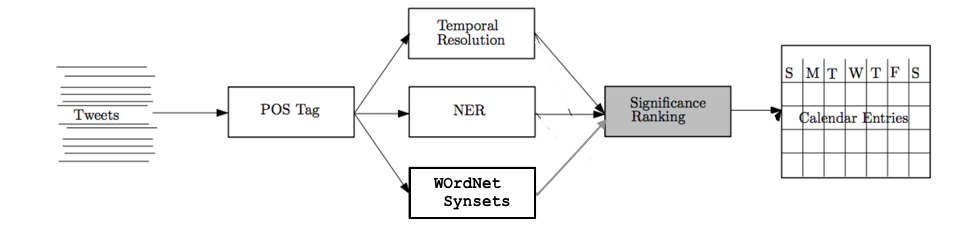
Project Report

Event Extraction from Twitter

Objective: Build Accurately extracting calendar of events - TWICAL

Current Challenges:

1. Twitters Noisy, disorganized data and informal way of writing tweets makes existing NLP tools to perform poorly on Tweets as opposed to well-formed News articles.
2. To differentiate between strong and week events i.e. someone posting a family gathering may not be contextual to everyone. So we have to identify such mundane events (which are tweeted a lot).
3. Managing the redundancy of Information for such a big scale as Twitter is a challenge as well.
4. To reduce complexity, we are focusing on most commonly occurring events on twitter like Birthday, Meeting, World Summit, Tech Conference, celebrity Death, launch/release etc.



*Fig 1. This is a block Diagram given in the Paper a little modified in my implementation.*

Steps for Event extraction Implementation:

1. **Tweet Extraction**: Using the Twitter stream API accept the broad continuous stream of Tweets, preprocessed tweets for mention of event date or any time suggesting tag since we are using events which are clearly associated with date rather than a span of time.
2. **POS tagging**: Tweets are needed to be tagged with appropriate tags. Here, we use NER with POS tagger from Stanford NER.

Event tagged like VBN, VBD are also removed since they happened in past.

Get location from Stanford Named Entity recognition or from Twitter geolocation.

1. **Events matching with our categories based on Synonymy from WordNet**:

Since I do not have access to already annotated data for event extraction, I am using bag of words approach to detect the event in the tweet. WordNet is used to handle the Synonymy for the considered events just as a semantic improvement over regular expression or simple POS tagging approach.

1. **Extracting & Resolving Temporal Expressions**: To get clear idea about when the event happened. This can be very tricky in 140 character Tweets since user never mentions whole context. I Used POS tagger & NLTK TempEx to get the absolute date mentioned in a phrase.

E.g. Takes Next Friday, August 12, tomorrow, today etc. and gives fixed date.

Also Filtered Events happened in Past since we have exact dates of such events.

Eg -> *this Saturday night at 9.30pm -> “2017-11-25”*

1. **Event Ranking and Categorization**: Strong association with a date and number of times same event occurs increases the event ranking. Categorization could be easily done using a classifier trained on annotated data or using supervised learning but it is outside the scope of the project.

TWICAL output format - Event\_Entity(Entity, phrase, date, type) and Yes or No for the event type classification.

Conclusion:

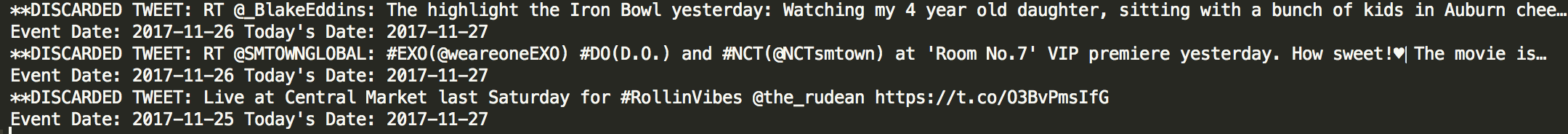
Twitter data could be tricky to fit into NLP tools trained on properly articulated data such as News articles. Accuracy on this could be improved using techniques like POS tagging trained on domain data, Semantic improvements like using WordNet knowledge and resolving temporal expressions for time etc.

