Event Detection & Retrieval from Twitter

Srinjay | Datta | Sanjeev | Robin | Sujan | Bodhi | Niten
Mentored by – Sankarshan Mridha

Motivation

Twitter emerged from social network to a news media for tracking real-world events, like-


NWS New York NY @NWSNewYorkNY · 13h

April 7th marks the earliest 90° day ever recorded in Central Park. It was back in 2010 when the temperature reached 92°. Today's high 58°.

17 21

What? Earliest 90* day
When? April 7
Where? Central Park
Today’s high 58*

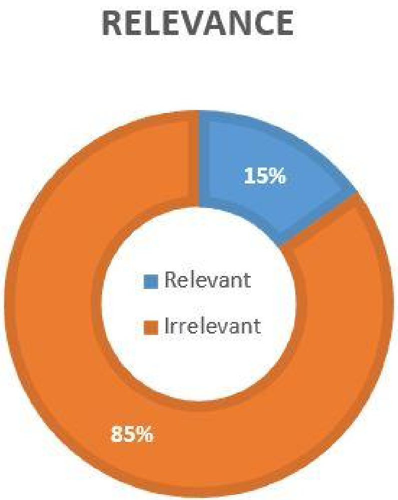
Motivation

Some relevant twitter handles for extracting meaningful events-

Politics	NYC Politics (@PoliticsInNYC)
Sports	BBC Sport (@BBCSport) Star Sports (@StarSportsIndia)
Natural Disasters	Get Ready Get Thru (@NZGetThru)
Events Organized	New York Nightlife (@NYNightlife) Fashion and Style (@FashionAndStyle)
Taxi Availability & Planning	YellowCabNYC (@YellowCabNYC) taxiNYC (@taxiNYC)
Public Departments of US (Govt.)	NYC DOT (@NYC_DOT) NYC Finance (@NYCFinance)

Chosen

Motivation



254.5
tweets/month

[NYC DOT (@NYC_DOT)]

Twitter Handle:
- Information related to Transportation at NYC

Primary objective:
- Extracting Eventful Information.

Event:
- NYC Road-Block/Closure

Information:
- Event Tag
- Date
- Time
- Location

@CentralParkNYC #65thStTransverse
10th April, 2016 12:01-6AM

Data

TWITTER → API → JSON → Parser → CSV

Tweets

From NYC DOT Twitter Handle

Around 2000 tweets, with –

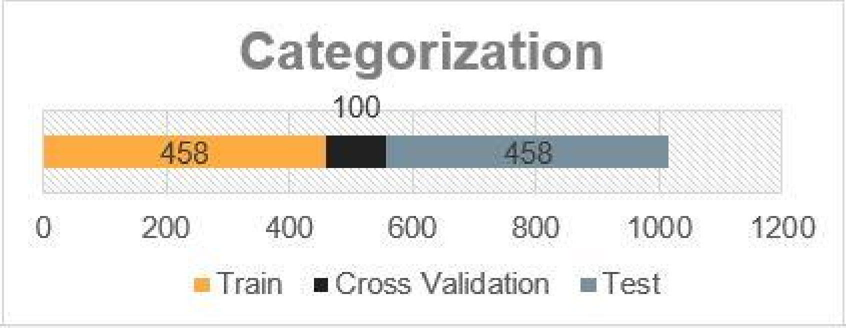
- Time of tweet
- Tweet Description
- Hashtags

Manual labeling of all tweets are done (Relevant / Non-Relevant)

