Extracting Knowledge Base from Text

Charu Mittal
Nikhil Jukar
Nilesh Singh

Agenda

- What is Machine Reading?
- Text Extraction
- Parts of speech tagging
- Classification
- Demo and Explanations

Introduction

- We seek to apply Natural language, Information Extraction, and Machine learning concept to build TwitterWordCloud.
- TwitterWordCloud is an application that takes the data from twitter and process it to gather important information.

https://github.com/nilesh892003/TwitterWordCloud

Text Extraction

Generating a large-scale knowledge base from the text

- Information extraction (IE) [6]
- Information Retrieval

Need of Text Extraction

- Text is Fundamental repository of human knowledge
- Most updated data source
- Enormous text data sources available
- Additional text production engine i.e Twitter, Facebook
- Finding relevant text is challenge

Resources/Tools

- Twitter data

- Twitter Natural Language Processing Tool

- Processing (Word Cloud)

Why it is Interesting

What are we doing

- Real Time Twitter based Machine reading System.
- Extracting useful structured representations of events.

Why It is Interesting

How are we doing

- Twitter NLP is used to process the tweet data.
- It extract the relevant words with number of occurrences in current tweets.
- Displays the data so the end result is clearly visible as word cloud

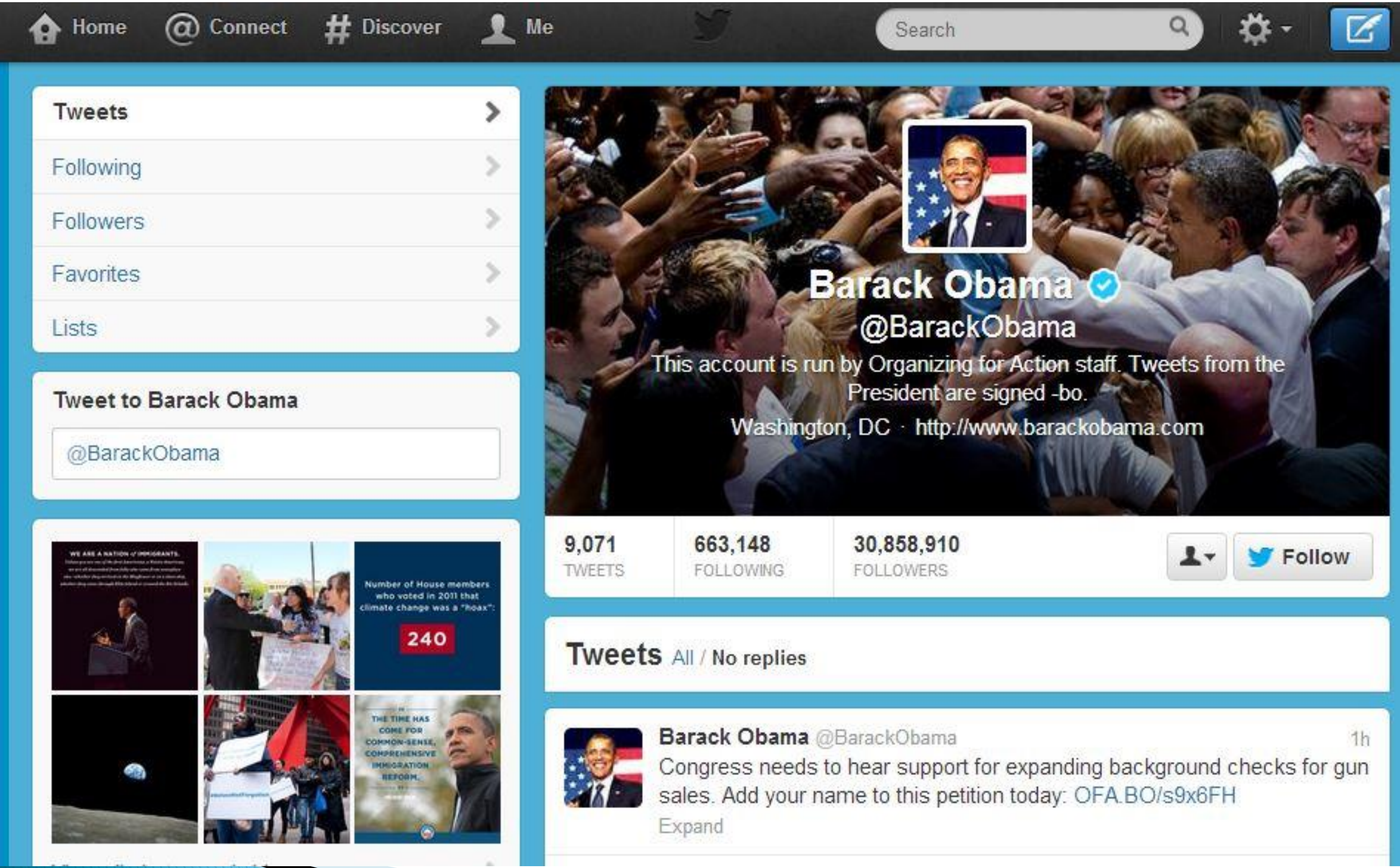
Why Twitter

- Tweets are the most up-to-date stream of information on current events
- short and self-contained nature of tweets
- The volume of Tweets is much larger than the volume of news article
- Free API to collect data

Tweet Format

- User can tag other user with ‘@’
- User can hash tag particular word with ‘#’
- Other user can see all tweets which have hash tagged words

Twitter Interface


Home
Connect
Discover
Me
Search
Tweets
Following
Followers
Favorites
Lists
WE ARE A NATION of IMMIGRANTS.
Number of House members who voted in 2011 that climate change was a "hoax":
240
THE TIME HAS COME FOR COMMON-SENSE, COMPREHENSIVE IMMIGRATION REFORM.
Follow

Tweet to Barack Obama

@BarackObama

Barack Obama
@BarackObama
This account is run by Organizing for Action staff. Tweets from the President are signed -bo.
Washington, DC · http://www.barackobama.com

9,071	663,148	30,858,910
TWEETS	FOLLOWING	FOLLOWERS

Tweets All / No replies

Barack Obama @BarackObama 1h
Congress needs to hear support for expanding background checks for gun sales. Add your name to this petition today: OFA.BO/s9x6FH
Expand

Challenges [1,2]

- Twitter users frequently mention mundane events in their daily lives
- Individual tweets often lack sufficient context to be categorized
- Temporal expressions: User can refer to same calendar date in different ways

  “Yesterday”, ‘Tomorrow”, “Last Friday”, “August 12th”
- Lexical Variation in Words:

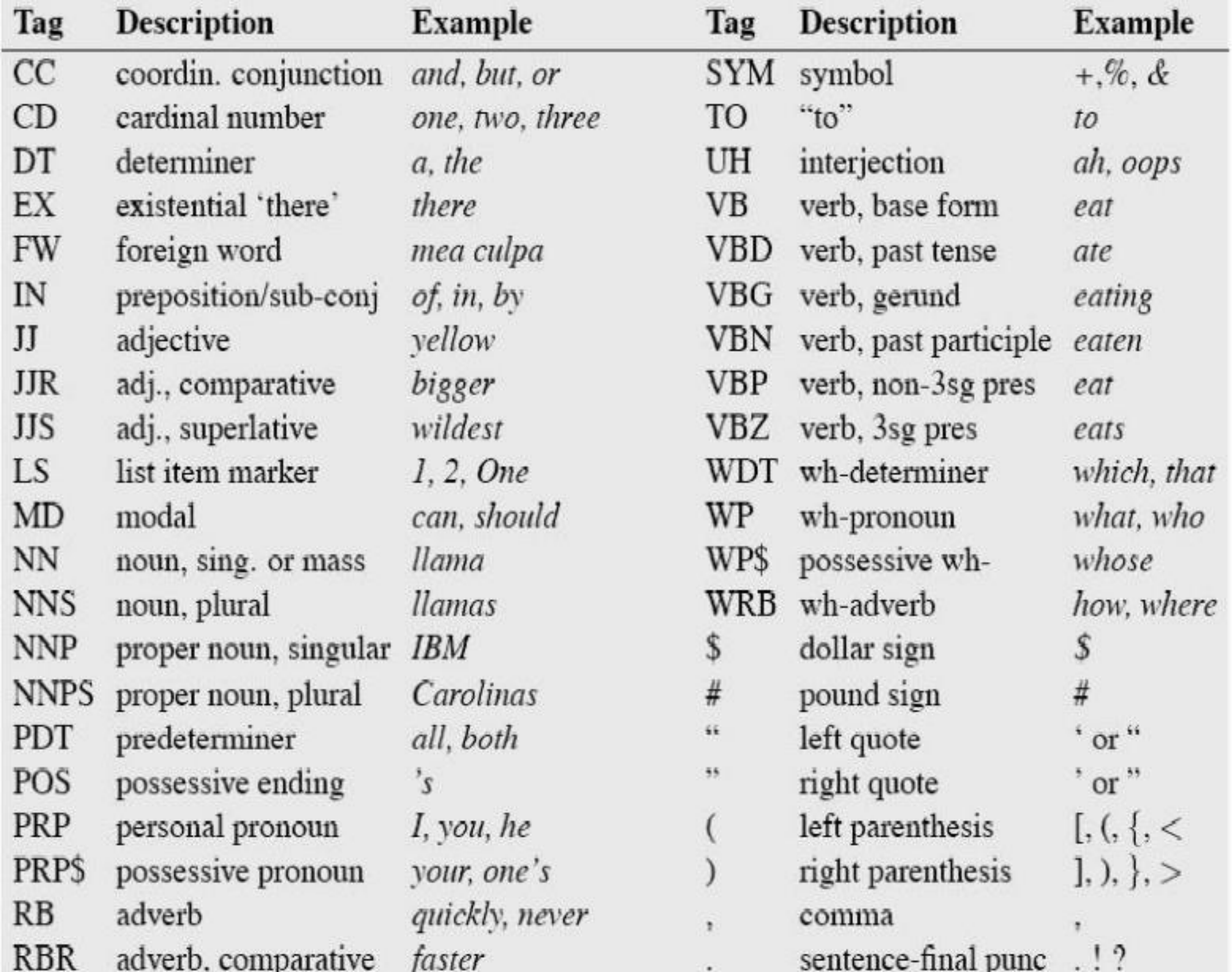
  Tomorrow : 2morrow, 2mro

  Wow: Woww, woow
- Poor performance for capitalization

Part of Speech(POS) Tagging

- Assign the correct part of speech (word class) to each word/token in a document

- We have used the Penn Treebank tag set which consist of 45 POS tags.[1]



Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	+,%, &
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	“	left quote	‘ or “
POS	possessive ending	*'s*	”	right quote	’ or ”
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?

Example of POS Tagging

“The/DT planet/NN Jupiter/NNP and/CC its/PPS moons/NNS are/VBP in/IN effect/NN a/DT mini-solar/JJ system/NN ,/, and/CC Jupiter/NNP itself/PRP is/VBZ often/RB called/VBN a/DT star/NN that/IN never/RB caught/VBN fire/NN ./.”

What are we interested in ?

- Nouns (Singular or plural)

eg: Jupiter,fire,Planet

- Adjectives(basic, comparative,superlative)
- Eg:Mini

- Adverbs

Eg:Beautifully

Ambiguities in POS tagging

- Current tools give almost a 90% accuracy
- “Around” can be a preposition, particle, or adverb
  - I bought it at the shop around/IN the corner
  - I never got around/RP to getting a car
  - A new Prius costs around/RB $25K

POS Tagging Approaches

- Rule-Based: Human crafted rules based on lexical and other linguistic knowledge.

