



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-



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

NYC DOT (@NYC_DOT)

Twitter Handle:

- Information related to Transportation at NYC

Primary objective:

- Extracting Eventful Information.

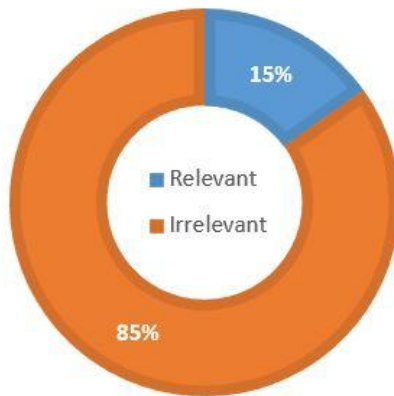
Event:

- NYC Road-Block/Closure

Information:

- Event Tag
- Date
- Time
- Location

RELEVANCE



254.5
tweets/month



NYC DOT @NYC_DOT · Apr 4

#65thStTransverse in **@CentralParkNYC**
closed both directions (except for
emergency vehicles) on 4/10 from 12:01-
6AM.

New York City 311 and NYCEM - Notify NYC

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

Data



Tweets

From NYC DOT Twitter Handle

Around 2000 tweets, with –

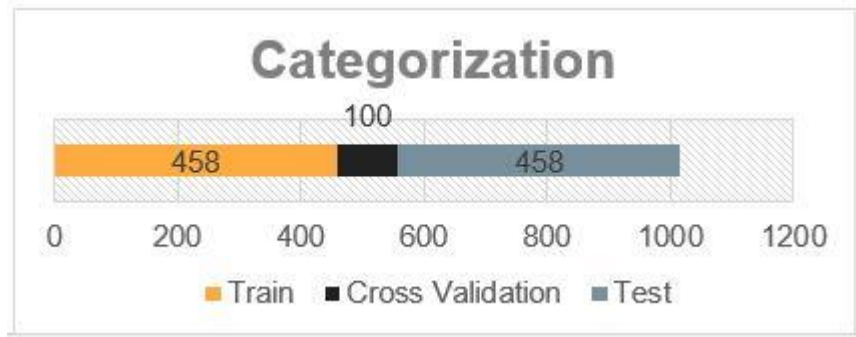
- Time of tweet
- Tweet Description
- Hashtags

