 앱이 없는 상황. 6월은 되야 풀
린다고(cnet참조.
0 Retweets[...]/7qS2rdHou) #애플워치
Sun Apr 26 02:12:12 +0000 2015
#AppleWatch

yhy
yyhhyy

RT @Pirate_Radi
めてさっき見た。さ
イエットの話をしつ
並んでる女の子が
見た。1999年に地
[...]聞いてた僕
Sun Apr 26 02:12:[...]
つもちょっと怖かっ

Figure 29: Wall for Event "Apple Watch" on 25th March

34

Figure 30: Wall for Event "Apple Watch" on 26th March

Figure 31: Wall for Event "Apple Watch" on 27th March

Figure 32: Wall for Event "Apple Watch" on 28th March

Figure 33: Wall for Event "Apple Watch" on 30th March