- Learning-Based: Trained on human annotated corpora like the Penn Treebank.

  Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF), Transformation Based Learning (TBL)

Sequence Tagging (BIO encoding)

- B - Begin (Beginning a phrase)
- I –Inside (Inside a phrase)
- O –Outside (Outside a phrase)

- BIO encoding is used for encoding phrases (Named Entities, event phrases, and chunks)
- Assign the category to the word depending on the context. [2]

Example

- “The Town might be one of the best movies I have seen all year”

- After Applying BIO encoding:

  The/B Town/I might/B be/I one/B of/B the/B best/I movies/I I/B have/B seen/I all/B year/I

How To handle common Shortforms

- “fb” is commonly used for face book
- “ikr” is commonly used for I know right

- A bag of such words can be found at

http://www.ark.cs.cmu.edu/TweetNLP/

Twitter NLP

- Available at, http://github.com/aritter/twitter_nlp
- Tokenizes tweets
- Classify events
- Tag POS
- NER
- Tokenize events.

Combining all above concepts

- The output of our tool has the following format:
- GivenWord /Classifier with BIO /POS /BIO /Event relation
- Example:
- Town /0 /NNP /I-NP /I-movie
- Knicks /I-sportsteam /NNP /I-NP /O

Standard NLP vs Twitter NLP

- Degraded performance of standard NLP on twitter data.

- Twitter NLP redesigns the NLP pipeline by POS tagging, Classification, Chunking and NER.