Data

TWITTER → API → JSON → Parser → CSV

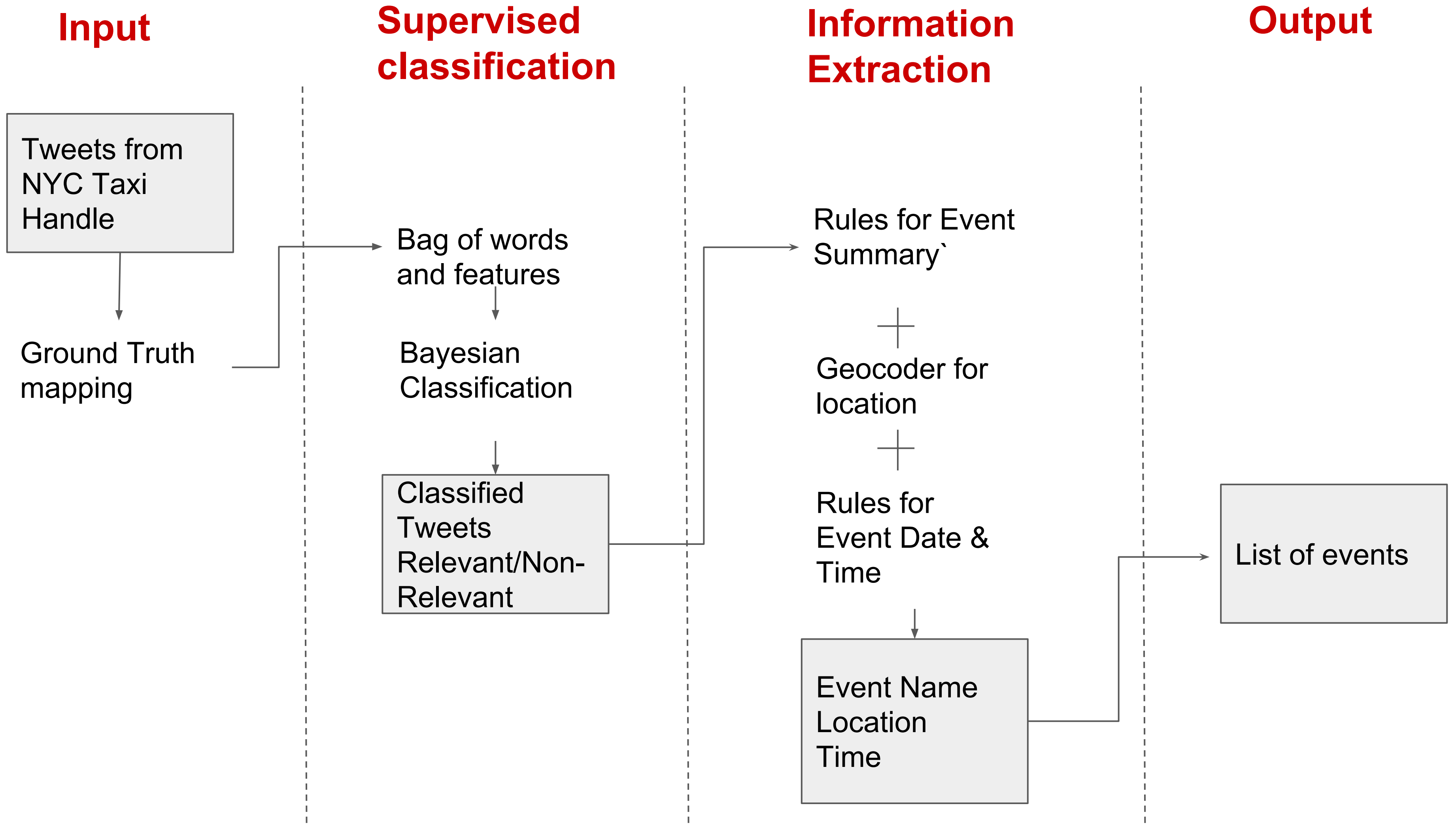
Tweets



Problems faced –

- Mostly Irrelevant Data (wrt Roads)
- Out of Vocabulary Words used
- High amount of noise and ambiguity of semantic

Process Flow Diagram


Input
Supervised classification
Information Extraction
Output
Tweets from NYC Taxi Handle
Ground Truth mapping
Bag of words and features
Bayesian Classification
Classified Tweets Relevant/Non-Relevant
Rules for Event Summary`
Geocoder for location
Rules for Event Date & Time
Event Name Location Time
List of events

Ground Truth Marking

Preliminary Marking
- Relevant
- Non-Relevant

→ Final Marking

- Preliminary marking done by all the team members

- The marking process gave insights into rules for extraction of date and time and location

- Inter Annotator Agreement is measured among

Tweet Pre-Processing

To prepare data for classification, the following is done consecutively -

1. **Normalization**
   a. Punctuation Removal
   b. Squeezing of Whitespace
2. **Encoding**
   a. Twitter API introduces some characters outside UTF-8
   b. the others are generated from the csv
3. **Date and Time Attribute** Introduced
   a. A new feature added, instead of treating all dates individually
   b. Same applied for time.
4. **Numeric Removal**, since it unnecessarily increases the number of tokens
5. **Tokenization** (Tweets broken into words)
6. **Stop Words Removal**
7. **Stemming** (As bag of words are considered as feature, stemming is important)

Feature Engineering

Step 1 :

$$Word\ Score = \left(\frac{Occurence\ in\ Relevant\ Tweets}{Total\ \#\ of\ Relevant\ Tweets}\right) - \left(\frac{Occurence\ in\ NonRelevant\ Tweets}{Total\ \#\ of\ NonRelevant\ Tweets}\right)$$

