Finding Trending Hashtags and Time Series Analysis in Twitter

Project Report

Team Members:
KRISHNA GAPILESWAR V R – 2015103513
NAVEEN N S – 2015103561

Abstract:

With advent of micro blogging sites, the data that is generated is humungous. Twitter is one of the major players of micro blogging sites. Every second, on average, around 6,000 tweets are tweeted on Twitter, which corresponds to over 350,000 tweets sent per minute, **500 million** tweets per day and around **200 billion** tweets per year. On average, around **6,000** tweets are tweeted on Twitter every second. These tweets are a treasure of data that can be analysed to find trending topics, mindset of the masses and even influence the most sought after political elections of the world. Our project is aimed at finding the hashtags in the tweets both in our home timeline and in a specific user's timeline. The hashtags are obtained from tweets and analysed for their frequency and the time period in which the users tweet the most.

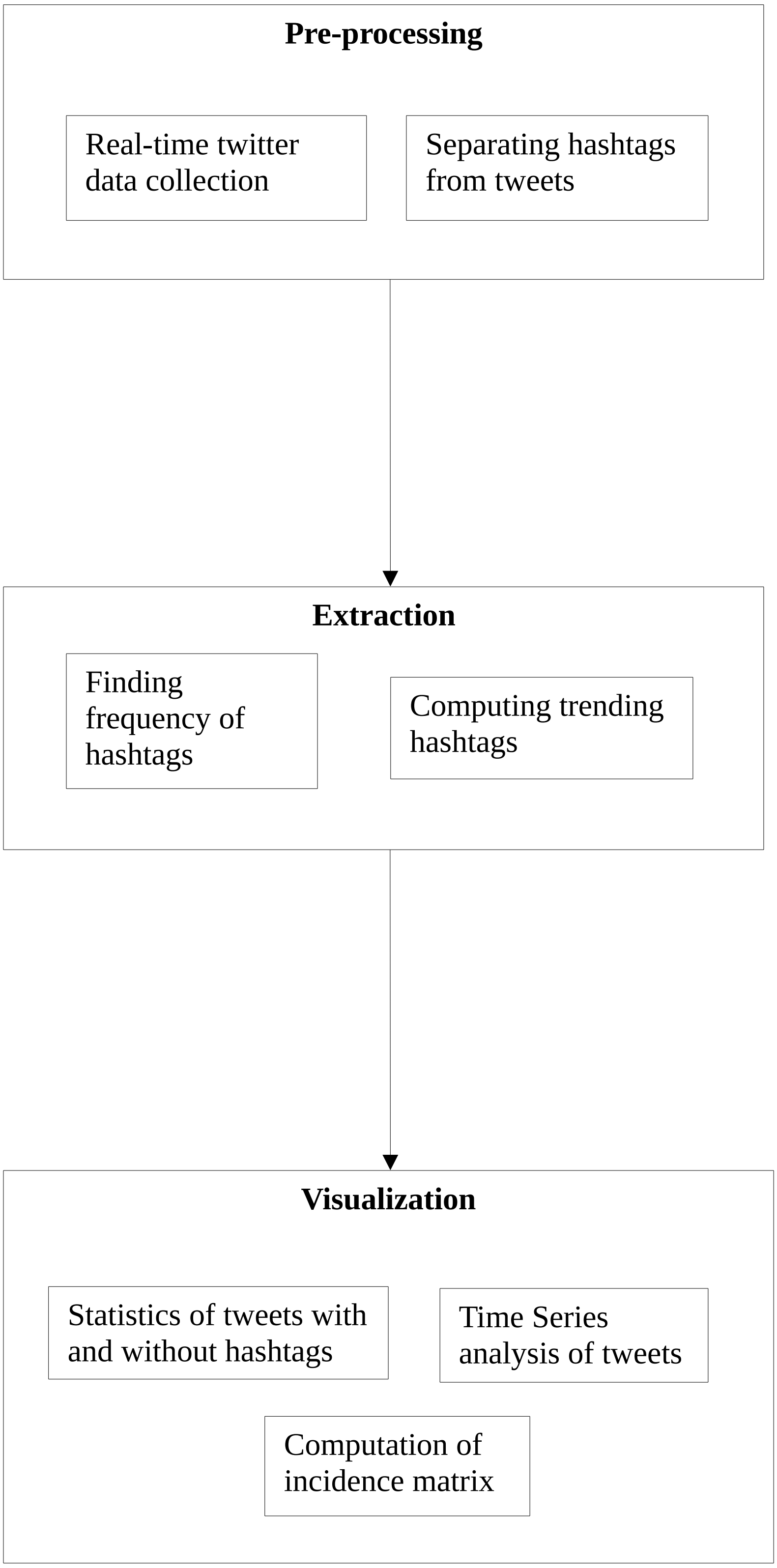
Existing methods:

- Twitter Developer API.
- Python offers tweepy module for this purpose.
- In general, several websites like sprout social, hashtagify to find the trending hashtags.

