Personalized Highlights Recommendation System

Muhammad Talha Khan {Roll #18100267 }
Ahmed Muqaddas {Roll #16030039 }
Syed Mohsin Bukhari {Roll #16030038 }

March 25, 2018

1 General Information about data

When there is a cricket or football match going on, there are tweets related to certain hashtags. This data is generated on user's Twitter profile. This kind of data can be streamed live when the matches are going on. Twitter's Streaming API is for this purpose.

To the best of our knowledge there is no work done regarding recommending personalized highlights for an event or stream. However, there has been work to detect highlights.
Lanagan, James and Smeaton, Alan F. Using Twitter to Detect and Tag Important Events in Live Sports. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011

Collected tweets are mostly text. Further processing is required to extract insights from this text data. Sentiment for each tweet has to be extracted and then the data has to be restructured in form of a matrix. This will help fill out missing values and predict sentiment of new streams for users.

A single data object corresponds to a tweet which is a sentiment score of a user at a particular time. This object will be processed and it will correspond to an element in a matrix where rows are users and columns are time slot of highlights.

The data is in form of streams. Data is captured on the go. That's nature of the problem. However, the dataset needed for sentiment analysis is available and downloaded. The size of one data stream is typically 2MB to 10MB. The overall size of the data is not known and irrelevant because the analysis that is performed needs a stream of data and does not require to data other than what the algorithm produces itself.

Each stream of data is four dimensional. One dimension corresponds to users, the other to different streams, the third to time and the fourth to sentiment embedded in the tweet. However, for preprocessing each data object in order to perform sentiment analysis, the dimension of data is the number of words in vocabulary. This is typically between ten to thirty thousand. For now, we will stick to the lower bound of the range, i.e. ten thousand. Other size related properties include the length of each tweet, which Twitter allows to be 240 characters. This means that a typical Tweet has between 25 to 50 words.

2 Data Format, Types and Description

- The file type is *.txt* for both streamed data and unstructured text data.
- For streams the algorithm stores following attributes.
  - user_id (to identify user)
  - tweet_id (to identify tweet and fetch more data if needed in future)
  - tweet_content (the content of the tweet to perform analysis on)
  - date_time (time at which tweet was Tweet was generated)
- Detailed description of data
  There are typically ten to twenty thousand tweets for a PSL cricket match. The data for Personalized Highlights Recommendation System needs to be four dimensional. However, each data object from the stream needs to be preprocessed in order to represent it in the desired four dimensions (users, streams, time, sentiment).
  - Data types and description of each feature:

Attribute	**Data Type**	**Description**
user_id	Numerical	Unique identifier for a user
tweet_id	Numerical	Unique identifier for a tweet
tweet_content	Text	The actual content that is used to determine embedded sentiment
date_time	DateTime	The time at which the tweet was generated

  - Data types and description of each dimension:

Dimension	**Data Type**	**Description**
user_id	Numerical	Unique identifier for a user
stream_id	Numerical	Unique identifier for a stream
sentiment	Numerical	A number representing sentiment of a user at this particular time
date_time	Numerical	The time at which the sentiment was generated

- Each data object is a tweet, therefore, for sentiment analysis, tweet specific grammar and vocabulary must be used. This grammar and vocabulary is very different from that of actual language.

3 Data preprocessing

The tweets have to be processed in order to extract sentiment from them. Each tweet has to go through a pipeline to be available for sentiment extraction.

1. Stopword Removal
2. Lemmatization
3. Drop very common and rare words

After these steps, each tweet can be expressed as a vector of TF-IDF values where each element corresponds to one word. An alternative to this can be the usage of word embeddings like *Word2Vec* or *GLoVe*.

There will be a lot of missing data once the sentiment has been extracted. For this, a search will be required for similar users and similar streams in order to fill in the missing values.

There might be some noise in the data. In order to deal with that, there can be a range of techniques that can be used to detect this noise. For example, histogram of words can be matched or Hamming distance can be computed to find outliers.

