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 successfully 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. 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-afterboiler platetextfiles" located inside "q1-A4" folder

I have written a java code named "UniGram.java", "BiGram.java", "Tri-Gram.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 "unijackard-sorted.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 , UniGrams, 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 "0.95833333333" 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 $0.01736111111111,\ 0.0138248847926,\ 0.0134831460674$ 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 $\rm Q1/\rm Q2$ 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".

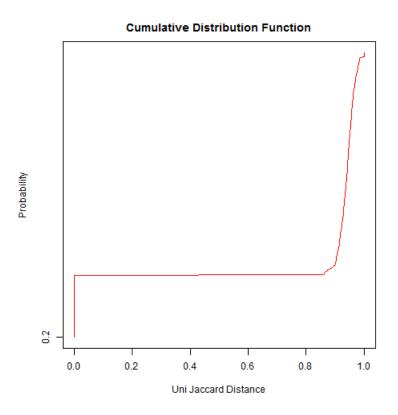


Figure 1: CDF for Unigram Jaccard Distance

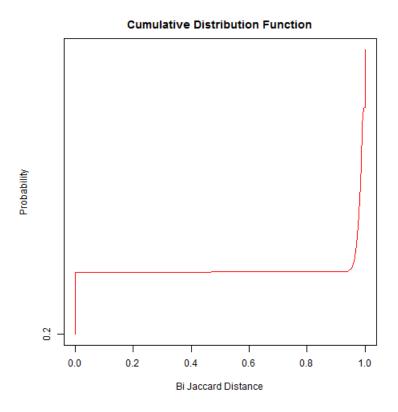


Figure 2: CDF for Bigram Jaccard Distance

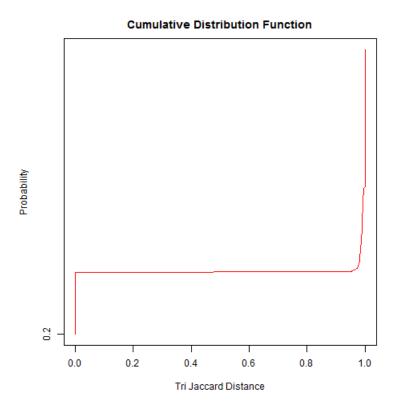


Figure 3: CDF for Trigram Jaccard Distance

CDF for Number of Mementos per URI

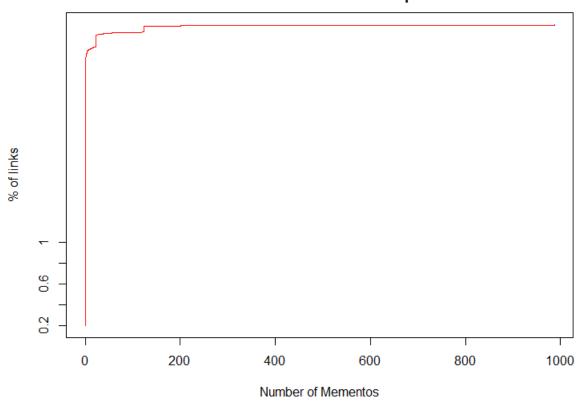


Figure 4: CDF for number of mementos for each original URI

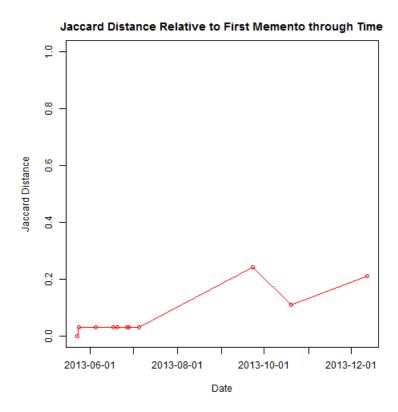


Figure 5: Jaccard Distance Relative to First Memento

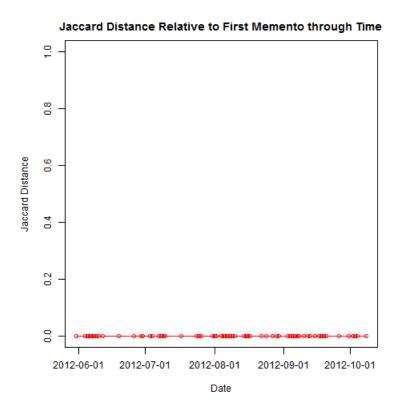


Figure 6: Jaccard Distance Relative to First Memento

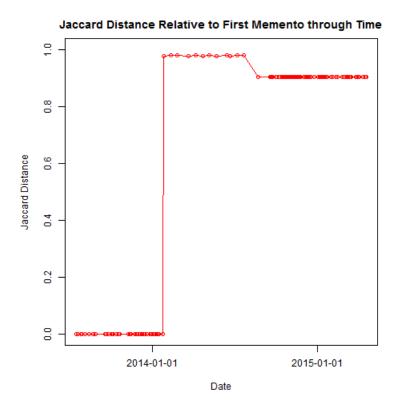


Figure 7: Jaccard Distance Relative to First Memento

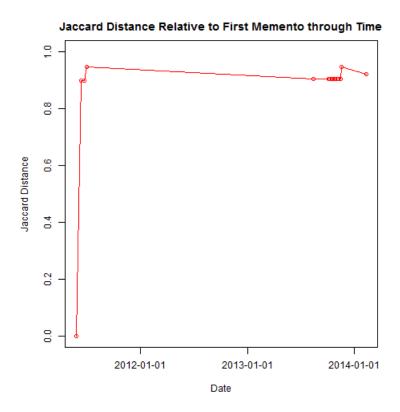


Figure 8: Jaccard Distance Relative to First Memento

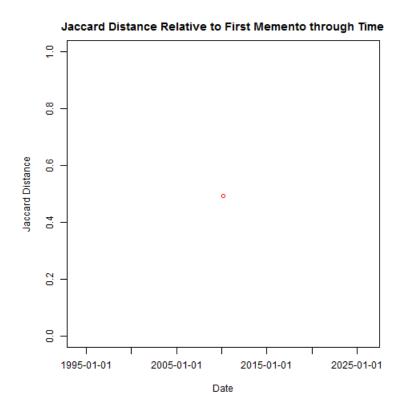


Figure 9: Jaccard Distance Relative to First Memento