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

Data



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

Tweets from
NYC Taxi
Handle

Ground Truth
mapping

Supervised classification

Bag of words
and features

Bayesian
Classification

Classified
Tweets
Relevant/Non-
Relevant

Information Extraction

Rules for Event
Summary`

+

Geocoder for
location

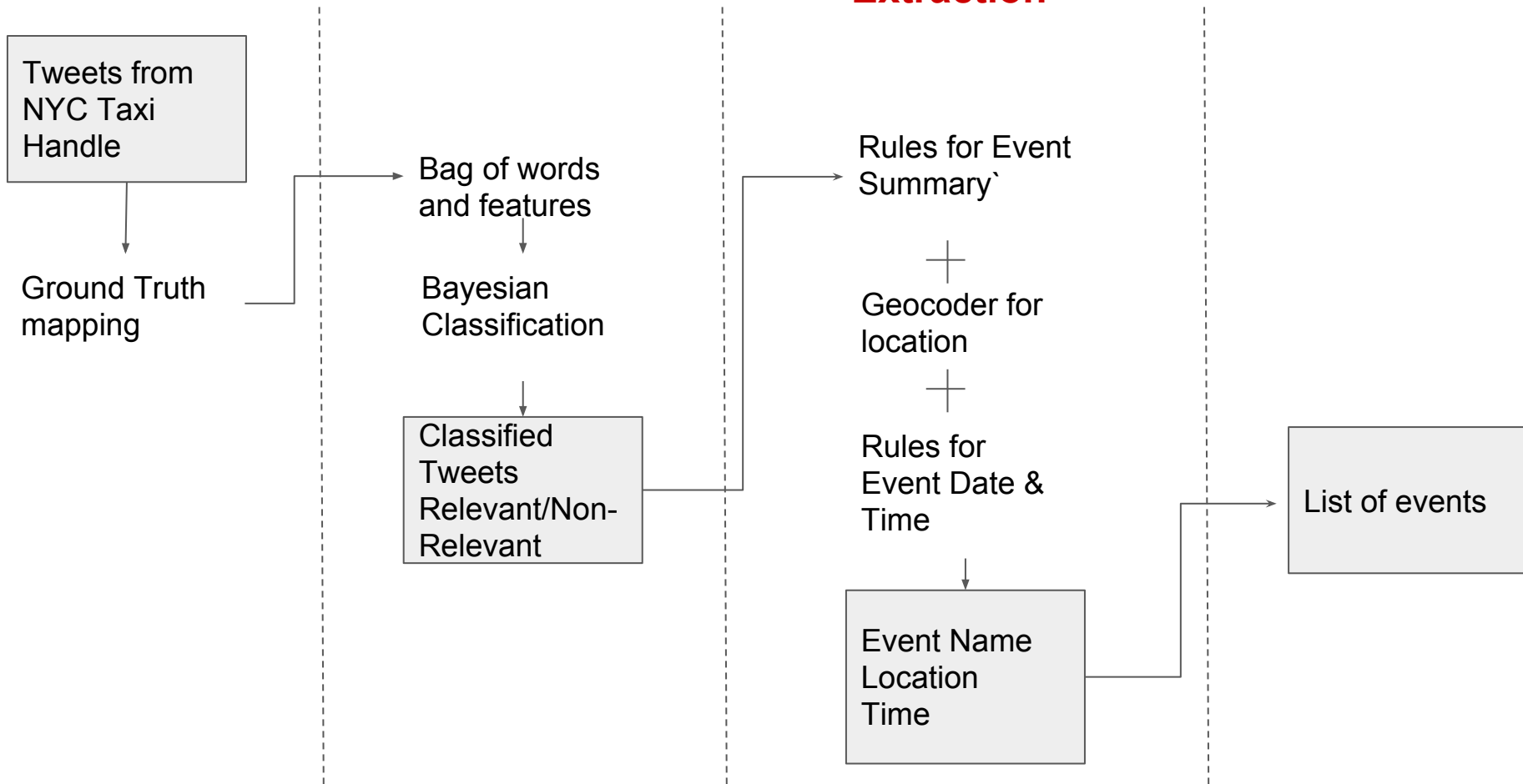
+

Rules for
Event Date &
Time

Event Name
Location
Time

Output

List of events



Ground Truth 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 Preprocessing

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 :