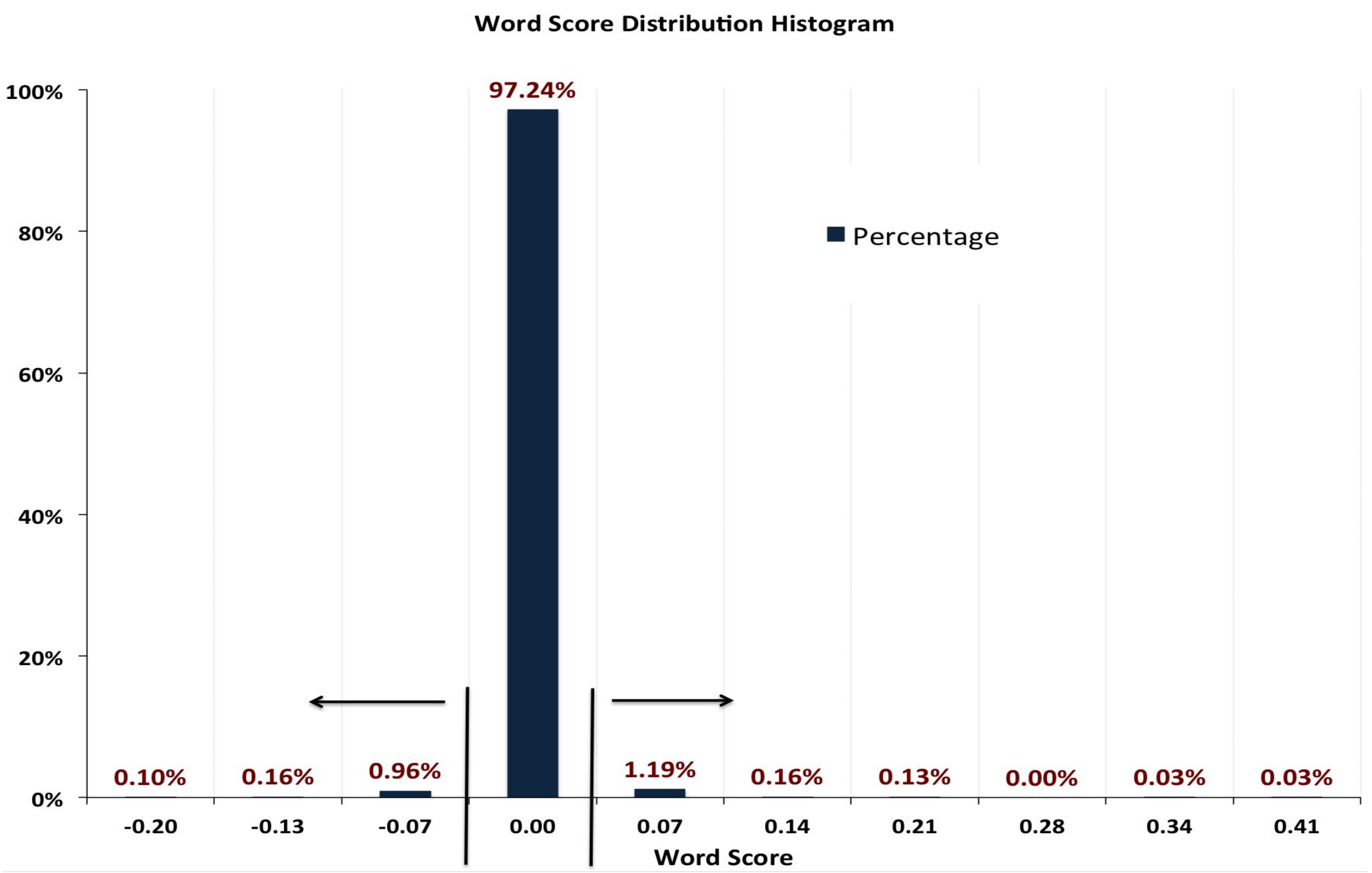
- Score indicative of Relevance prediction ; Appropriate Normalization

Step 2 :

Distribution of the word score for all the words plotted. [in the next slide]

Step 3 :

Extracting words with outlier scores [say, outside the 95% confidence interval]



Word Score Distribution Histogram
100%
80%
60%
40%
20%
0%
97.24%
Percentage
0.10%
0.16%
0.96%
1.19%
0.16%
0.13%
0.00%
0.03%
0.03%
-0.20
-0.13
-0.07
0.00
0.07
0.14
0.21
0.28
0.34
0.41
Word Score

Feature Engineering

Positive Words:

['bridge', 'sun', 'fri', 'bound', 'thru'
'one', 'sat', 'closed', 'closure', 'closures',
'lanes', 'reminder', 'full', 'overnight', 'lane', 'from', 'will']

Negative Words
['here', 'office', 'free',
'report', 'contact', 'the',
'pls', 'commissioner', 'your', 'with', 'you',
'share', 'please', 'location', 'dot', 'our']

Bayesian Text Classification Incomplete Slide

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

1. It is a conditional probability model: given a problem instance to be classified, represented by a vector x=(x1,...,xn) representing some n features (independent variables).

$$P(c|x) = P(x|c).P(c)/P(x)$$

Event Tags

Objective:

Understand the type of event.