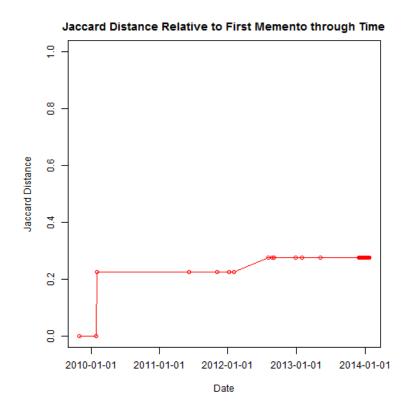


Figure 10: Jaccard Distance Relative to First Memento

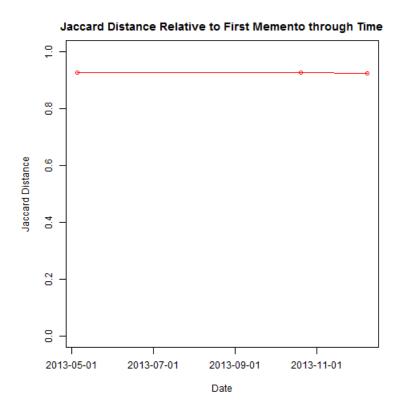


Figure 11: Jaccard Distance Relative to First Memento

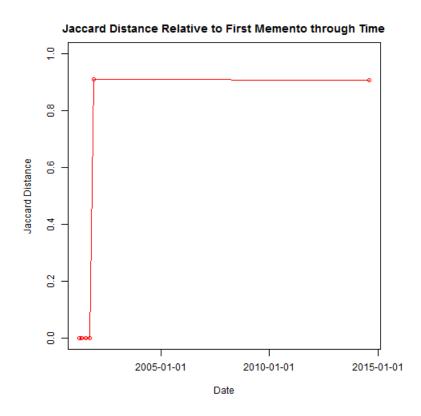


Figure 12: Jaccard Distance Relative to First Memento

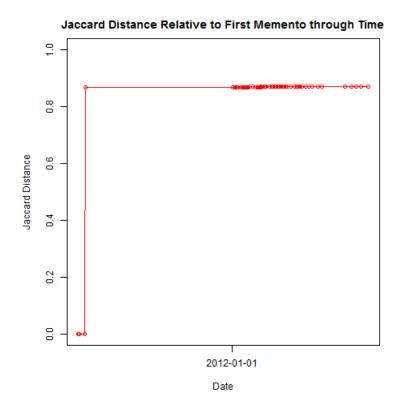


Figure 13: Jaccard Distance Relative to First Memento

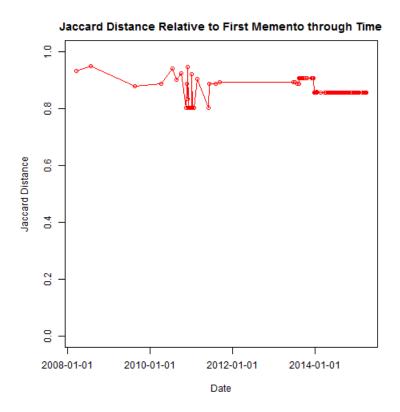


Figure 14: Jaccard Distance Relative to First Memento

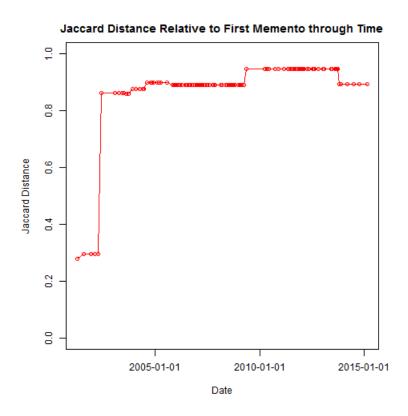


Figure 15: Jaccard Distance Relative to First Memento

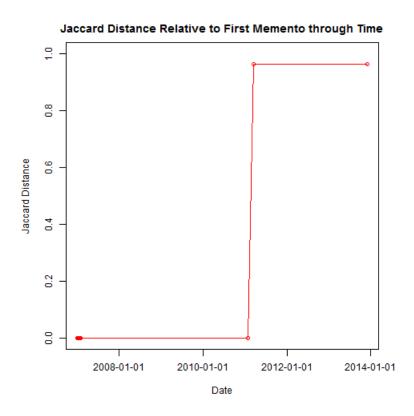


Figure 16: Jaccard Distance Relative to First Memento

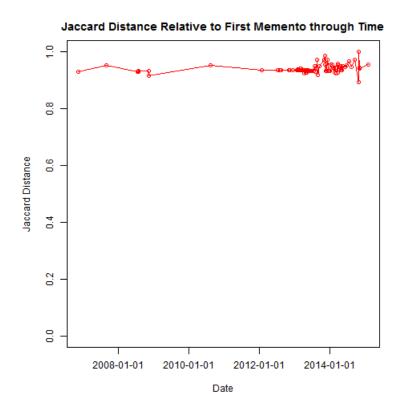


Figure 17: Jaccard Distance Relative to First Memento

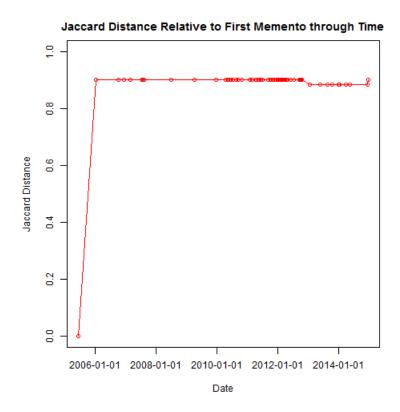