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

- Score indicative of Relevance prediction ; Appropriate Normalization

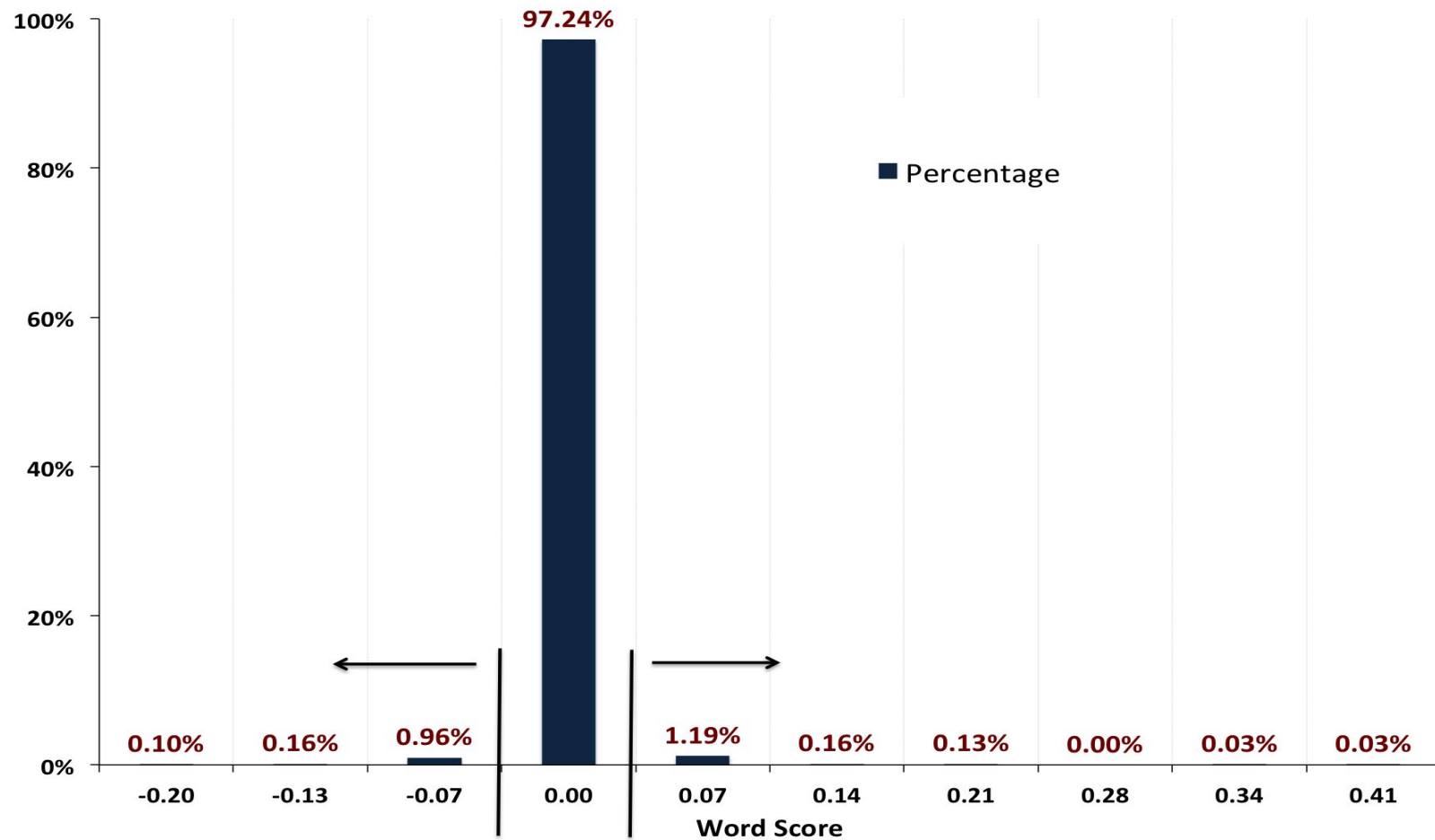
Step 2 :

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



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 $x=(x_1,\dots,x_n)$ representing some n features (independent variables).

$$P(c|x) = P(x|c) \cdot P(c) / P(x)$$

Event Tags

Objective:

Understand the type of event.

Example:

1. Full Closure
2. Both Direction



The image shows a tweet from NYC DOT (@NYC_DOT) dated 1 Sep 2015. The tweet text is: "#CentralPark #65thStreetTransverse full closure, both directions 9/2 & 9/3, 11pm-5am. Details below:". Below the tweet is a graphic from the NYC Department of Transportation. The graphic features the NYC DOT logo, the text "Department of Transportation", and "POLLY TROTTERBERG, Commissioner". A blue arrow with the word "EXTENDED" points to the title "Important Notice" in red. Below this is the title "Central Park – 65th Street Transverse" in blue. At the bottom, a table lists "Manhattan", "Community Boards 7 and 8", and "August 2015". The main text of the notice states: "On Wednesday and Thursday nights, September 2 and 3, the NYC Department of Transportation will finish replacing the barriers under the bridge carrying Center Drive over the 65th Street Transverse in Central Park, just east of the Carousel. The transverse will be fully closed to traffic from 11:00 p.m. until 5:00 a.m. the next day."