- Twitter NLP`s F1 score is double than Standford NLP.

System requirement

- Operating System : Linux (Ubuntu), Windows 7
- Python IDE
- Java JDK
- Processing tool (http://processing.org/download/)

Twitter Data(REST Api)

- Twitter gives out recent tweets using RESTful web services
- It returns data in JSON format
- We have to parse this JSON file to text and remove important information i.e. Tweet text

Coding: Java

Twitter Data

- JSON file: 1 tweet record

{"completed_in":0.093,"max_id":328586609660399616,"max_id_str":"328586609660399616","next_page":"?page=2&max_id=328586609660399616&q=boston&rpp=100","page":1,"query":"boston","refresh_url":"?since_id=328586609660399616&q=boston","results":[{"created_at":"Sun, 28 Apr 2013 19:08:47 +0000","from_user":"EricGarment","from_user_id":285770446,"from_user_id_str":"285770446","from_user_name":"Eric Garment", "text":"RT @Mvlique: Its been 17 years and they still haven't found Tupac's or Biggie's killer\ud83d\udc82 but in 3 days they found the Boston bomber\ud83d\udca3, Oh.\ud83d\ude10\ud83d\ude12\u270b"}],"results_per_page":100,"since_id":0,"since_id_str":"0"}

From Tweet Text

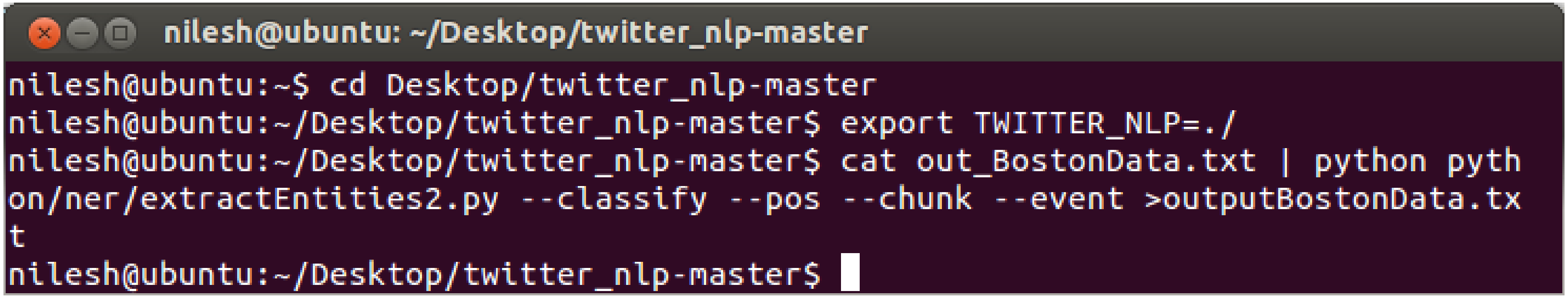
- Use Eclipse to setup and run twitterTweet.java
- The program will parse the JSON file into the normal tweet text file and save it as output.txt

- Output(Single tweet):

```
@stephenasmith lol who said the Knicks were too old, I guess Boston older then them cause look what they doing to Boston haha.
```

command

1. export TWITTER_NLP=./
2. cat TwitterBostonData.txt | python python/ner/extractEntities2.py --classify --pos --chunk --event >outputBoston.txt



```
nilesh@ubuntu:~$ cd Desktop/twitter_nlp-master
nilesh@ubuntu:~/Desktop/twitter_nlp-master$ export TWITTER_NLP=./
nilesh@ubuntu:~/Desktop/twitter_nlp-master$ cat out_BostonData.txt | python pyth
on/ner/extractEntities2.py --classify --pos --chunk --event >outputBostonData.tx
t
nilesh@ubuntu:~/Desktop/twitter_nlp-master$ 
```