Example:

1. Full CLosure
2. Both Direction

Event Tags

Input:

full closure, both directions 9/2 & 9/3, 11pm-5m. Details below:

Output:

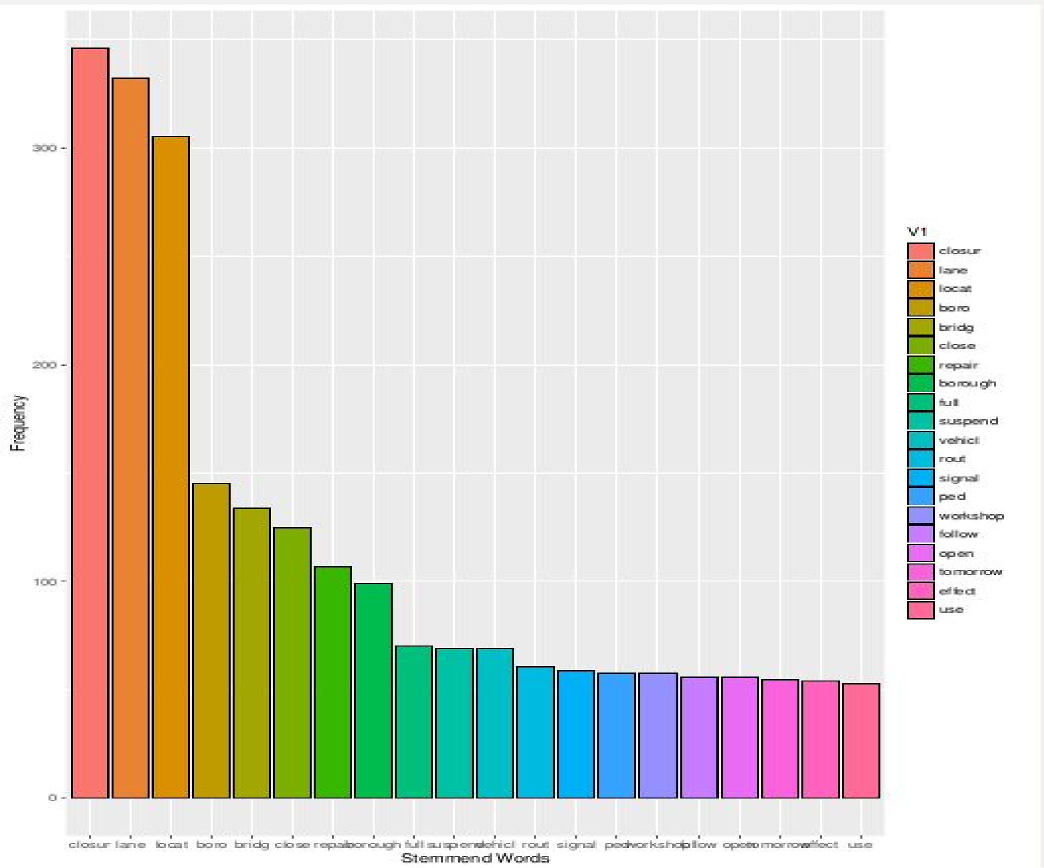
Full closure both directions

Event Tags

1. Build Event Vocabulary
    a. Take all the relevant tweets.
    b. Take out the most frequent words (using a threshold frequency) Based on their stemmed counterparts
    c. Inference from Data:
    close(5) , closed(114), closes(1), closing(2), closely(3) = close(125)
    Closures(106),closure(240) = closur(346).
    d. Create a Vocabulary of frequent words.
    e. Manually remove the non relevant words related to road closure or

Stem	Words	Frequency
pleas	please pleased	670
report	report reported reporting reports	622
dot	dot	360
closur	closures closure	346
lane	lane lanes	332
locat	location locations located	305
contact	contact contacting contacted	287
pl	pl pls	277
free	free	233

Event Tags

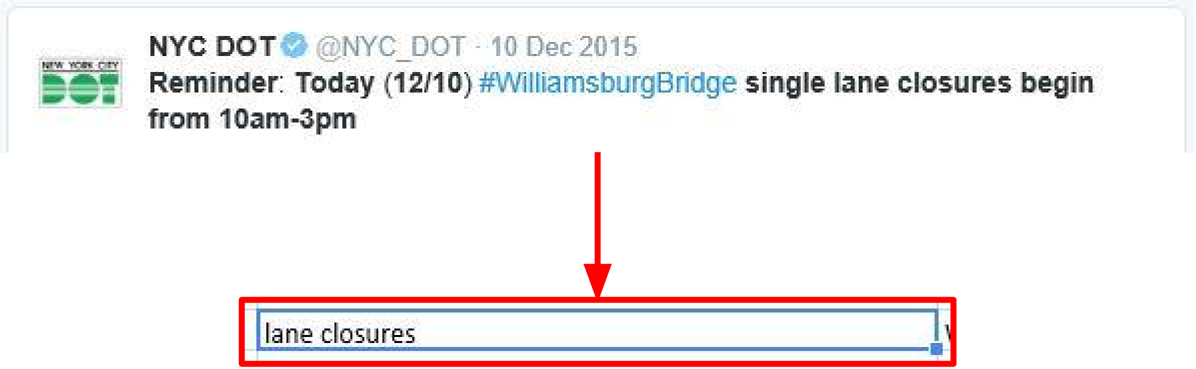


Frequency Distribution of Stemmed Words

Event Tags

2. Compare Incoming Tweet
   a. Tokenize, Normalize Tweet
   b. Check if each token is in Vocab
   c. If yes, include in 'name'
   d. List tokens in the order of occurrence.