NYC DOT @NYC_DOT · 1 Sep 2015
#CentralPark #65thStreetTransverse full closure, both directions 9/2 & 9/3, 11pm-5am. Details below:

NEW YORK CITY
DOT
Department of Transportation
POLLY TROTTERBERG, Commissioner

EXTENDED **Important Notice**
Central Park – 65th Street Transverse

| | | |
|-----------|--------------------------|-------------|
| Manhattan | Community Boards 7 and 8 | August 2015 |
|-----------|--------------------------|-------------|

On Wednesday and Thursday nights, September 2 and 3, the NYC Department of Transportation will finish replacing the barriers under the bridge carrying Center Drive over the 65th Street Transverse in Central Park, just east of the Carousel. The transverse will be fully closed to traffic from 11:00 p.m. until 5:00 a.m. the next day.

Event Tags

Input:

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

Output:

Full closure both
directions



The image shows a screenshot of a tweet and a notice from the NYC Department of Transportation. The tweet, from @NYC_DOT, is dated 1 Sep 2015 and mentions a full closure of the 65th Street Transverse in Central Park on 9/2 and 9/3 from 11pm to 5am. Below the tweet is a notice from the NYC DOT, titled 'Important Notice' and 'Central Park – 65th Street Transverse'. The notice is dated August 2015 and mentions that the closure will occur on Wednesday and Thursday nights, September 2 and 3, from 11:00 p.m. to 5:00 a.m. The notice is part of a series of notices for Manhattan, Community Boards 7 and 8.

NYC DOT @NYC_DOT · 1 Sep 2015
#CentralPark #65thStreetTransverse full closure, both directions 9/2 & 9/3, 11pm-5am. Details below:

NEW YORK CITY
Department of Transportation
POLLY TROTTERBERG, Commissioner

EXTENDED **Important Notice**
Central Park – 65th Street Transverse

Manhattan Community Boards 7 and 8 August 2015

On Wednesday and Thursday nights, September 2 and 3, the NYC Department of Transportation will finish replacing the barriers under the bridge carrying Center Drive over the 65th Street Transverse in Central Park, just east of the Carousel. The transverse will be fully closed to traffic from 11:00 p.m. until 5:00 a.m. the next