Figure 18: Jaccard Distance Relative to First Memento

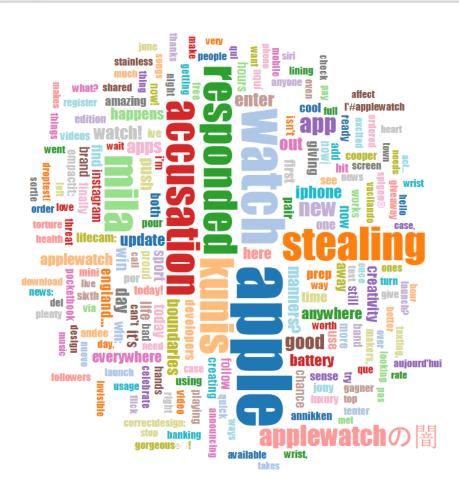


Figure 19: WordCloud for Event "Apple Watch" on 25th March

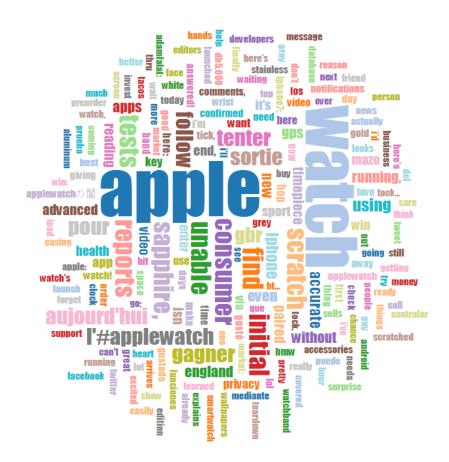


Figure 20: WordCloud for Event "Apple Watch" on 26th March



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

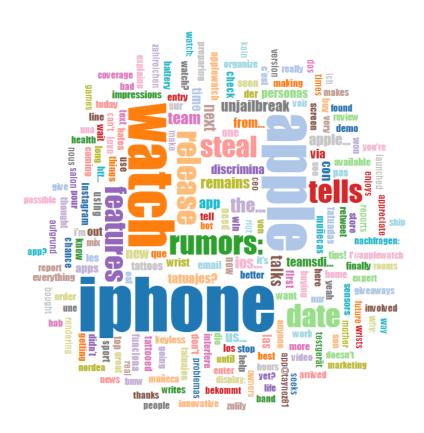


Figure 22: WordCloud for Event "Apple Watch" on 28th March



Figure 23: WordCloud for Event "Apple Watch" on 30th March

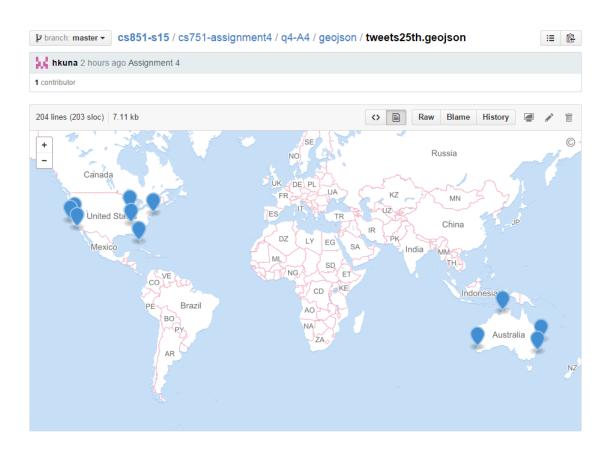


Figure 24: Geo-Coordinates for Event "Apple Watch" on 25th March

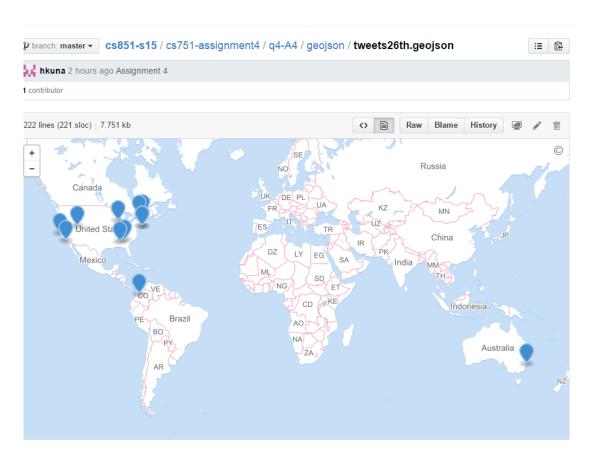


Figure 25: Geo-Coordinates for Event "Apple Watch" on 26th March

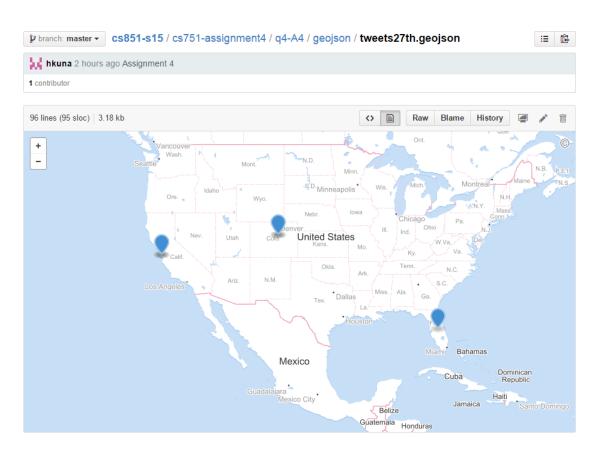


Figure 26: Geo-Coordinates for Event "Apple Watch" on 27th March

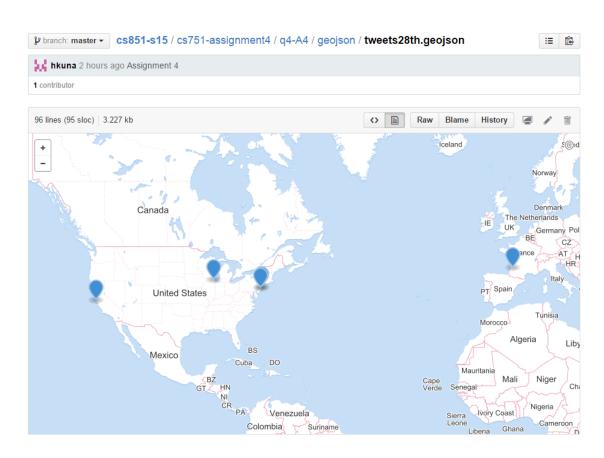