Stem	Words	Frequency
closur	closures closure	346
lane	lane lanes	332



Event Date - Time Extraction

Example Tweets with Date/Time

1. 12/5 thru 12/20, 7am-3pm.
2. Today lane closures begin from 10am-3pm
3. 12/16-12/18 12:01-5AM & 12/19 1-6AM

Step 1	Retrieve dates by REGEX + NER
Step 2	Retrieve time by REGEX
Step 3	Delete duplicate dates
Step 4	Assign times to relevant dates

NYC DOT @NYC_DOT · 7 Dec 2015
Nightly #ManhattanBridge vehicular lane closures next week: **12/14+12/15 10PM-5AM**, **and 12/16 12:01AM-5:40AM**

14-12-2015 and 15-12-2015	10pm-5pm
16-12-2015	12.01am-5:40am

Event Date - Time Extraction

Pain points -

1. 12am - Noon
2. Two location, two date/time

NYC DOT @NYC_DOT · 11 Nov 2015
#PelhamBridge **11/13 Partial Lane Closures: N/B 10AM-Noon; S/B 12-3PM.**
Alt Route: #HutchinsonRiverParkway

3. One date, two time

NYC DOT @NYC_DOT · 20 Nov 2015
#RooseveltIslandBridge maintenance on 11/24: lane closure **from** 7am-3pm,
15min full bridge closures from 10am-2pm

4. Time ranging from one day to another
5. 1 date , 1 weekday mentioned

NYC DOT @NYC_DOT · 12 Nov 2015
#BrooklynBridge pedestrian/#bikenyc path **will be closed tonight 9PM until Fri 6AM**. Please use #ManhattanBridge as alternate route.

6. Time Mentioned before a Date

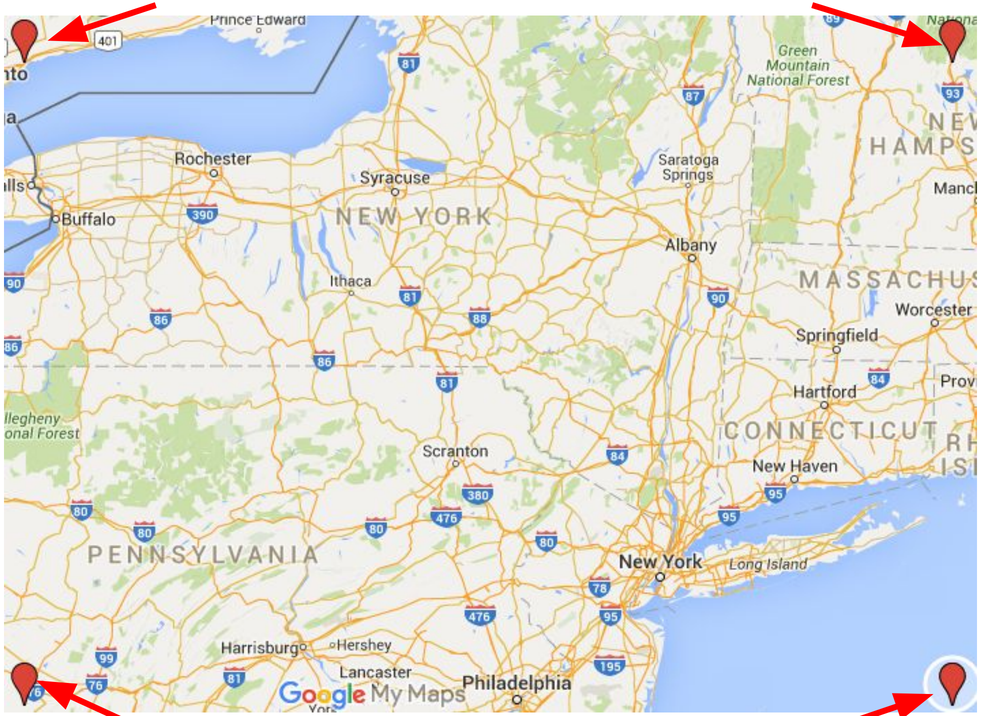
NYC DOT @NYC_DOT · Apr 1
Extended: #BQE nightly lane closures from 12:01-5AM in both directions will now continue through April 30.

Event Location Extraction Algorithm

Step 1	**Relevant Tweets** coming from classifier
Step 2	Generate **Candidate words** from - Hashtags - Check text words in online Library
Step 3	Generate possible **Geographical Coordinates** corresponding to all the candidate points
Step 4	Check the Coordinates against the **boundary coordinates of New York**
Step 5	Return the candidate words which fall **within the coordinates of New York**
Step 6	Return the **Actual Address** corresponding to the coordinates if the subsequence matches.