4 EDA

There are no numerical attributes in this dataset for which measures of centrality and spread need to calculated at this stage of the analysis. However, histograms for number of tweets can be plotted along time axis. It is clearly visible that there are some humps in the histogram. These events correspond to the real time highlights when something important happened in the match.

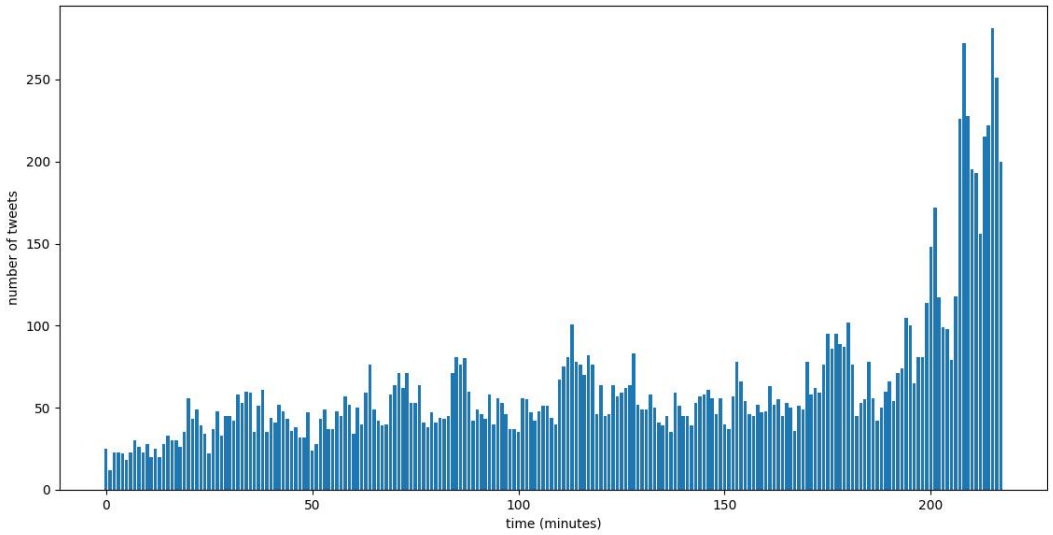

Figure 1: Match 1 (16th March, 2018)

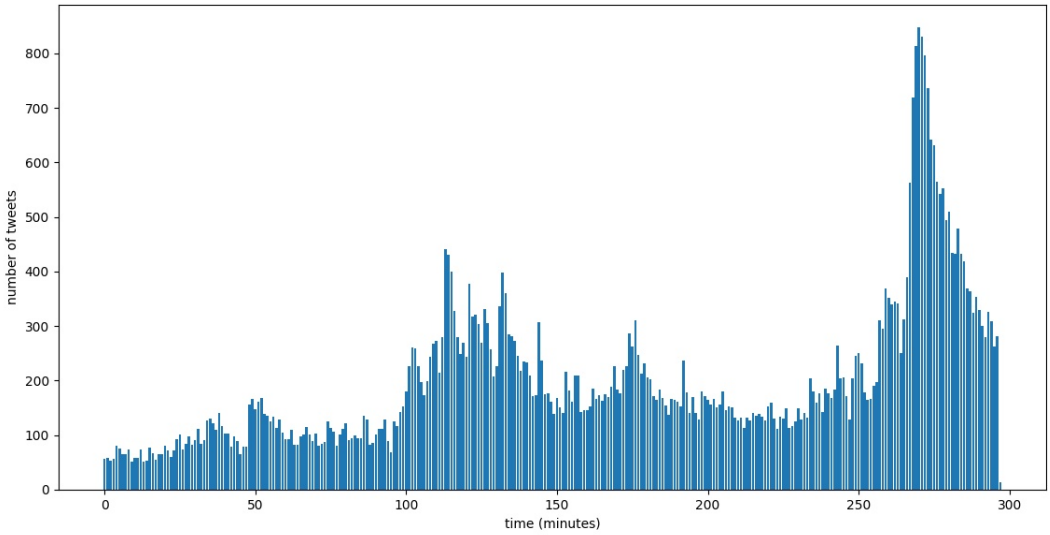

Figure 2: Match 2 (21st March, 2018)