Event Tags

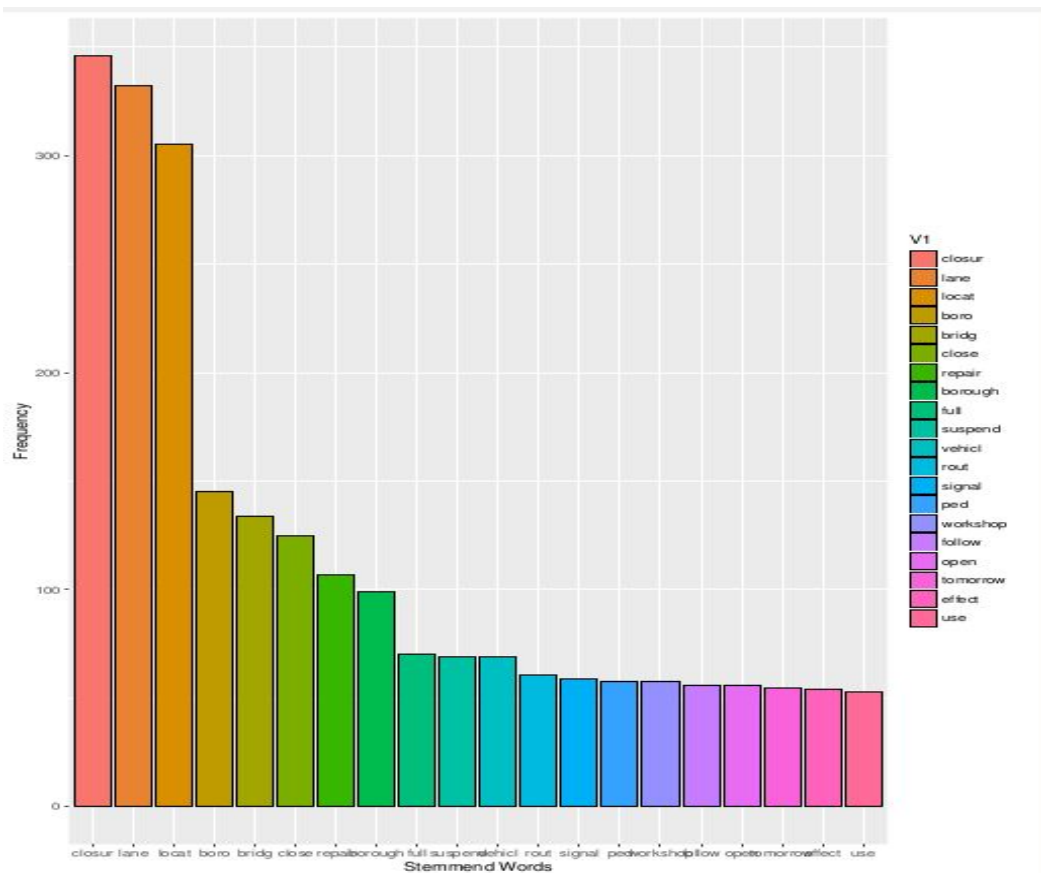
1. Build Event Vocabulary

- 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 blockage event

| 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 |

**Number in brackets represents their occurrence frequency.*

Event Tags



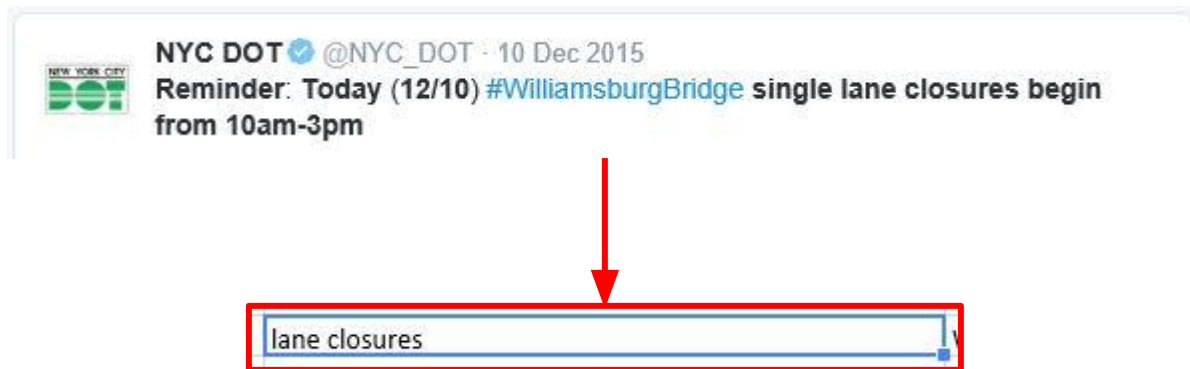
Frequency Distribution of Stemmed Words

Event Tags

2. Compare Incoming Tweet

- Tokenize, Normalize Tweet
- Check if each token is in Vocab
- If yes, include in 'name'
- List tokens in the order of occurrence.

| Stem | Words | Frequency |
|--------|------------------|-----------|
| closur | closures closure | 346 |
| lane | lane lanes | 332 |



The image shows a tweet from NYC DOT (@NYC_DOT) dated 10 Dec 2015. The tweet text is: "Reminder: Today (12/10) #WilliamsburgBridge single lane closures begin from 10am-3pm". A red arrow points from the tweet to a search bar below it. The search bar contains the text "lane closures".