[43.818746, -79.133970]

[43.818746, -71.653065]



[39.914668, -79.133970]

[39.914668, -71.653065]

Event Location Extraction Algorithm

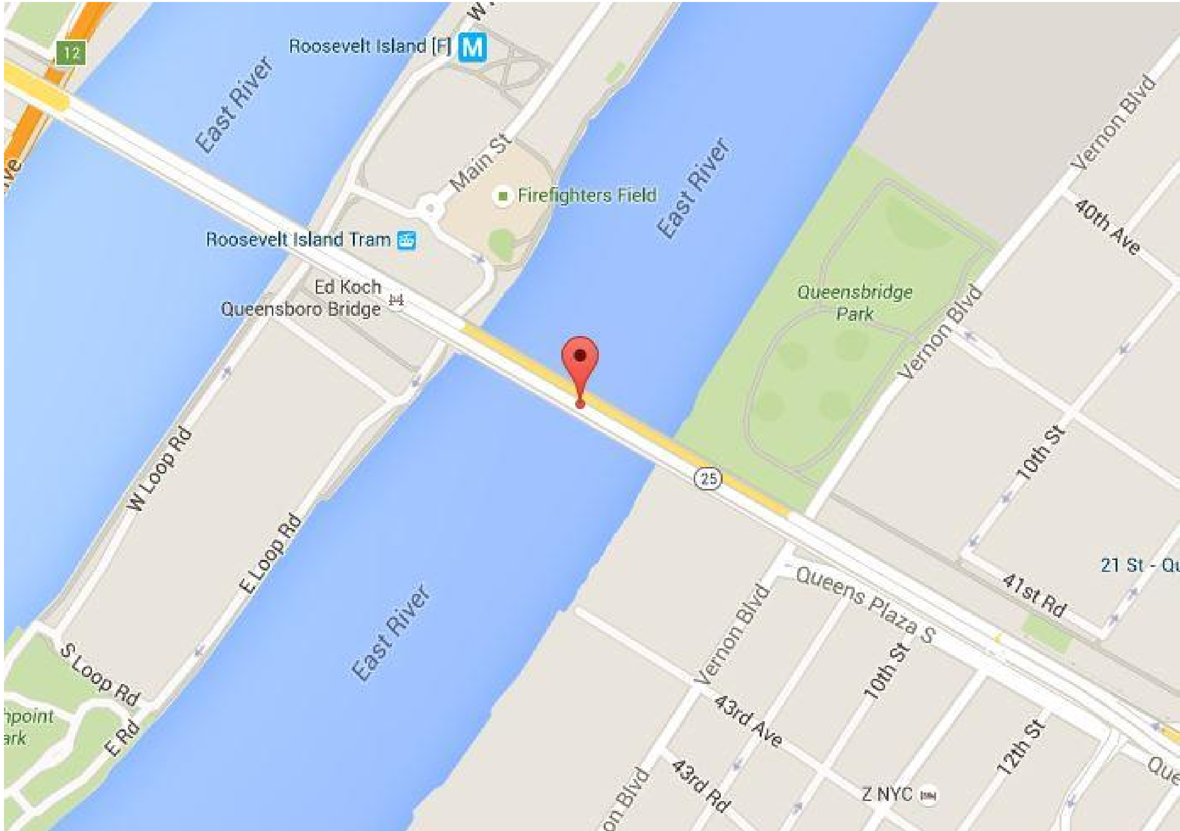
QueensboroBridge

↓

(40.75573455, -73.9518931501921)

↓

Ed Koch Queensboro Bridge Upper Level



Evaluation

Inter-Annotator Agreement

Agreement	0.9191919192	Sanjeev		
Expected Agreement	0.6510560147	Relevant	Non-Relevant	
Datta	Relevant	18	8	26
	Non-Relevant	0	73	73
Kappa	0.7684210526	18	81	99

Agreement	0.9090909091	Sujan		
Expected Agreement	0.6558514437	Relevant	Non-Relevant	
Datta	Relevant	17	9	26
	Non-Relevant	0	73	73
Kappa	0.7358434628	17	82	99

Average Kappa
0.8231880667

Landis and Koch (1977)
0.0 – 0.2 : slight
0.2 – 0.4 : fair
0.4 – 0.6 : moderate
0.6 – 0.8 : substantial
0.8 – 1.0 : perfect