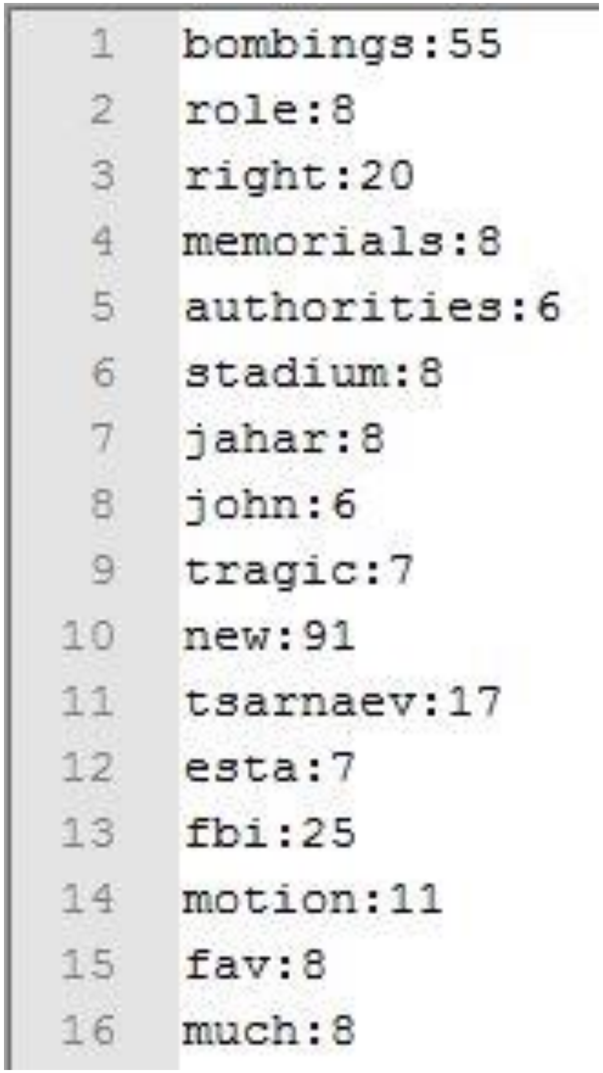
NLP output

- RT/O/RT/B-NP/O @olimpycs/O/USR/I-NP/O :/O//O/O Remember/O/VB/B-VP/B-EVENT when/O/WRB/B-ADVP/O people/O/NNS/B-NP/O made/O/VBN/B-VP/O manips/O/NNS/B-NP/B-EVENT with/O/IN/B-PP/O the/O/DT/B-NP/O boston/O/NN/I-NP/O bomber/O/NN/I-NP/O same/O/JJ/I-NP/O

Java Program (twitterWords.java)

- Extract information from the tagged NLP output
- Pass NLP output file to this program to extract personal Nouns from the tweet data
- Code calculated the occurrences of the Nouns and write the number of occurrence and noun in pair on .txt

Output



```
bombings:55
role:8
right:20
memorials:8
authorities:6
stadium:8
jahar:8
john:6
tragic:7
new:91
tsarnaev:17
esta:7
fbi:25
motion:11
fav:8
much:8
```