Figure 27: Geo-Coordinates for Event "Apple Watch" on 28th March

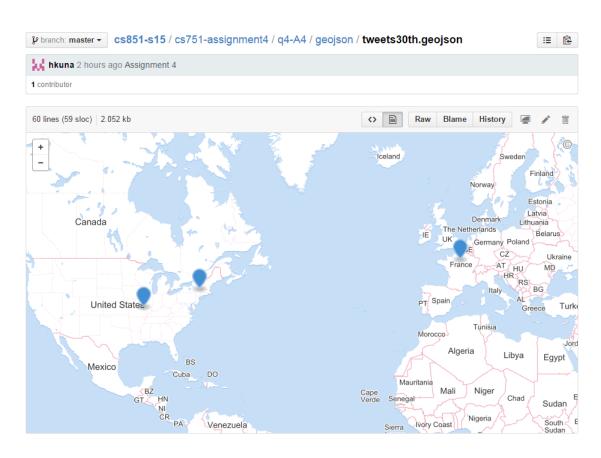


Figure 28: Geo-Coordinates for Event "Apple Watch" on 30th March

戻 태현でなおべいびー

の闇 http://t.co/YKyoxsvFI8

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

Title Here

created on the command line with twarc

특화된 앱이 없는 상황. 6월은 되야 풀

린다고(cnet참조. NRetweets/17qs2/zqHow)2#앵플워치 #AppleWatch

Alex Fayma

GO INSIDE INTE

W/ #CIA #DATAE

http://t.co/wk2bwy

#AppleWatch #io

#BookBoost http:

Sun Apr 26 02:14:21

take_kagoMl

@Badboys2Men

ずかー 私的にはA

から盛り上がってく

ます。分かりにくし

の良かった体験が

ERetweets) ± ± Sun Apr 26 02:13:31

🎅 opal8877

opal8877

ニコニコ超会議20

AppleWatchをつけ

でも評判悪し! ht

http://t.co/XWCk5

Sun Apr 26 02:12:50

yyhhyy

RT @Pirate_Rad 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年に対 **発力的関係**でた僕 つもちょうと怖かっ