Modules:

- Pre-processing
- Extraction
- Visualization

Block Diagram:



Pre-processing
Real-time twitter data collection
Separating hashtags from tweets
Extraction
Finding frequency of hashtags
Computing trending hashtags
Visualization
Statistics of tweets with and without hashtags
Time Series analysis of tweets
Computation of incidence matrix

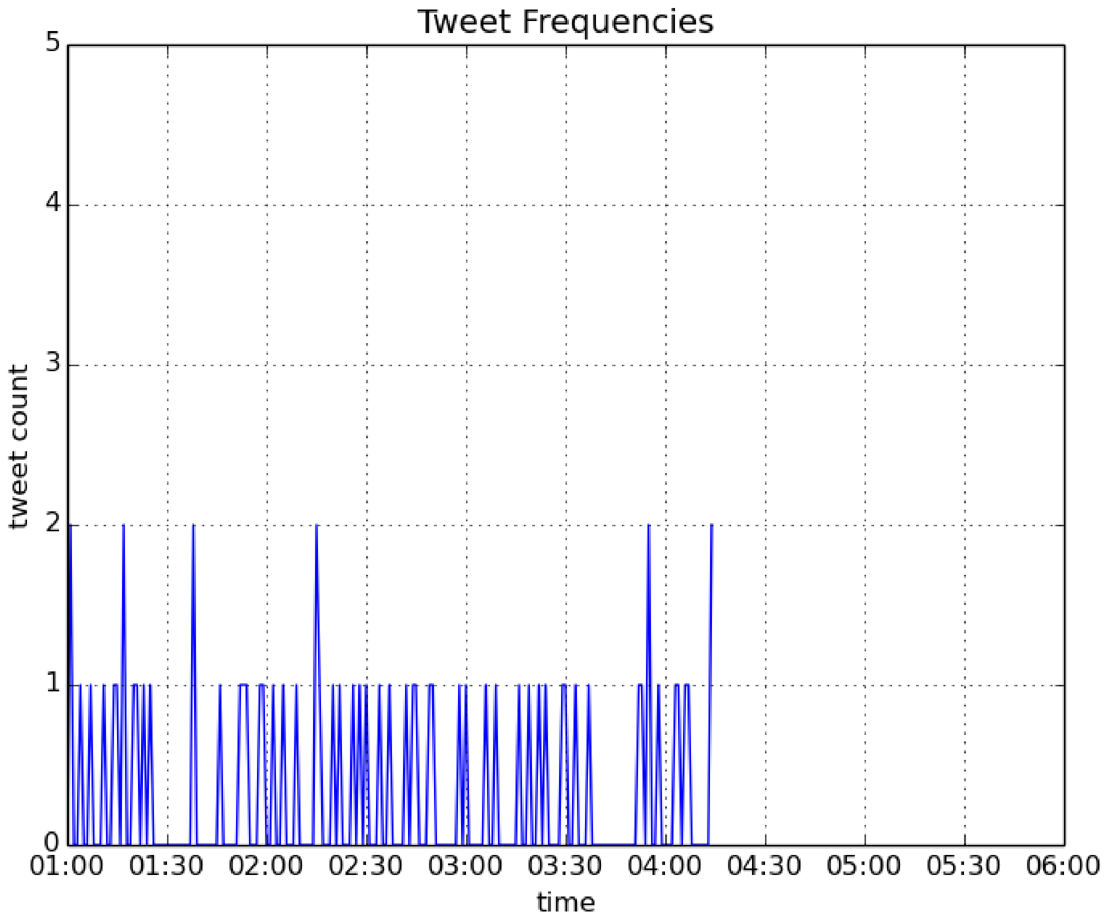
Details:

Module Name	**Responsibility**
Real-time twitter data collection	Input: Twitter access tokens and user handle Output: Tweets from timeline stored in a jsonl format
Separating hashtags from tweets	Input: Extracted tweets Output: Hashtags present in each tweet
Finding frequency of hashtags	Input: Extracted hashtags Output: Frequency of each hashtag
Computing trending hashtags	Input: Frequencies of hashtags Output: Sorted in desc order
Statistics of tweets with and without hashtags	Input: Extracted tweets Output: Composition of tweets with and without hashtags
Computation of incidence matrix	Input: Extracted tweets Output: Incidence matrix
Time series analysis of tweets	Input: Extracted tweets Output: Graph

Process flow:

With the help of Twitter developer credentials, access to the Twitter API is obtained. Then, a home timeline or particular user's timeline tweets are obtained in the jsonl format or tweets embedded with specific hashtags can be obtained and stored in jsonl format. Tweets are then pre-processed for case sensitivity and frequency of hashtags are obtained. Statistics about the tweets with and without hashtags is obtained and incidence matrix is computed. Finally, the time series analysis is done and the output is plotted in the form of graph.

Sample Output:


Tweet Frequencies
tweet count
01:00
01:30
02:00
02:30
03:00
03:30
04:00
04:30
05:00
05:30
06:00
time