Detecting New Extraordinary Events from Twitter in Real-time

TANG Ka Ki, WONG Cheuk Hei
Advised by Prof Pan HUI

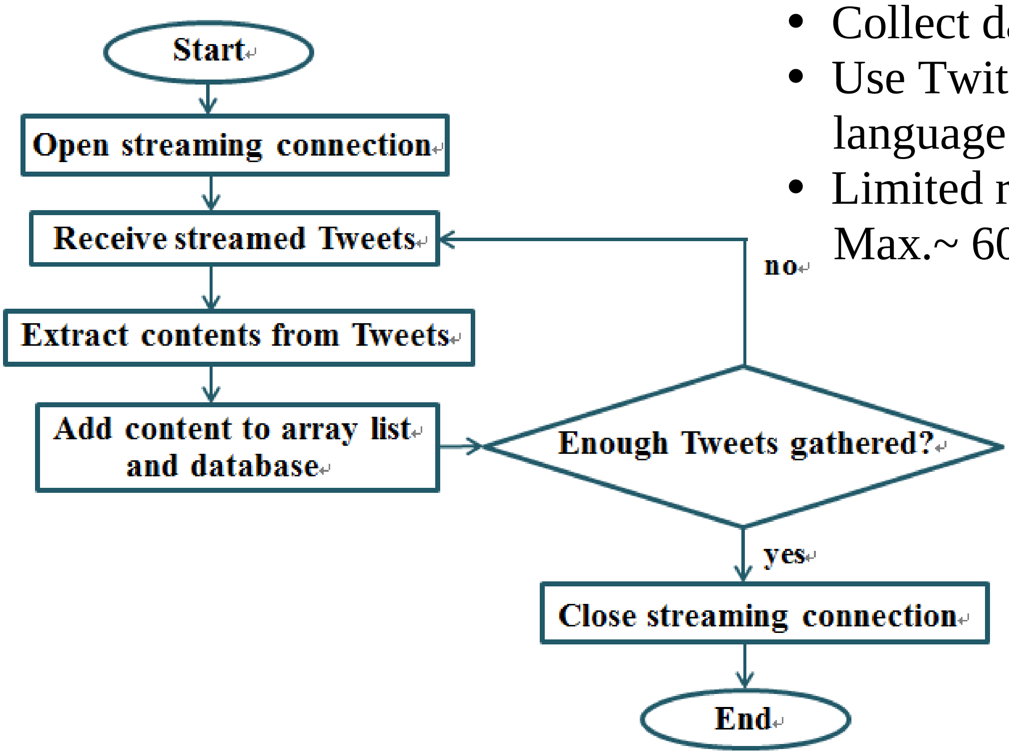
TWEETS	FOLLOWING	FOLLOWERS
500 million	2014	2015

Overview

Real time event detection system is an online analytics platform for Twitter. Twitter has enormous temporal dataflow. It contains 284 million monthly active users and about 500 million tweets flow per day. With the process of discovering interesting and **useful patterns and relationships** in large volumes of data, people can keep track of any special current new events happening in **New York** and **San Francisco** without scrolling down the Twitter’s browser all the time.

Data Grabber and Database

- Collect data by Twitter API
- Use Twitter 4J to filter language and location
- Limited rate of streaming Max.~ 60000 tweets each time



Start↵

Open streaming connection↵

Receive streamed Tweets↵

Extract contents from Tweets↵

Add content to array list↵ and database↵

Enough Tweets gathered?↵

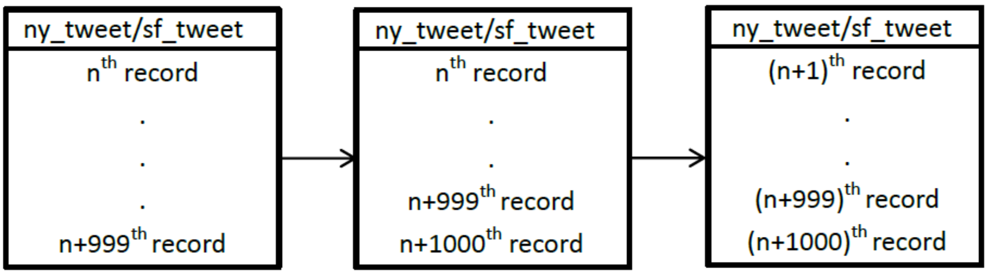
no↵

yes↵

Close streaming connection↵

End↵

- Store tweets into **database** according to locations
- Two tables’ size is fixed with 1000 records



ny_tweet/sf_tweet
n th record
.
.
.
n+999 th record

ny_tweet/sf_tweet
n th record
.
.
n+999 th record
n+1000 th record

ny_tweet/sf_tweet
(n+1) th record
.
.
(n+999) th record
(n+1000) th record

What's happening?