0 Retweets

yhy yhy

1 Retweets

🎅 たけ

AlexFayman

房 J.R. Wright



Rene Limberger

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

iFORCETOUCH TEST!

https://t.co/Mi4FAJtPf0

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

Title Here

created on the command line with twarc

Eder Bustos アップルウォッチ最新情報 RedeMarquez applewatch_news Me gustó un video de @YouTube de ITmedia Mobile: 予約制でも行列: @marcianophone http://t.co/QAttbTchpS Mazo vs Apple Watch (別のでは、 ーストインブレッション http://t.co/PRMgny1bJv 0 Retweets Mon Apr 27 02:57:18 +0000 2015 #AppleWatch - THE END, #HammerTest 0 Retweets Mon Apr 27 02:57:21 +0000 2015 Thomas Raia Andika Yuda siholand Yuda_siholahans mocharaia

RT @AppStore: Grab your bags and go; the world is waiting. #AppleWatch http://t.co/X2hKxwC46Z when using it without its paired http://t.co/1KRaXsdj6O iPhone for GPS. http://t.co/Wjy2uP...

151 Retweets Mon Apr 27 02:57:12 +0000 2015 Mon Apr 27 02:57:07 +0000 2015

📝 よびよび バハ鯖 yopiyopi120 JohnNoonan

RT @kabubunet: こんなアプリがあれ ば急なトイレも安心! #FF14 #AppleWatch http://t.co/OCIJnf8zpn

140 Retweets Mon Apr 27 02:56:55 +0000 2015

Red Lips markefosegah

iPhone7: 日本初! AppleWatch開発者イベント開催 ウォッチアブリ設計の苦労 と工夫と期待を訊く http://t.co/JMFLGr9osL#iPhone

0 Retweets Mon Apr 27 02:56:19 +0000 2015

「Apple Watch at Isetan Shinjuku」で Apple Watch販売開始 開封からファ

RT @hverlind: This is how accurate the #AppleWatch is for running, even

1359 Retweets

John Noonan

RT @hverlind: This is how accurate the #AppleWatch is for running, even when using it without its paired iPhone for GPS. http://t.co/Wjy2uP.

1359 Retweets Mon Apr 27 02:56:37 +0000 2015

Black Heart markasaboput

iPhone7: 日本初! AppleWatch開発者イベント開催 ウォッチアブリ設計の苦労 と工夫と期待を訊く http://t.co/nWvqauB8Fs #iPhone

0 Retweets Mon Apr 27 02:56:18 +0000 2015

🎅 ds20u2 ds20u2

RT @priceminister: #jeu Sortie de I'#AppleWatch aujourd'hui ! #Happy ? Follow @PriceMinister + RT pour tenter de la gagner http://t.co/dO3y.

2480 Retweets Mon Apr 27 02:57:15 +0000 2015

BPI Solutions **BPIsolutions**

RT @iotattack: Can #AppleWatch really help you get fit? http://t.co/hAeZOx639R #loT #WearableTech http://t.co/2k36je3qlr via joanaw

2 Retweets Mon Apr 27 02:57:05 +0000 2015

Alwaleed Khaled Alwaleed_Khaled

فيديو: تحمل ساعة أبل :RT @AlziadiQ8Blog ـ للماء والسقوط والدهس AppleWatch# http://t.co/MiyioCZy5b

1 Retweets Mon Apr 27 02:56:27 +0000 2015

戻 lex pon Dy lexpondy

iPhone7: 日本初! AppleWatch開発者イベント開催 ウォッチアブリ設計の苦労 と工夫と期待を訊く http://t.co/FyFvGYGrwW #iPhone