NYC DOT @NYC_DOT · 10 Dec 2015
Reminder: Today (12/10) #WilliamsburgBridge single lane closures begin from 10am-3pm

lane closures

Event Date - Time Extraction

Example Tweets with Date/Time

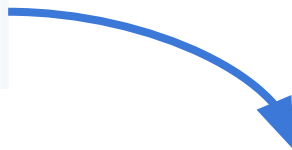
1. 12/5 thru 12/20, 7am-3pm.
2. Today lane closures begin from 10am-3pm
3. 12/16-12/18 12:01-5AM & 12/19 1-6AM

| | |
|--------|--------------------------------|
| 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
3. One date, two time
4. Time ranging from one day to another
5. 1 date , 1 weekday mentioned
6. Time Mentioned before a Date

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

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

 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.

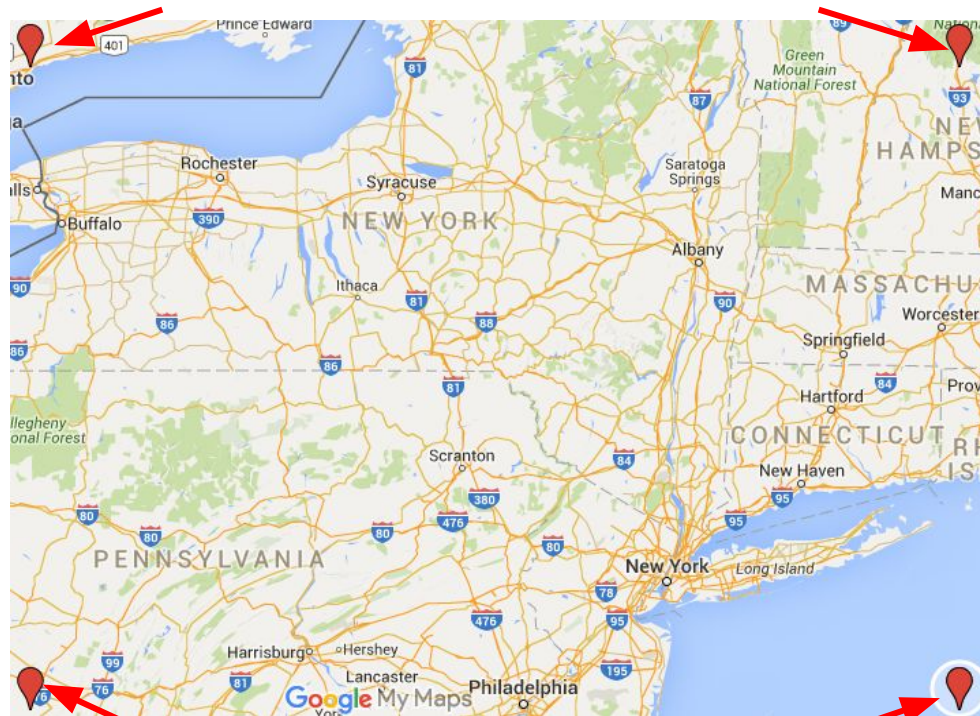
 NYC DOT @NYC_DOT · Apr 1
Extended: **#BQE** nightly lane closures
from 12:01-5AM in both directions will now
continue through April 30.

Event Location Extraction Algorithm

| | |
|--------|---|
| Step 1 | Relevant Tweets coming from classifier |
| Step 2 | Generate Candidate words from <ul style="list-style-type: none"> - 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]