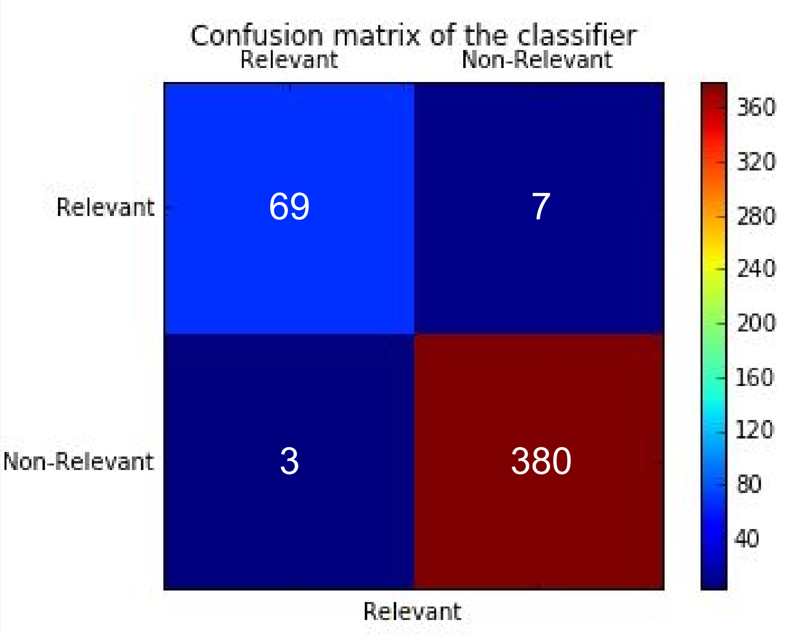
Result

Perfect

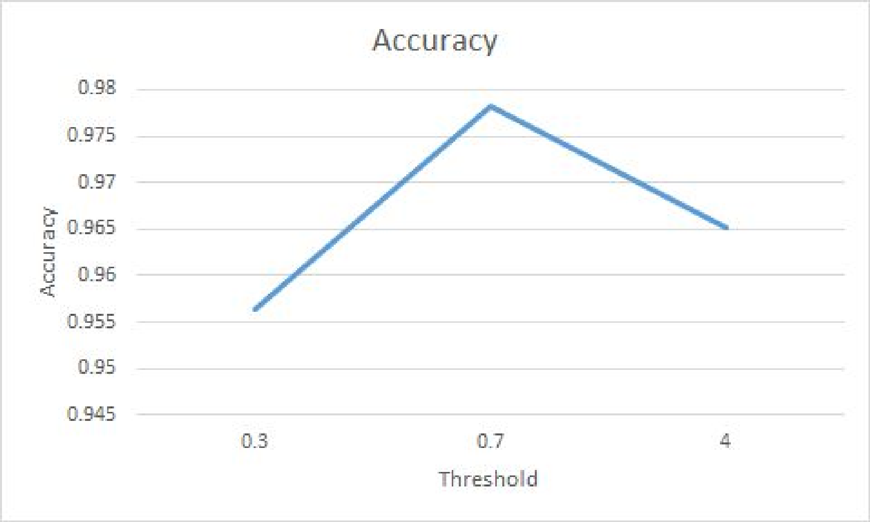
Classification Performance

Confusion Matrix

-

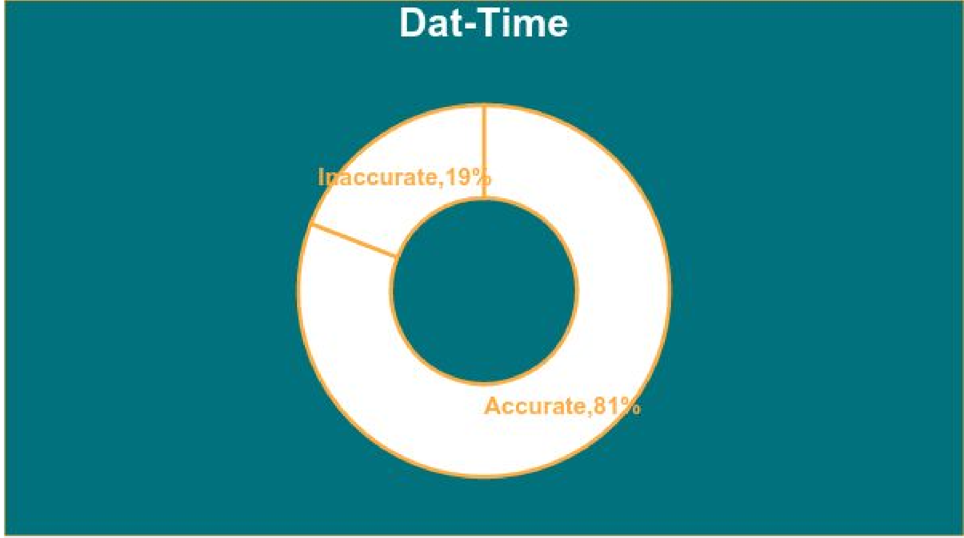


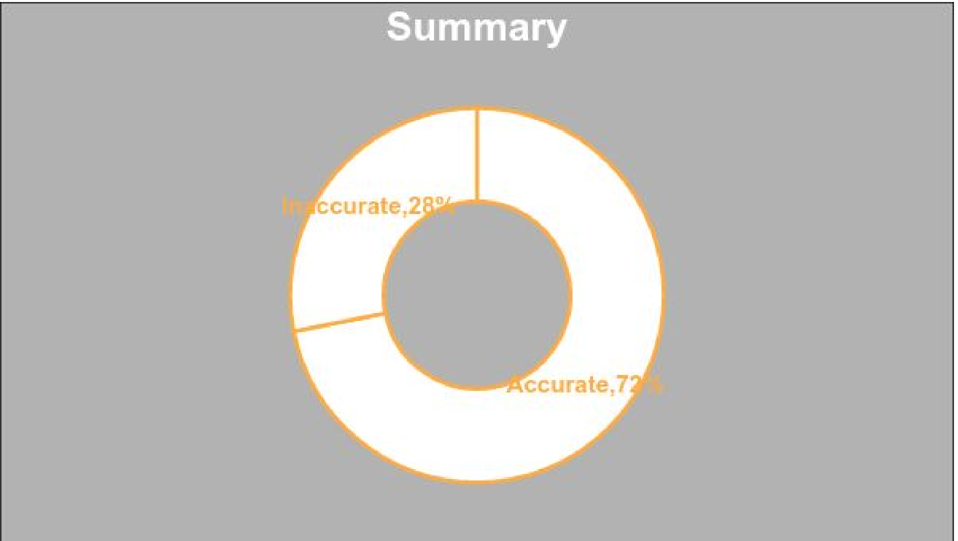
Accuracy Measure

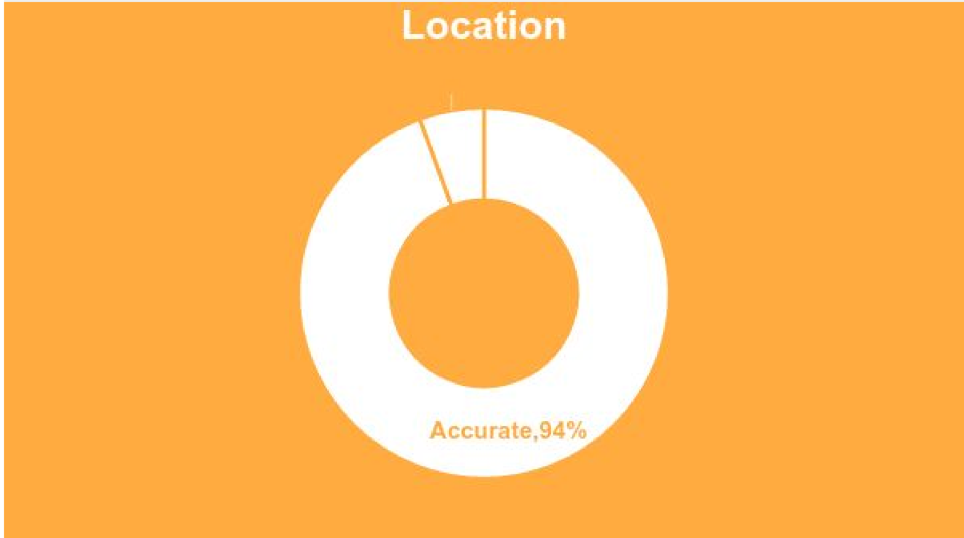


Threshold Selected	0.7
Cross Validation Accuracy	0.975
Test Accuracy	0.91

Extraction Measure







Extraction Measure

	Summary	Location	Date-Time	
Accuracy	Y	N	N	0
	N	Y	N	16
	N	N	Y	2
	Y	Y	N	17
	Y	N	Y	7
	N	Y	Y	31
	Y	Y	Y	104
	N	N	N	1
	128	168	144	178

Demo

References and Modifications

Appendix : Past Researches

Paper	Event Detection	Information Extraction
TEDAS: Twitter Based Event Detection and Analysis System[1]	- Twitter function based features - Crime and disaster specific words	- Rank information based on the user attributes, content
EvenTweet : Online Localized Event Detection from Twitter[2]	- Burstiness of word - Persistence of word for a significant time	- Entropy based Spatial signature - Cosine similarity for signature

1.TEDAS: a Twitter Based Event Detection and Analysis System , Rui LI, Kin Hou Lei, Ravi Khadiwala , Kevin Chen-Chuan Chang
2.EvenTweet: Online Localized Event Detection from Twitter,Hamed Abdelhaq, Christian Sengstock, and Michael Gertz

References

- Python - Pandas, NLTK , Geopy
- http://homes.cs.washington.edu/~mausam/papers/kdd12.pdf
- http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html