0 Retweets Mon Apr 27 02:56:16 +0000 2015

AppleWatch barna_clue

RT @antoniobart #AppleWatch... ¿ #WaterProof http

1 Retweets Mon Apr 27 02:57:14

📝 アップルウォ: applewatch_i

ボルシェがいちは 連携する http://t.c @clicccarさんから

0 Retweets Mon Apr 27 02:57:01

Negr NestorNegro

Ya tengo el #App falta el iPhone y l próximos 6 mese

0 Retweets Mon Apr 27 02:56:21

Bella Donna

sjomalee

iPhone7: 日本初! イベント開催 ウォン と工夫と期待を訊く http://t.co/IF3KGF 0 Retweets Mon Apr 27 02:56:14

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

Title Here

created on the command line with twarc

AppleWatchOwner AppleWatchOwner RT @W2511Danni: Op-Ed: Is the Apple Watch a necessity or luxurious accessory? #AppleWatch #Apple http://t.co/sC2HNcqNIR via @digitaljournal 1 Retweets Tue Apr 28 03:42:47 +0000 2015 Musics Store

Jessie Courser #AppleWatch Apple Watch Sport 38mm White *Brand New* Unopened http://t.co/fEY0KDgRJi #Apple

0 Retweets Tue Apr 28 03:42:09 +0000 2015

戻 ramiro gallegos RamiroAlcaza

#iWatch

(EXCLUSIVA) Así sera la App Store del #AppleWatch (Preview de las Apps) https://t.co/4Si6Hrv1ML vía @YouTube

0 Retweets Tue Apr 28 03:41:21 +0000 2015

Mark Sgarlato sgarlato

RT @AppleWatchGuru: #AppleWatch Users Report Unintended Purchases http://t.co/f7ST9JnHLc #UX #appdev #Wearables http://t.co/2qHvbexW1C

7 Retweets Tue Apr 28 03:40:38 +0000 2015 AppleWatch News barna_clue

RT @derek_arg: Big fan of the show Drunk Tech Review :) @DrunkTechReview http://t.co/dWcfMplfec check it out! #AppleWatch #Tech #YouTube #. 1 Retweets Tue Apr 28 03:42:21 +0000 2015

戻 takumix takumix30

AppleWatch、充電の持ちは? 実際に1 日使ってみた感想や充電に関する豆知 識まとめ。 - ソフトバンクiPhone最新情 報 http://t.co/k8p0Equ3vv

0 Retweets Tue Apr 28 03:42:05 +0000 2015

CKob24 CKob24

It finally arrived! #applewatch #38mm #spacegray https://t.co/NBsbf9cfpq

Tue Apr 28 03:41:17 +0000 2015

📝 樋口和輝 kazy_14

RT @Toyokeizai: 【AppleWatch、使って3日目に分かったこと】 要するに 「iPhoneとの付き合い方」が変わる: http://t.co/oG34eTXTC3 http://t.co/xNoaODoxnd 21 Retweets Tue Apr 28 03:40:07 +0000 2015

Derek Argueta derek_arg

Big fan of the show Drunk Tech Review:) @DrunkTechReview http://t.co/dWcfMplfec check it out! #AppleWatch #Tech #YouTube #Show #DTR

1 Retweets Tue Apr 28 03:42:20 +0000 2015

房 LiZa snlz0

http://t.co/CHOhWQ6W08 Apple Watch Problems: Connecting to Wi-Fi, Long Sync Times, Notifications 10 31 http://t.co/fL6R0B8zDw

0 Retweets Tue Apr 28 03:42:03 +0000 2015

🍃 かなまま Ryieco_open

AppleWatch、表参道のAppleStoreで 試着。悩んだけどしっくりきたのは 38mm。Sportsは夏場にバンドが汗で ベトベトしそうだし汚れそうなのでやめ #E! http://t.co/w0Lw50yt1S 0 Retweets Tue Apr 28 03:41:11 +0000 2015

戻 The Tin Woodman NickChopperDigg

RT @moneymakingmode 5 Smartwatches That are Smarter than #AppleWatch: http://t.co/ffRE96deKp via @TheTufiTech

0 Retweets Tue Apr 28 03:40:04 +0000 2015 The Carous carouselwor

Think you know about the #Apple - here's the dow low...http://t.co/1 http://t.co/fY0K9

0 Retweets Tue Apr 28 03:42:1

Workshop Bucket Doll

RT @psswstyles #Apple Apple W Space Gray Alu Sport Band (MJ http://t.co/ouCB 1 Retweets Tue Apr 28 03:41:56