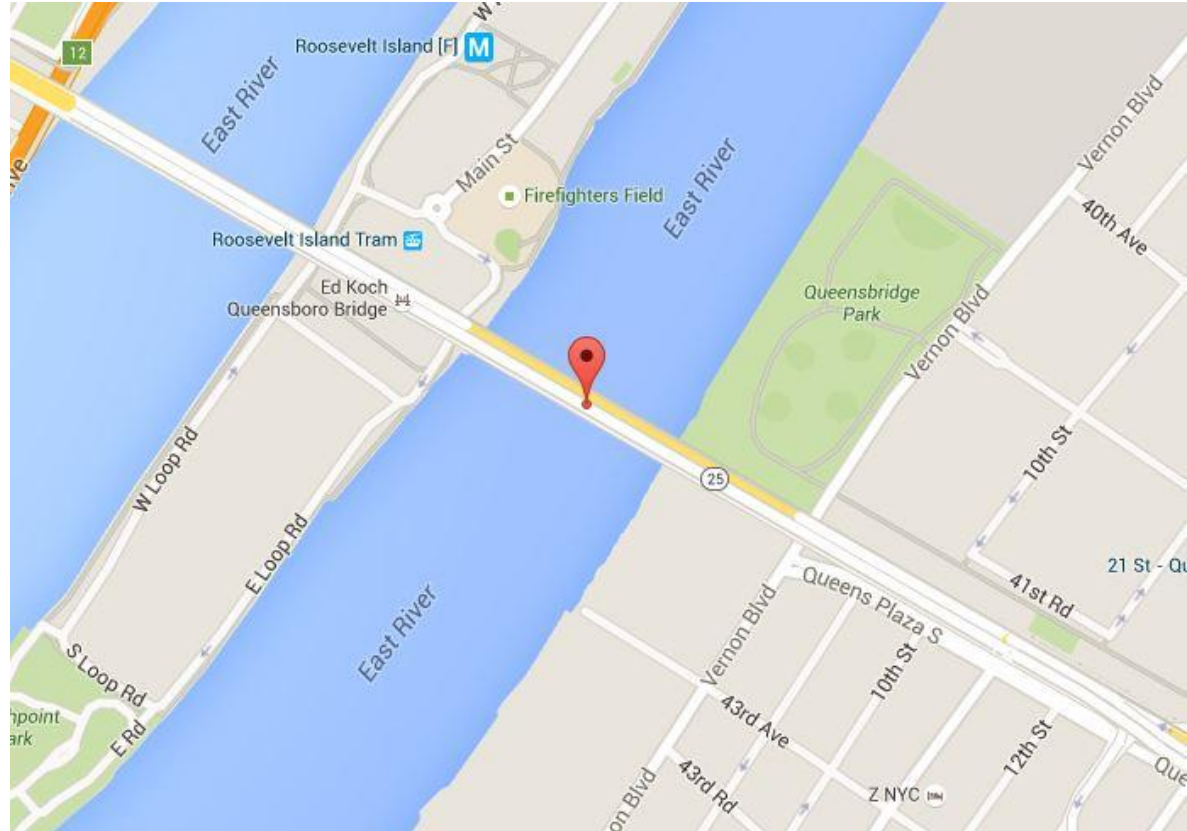
[43.818746, -71.653065]



[39.914668, -79.133970]

[39.914668, -71.653065]