- TEDAS: a Twitter Based Event Detection and Analysis System , Rui LI, Kin Hou Lei, Ravi Khadiwala , Kevin Chen-Chuan Chan
- EvenTweet: Online Localized Event Detection from Twitter,Hamed Abdelhaq, Christian Sengstock, and Michael Gertz

Thank You!!

Appendix : Past Researches - Incomplete Slide

1. Date Extraction: NER cannot be used because NUM format 12/12 (unstructured text)
2. TIme extraction: NER - 12:30 pm. Even then most of the times it recognises as JJ
3. Location of the tweet cannot be used. And OOV words used. BKLN Bridge, thus subsequence.
4. Classification - not only bag of words, date-time present, week present

Rough Slide : Not to be presented

Draw a tree diagram with 100% training data on the top

We are assuming, that the relevant tweets all have the Events, Location, Time information

Create a matrix which will have

The accuracy of information extraction, will determine which decisions can be taken based on the information

- For example without time information we cannot take a time related decision
- E.g. if we cannot get the location information, then we cannot take a judge on where to go

Examples of Extraction of Location from Tweets

At the end, reference of the papers which have been read. The papers were read we cannot use them

- Data set not matching
- Some greedy approach is not working
- Idea borrowed from there
- Not have enough stream of data for a particular method to be implemented