View 62 new Tweets

Data Preprocessing

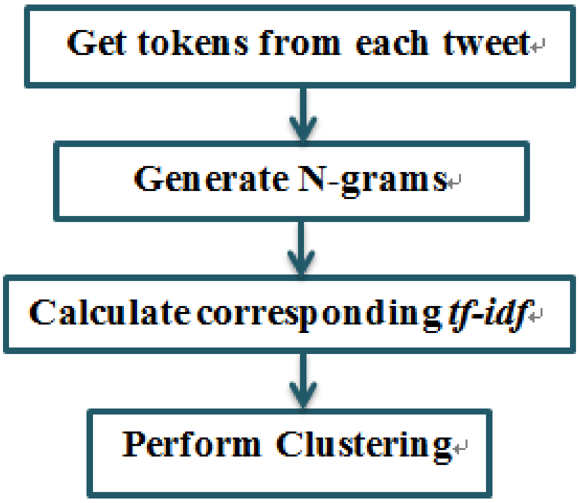
1. Tokenization

Remove Hashtags#, Addressing@, Uniform resources locator (URL), Punctuation characters, Unicode glyphs

2. Stemming

3. Remove Stop words

Data Analyzing



- **N-grams: (max. n=5)**

- Contiguous sequence of n items from a given sequence of tweet
- We use Java class String Builder to append the n-grams together

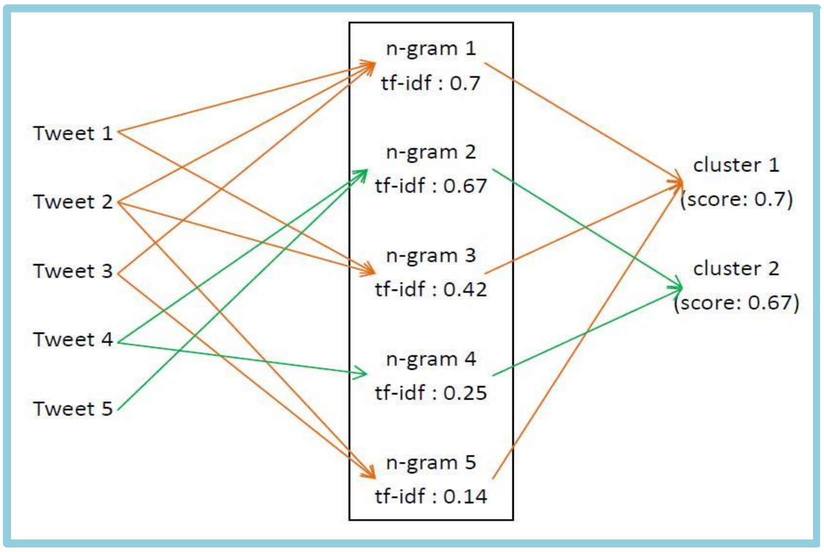
- ***Tf-idf* of an n-gram:**

- reflects the importance of that n-gram to that set of tweets

$$\boldsymbol{tfidf = tf \times idf}$$

tf : freq. of a token in its raw tweet
idf : no. of tweets in the set that contain that token

- higher Tf-idf -> more important word



Organization between Tweets, N-grams, Clusters

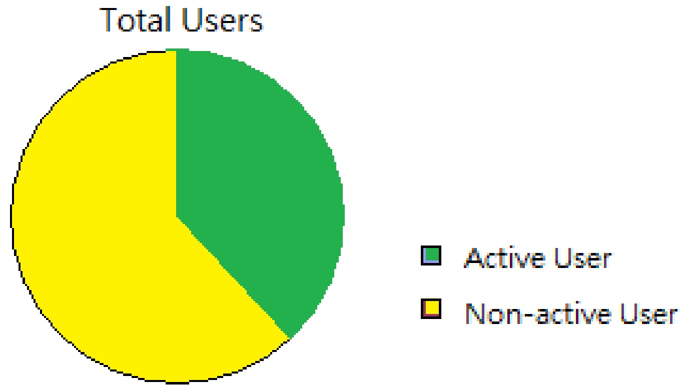
- Compare the *tf-idf* of tokens within a tweet to obtain the one and assign it to its corresponding n-gram.
- Obtain the maximum score within each n-gram and compare their similarities.
- Perform the agglomerative hierarchical clustering
- Each cluster represents a single topic.

9 15

Features

- Active users VS Inactive users

Active users: 1/3 of the users follow also follow he/her back



New York New Events!

San Francisco Hot Topics!

Evaluation

- Data grabber can extract **real time tweets** from Twitter
  100000 tweets from whole world without errors in 42 minutes

- Database can handle 100000 records without errors

- Data preprocessing tools can successfully **filter out more than 50% of noisy data** from the raw messages

- Discover many n-grams with *tf-idf* = 0

- Observe potential hot events

Conclusion

Real time event detection system successfully grabs and tokenizes data and detects potential new event every 16 minutes from Twitter. Due to the limited rate problem of grabbing data, system are recommended to restart again from 17:00 (UTC+8) to 22:00 (UTC+8) as least tweets grabbed at that time.