Event Location Extraction Algorithm



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

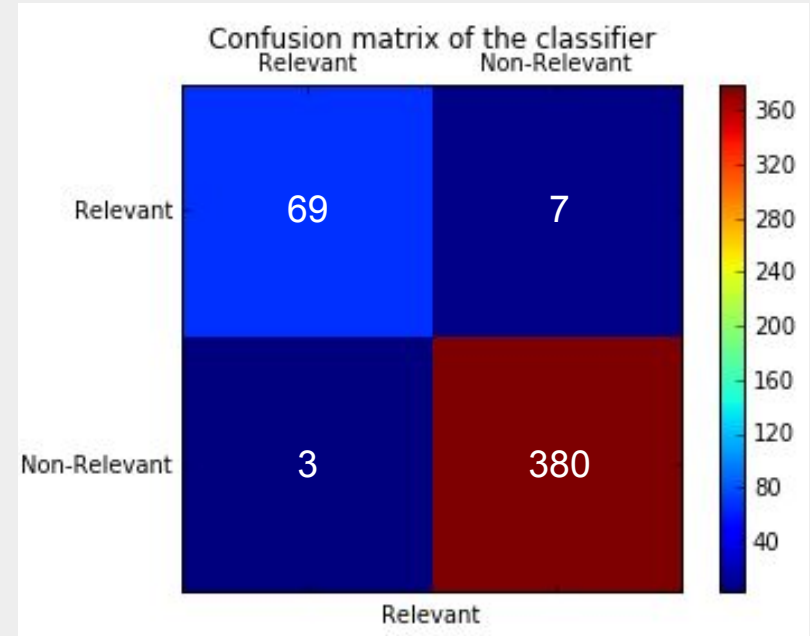
Result

Perfect

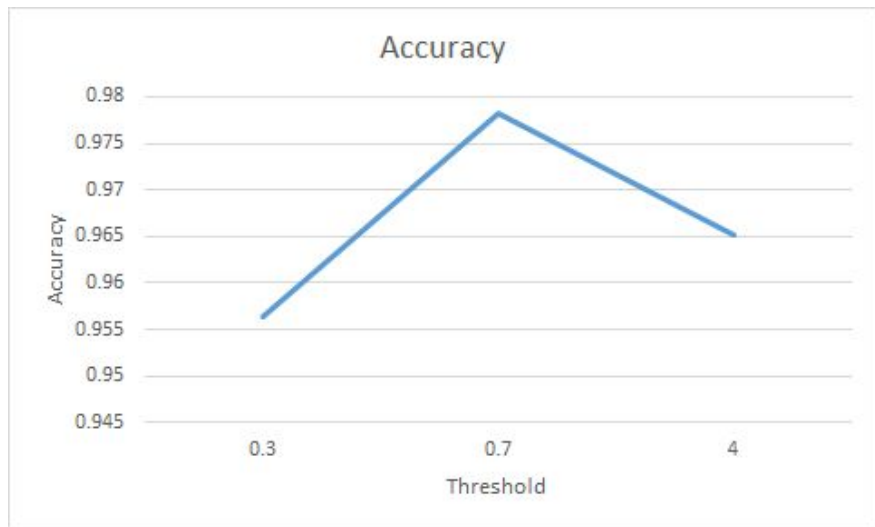
Classification Performance

Confusion Matrix

-



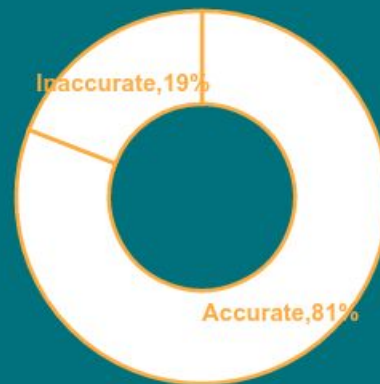
Accuracy Measure



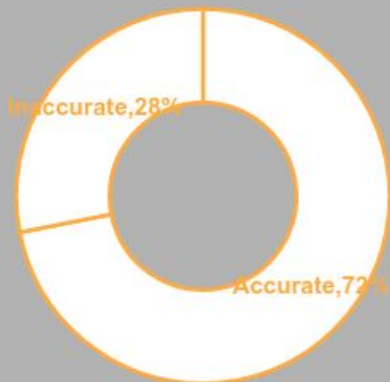
| | |
|---------------------------|-------|
| Threshold Selected | 0.7 |
| Cross Validation Accuracy | 0.975 |
| Test Accuracy | 0.91 |

Extraction Measure

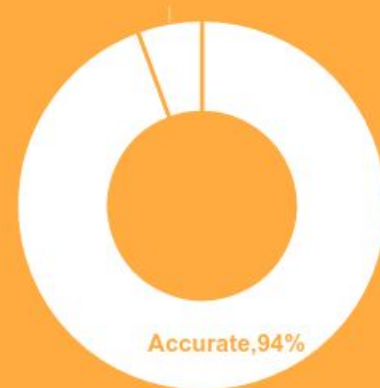
Dat-Time



Summary



Location



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 | <ul style="list-style-type: none">- Twitter function based features- Crime and disaster specific words | <ul style="list-style-type: none">- Rank information based on the user attributes, content |
| EvenTweet : Online Localized Event Detection from Twitter ² | <ul style="list-style-type: none">- Burstiness of word- Persistence of word for a significant time | <ul style="list-style-type: none">- 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!!