Since events are extracted from continuous stream of tweets the system performance cannot be measured unless large corpus of tweets is manually Annotated. Although the reference paper I used for this project explains a lot on how above techniques increase F1 score and performs better in twitter domain particularly than using regular NLP tools.

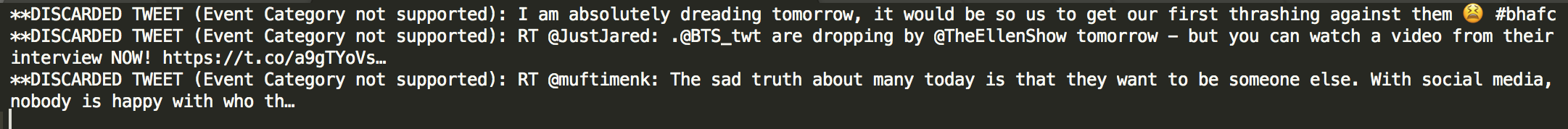
Paper referred: “Open Domain Event Extraction from Twitter” By Alan Ritter, Mausam, Sam Clark, Oren Etzioni.

Here are some of sample outputs of my program –

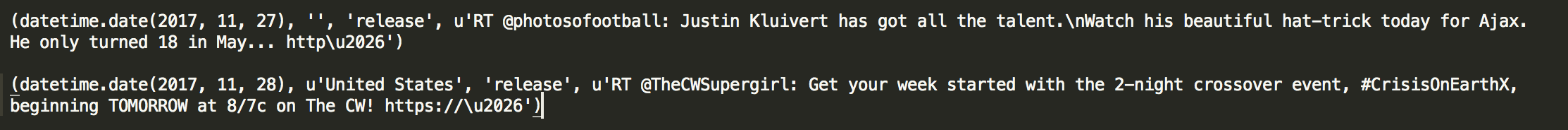
Dropping past events by temporal resolution of date from tweets:



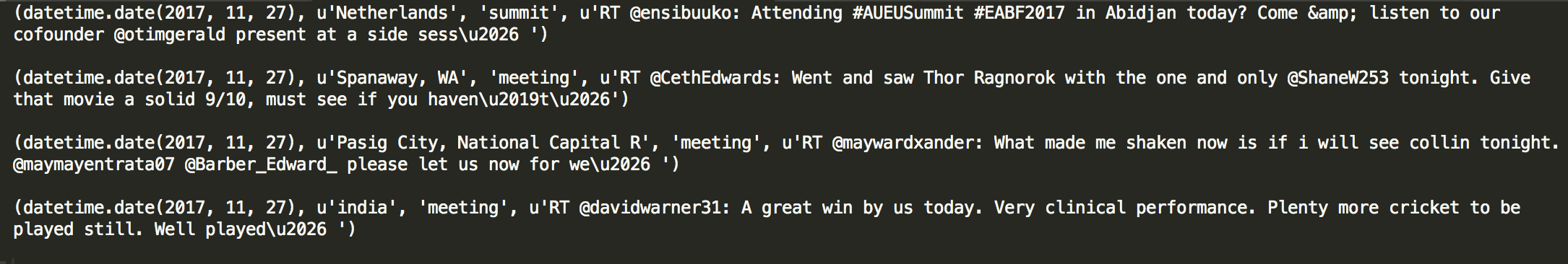
Dropping events that are not categorized as one of the events from considered event categories:



Improvements on event detection and Categorization with use of WordNet



Final Output of few Events:



References used for implementation:

Twitter Streaming API:

<http://piratefache.ch/twitter-streaming-api-with-tweepy/>

TimeX for temporal Expression Resolution:

<https://bebblebrox.wordpress.com/2011/01/23/temporal-expressions/>

<https://github.com/nltk/nltk_contrib>

POS tagging nltk:

<http://www.nltk.org/book/ch05.html>

WordNet used for Synonymy:

<http://www.nltk.org/howto/wordnet.html>

Stanford NER for Entity recognition

<https://nlp.stanford.edu/software/CRF-NER.shtml>

Demo of TWICAL by the Authors of Paper

<http://ec2-54-170-89-29.eu-west-1.compute.amazonaws.com:8000/>