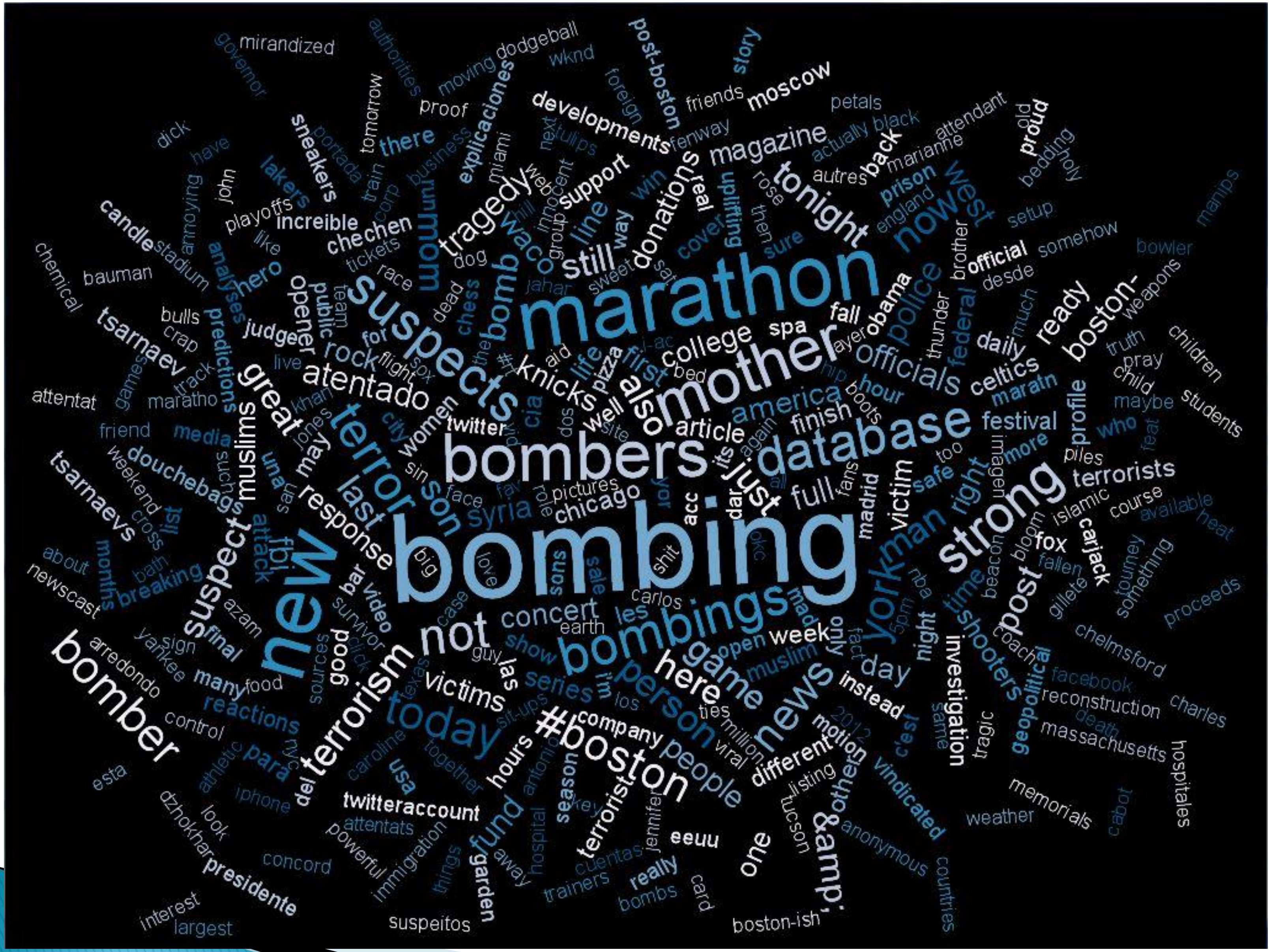
Processing IDE

- Processing is an open source programming language and environment
- Used to make Interactive programs using 2D, 3D and PDF output
- OpenGL integration for accelerated 3D

Setup Processing

- Copy processing setup folder anywhere in your computer.
- Open folder and start processing.exe Click file->preference and select your sketchbook location.
- Place processing folder to your sketchbook location.
- Click on open and import ".pde" file

Processing output



Future Scope

- Problem: In the TwitterWordCloud does not combine similar words into one.

  Eg: Similars and Similar

- The concept can be improved by taking consideration of lexical variation of the words.

Extracting and Resolving temporal expressions

- TempEx or SUTime can be used on the NLP processed data
- as input a reference date, some text, and parts of speech (from our Twitter-trained POS tagger) and marks temporal expressions with unambiguous calendar references.
- Efficiency: 94%

SUTime Output

- It provides output in XML format with temporal resolution.
- <TIMEX3 tid="t1" type="DATE" value="2013-04-26">4/26/13</TIMEX3>/O/NN/I-NP/O

Other Research

- Open Information Extraction[5]

  http://openie.cs.washington.edu/

- Twitter Calendar

  http://statuscalendar.cs.washington.edu/

- RevMiner: a novel smartphone interface that utilizes Natural Language Processing Techniques to analyze and navigate reviews

  http://jeffhuang.com/revminer_uist2012_preprint.pdf

What you should know?

- Extracting Knowledge Bases from Text is a concept of Machine Reading that on bigger scale is a concept of Machine Learning.
- It basically involves these process:

  1: Named Entity Segmentation

  2: Extracting Event Mentions

  3: Classification of Event Types

  4: Extracting and resolving temporal Expressions

References

1. Ritter, Alan and Clark, Sam and Mausam and Etzioni, Oren, Named Entity Recognition in Tweets: An Experimental Study, IN EMNLP,2011
2. Ritter, Alan and Mausam and Etzioni, Oren and Clark, Sam, Open Domain Event Extraction from Twitter, In KDD, 2012
3. I. Mani and G. Wilson. Robust temporal processing of news. In ACL, 2000.
4. http://nlp.stanford.edu/software/corenlp.shtml#SUTime
5. http://openie.cs.washington.edu/
6. Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni, Open Information Extraction from the Web, IJCAI 2011
7. Twitter Rest API:

   https://dev.twitter.com/docs/api