房 iPhone 最新 apple_ip_ne

Apple Watch iF (2) 2015/04/24 ついに発売になり で利用を考えた明 そなわっているの O Retweets http:// Tue Apr 28 03:41:0

TheAppTim

TheAppTime

RT @moneymal Smartwatches T #AppleWatch: https://doi.org/10.1003/10.1003 via @TheTufiTe

0 Retweets Tue Apr 28 03:40:03

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

Title Here

created on the command line with twarc

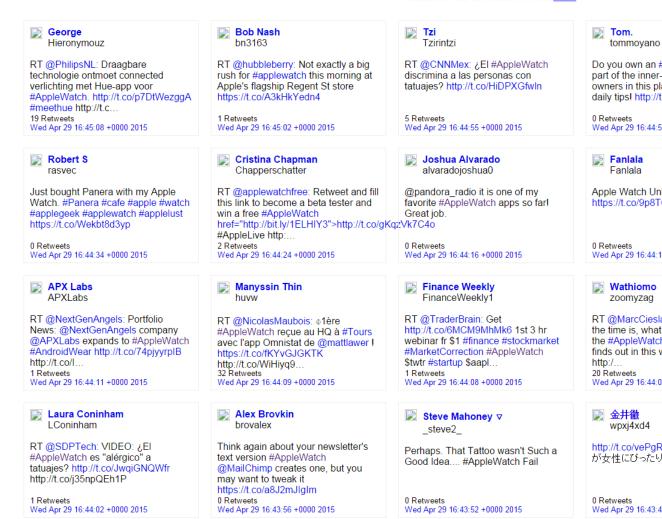


Figure 32: Wall for Event "Apple Watch" on 28th March

OrlandoBizJournal

Have tattoos? This could be bad

news for Mac-heads who want an

#AppleWatch http://t.co/iqwIPTvpI0

OBJUpdate

sunlight188

http://t.co...

24 Retweets

with one article a day on

Thu Apr 30 13:01:16 +0000 2015

RT @AppStore: .@YahooNewsDigest

#AppleWatch. http://t.co/ip0Yn0cAwf

continues to amaze us - this time

RIKKY

pad__ricky

RT @sonicdrum: 【悲報】

が発覚!供給を制限へ

『AppleWatch』の主要部品に不具合

Title Here

Nasri Zuhair

Problems with #WearableTech

http://t.co/ss2TCtSu0w #AppleWatch

NasriZuhair

harimiya

0 Retweets

今日の含み益分でapplewatch買おうと

思ったけど不具合とやらどうなんかなぁ

Thu Apr 30 13:01:11 +0000 2015

created on the command line with twarc

Nikki's Mart

RT @KidCudi: So sending some an using the Apple V

RT @3ajmee: أبك

#AppleWatch http

Thu Apr 30 13:01:09

35 Retweets

Andy_FromN

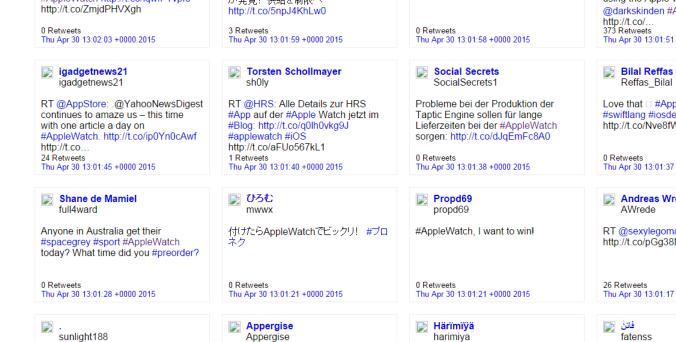


Figure 33: Wall for Event "Apple Watch" on 30th March

Appergise

http://t.co..

24 Retweets

Appergise

with one article a day on

Thu Apr 30 13:01:12 +0000 2015

RT @AppStore: .@YahooNewsDigest

continues to amaze us - this time

#AppleWatch. http://t.co/ip0Yn0cAwf