CS-5312 Big Data Analytics

Personalized Highlights Recommendation System

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

1 Problem Description

- Users on the Internet are not interested in watching whole videos (e.g. Cricket or Soccer matches), rather they just want to watch the parts they are interested in (e.g. Goals or Wickets). In our project we will recommend personalized highlights to a user based on his past profile (Interests/Sentiments). The input for the end user in this case will be text data (his/her tweets) and the output will be a set of segments of videos (e.g. Cricket matches sixes or Soccer match penalty saves). There has been work done regarding this in the past but it has not been used in the context of recommendation systems.

2 Motivation

- The motivation behind the work is that tagging segments of a video and recommending them is quite convenient for a user. Consider the case where there is a video lecture of duration 4 hours about vectors but you just want to find how the dot product is calculated. One way would be to watch the whole video but an elegant way would be to tag the segments of the video. The potential audience of this work could be sports websites such as (cricinfo or UEFA Champions league, twitch) and other websites too on which there is a temporal relation between comments and the video content.

3 (Potential) Tools and Techniques

- We will be using python and R studio mainly for our project. Python has rich libraries for for language processing which include implementations of *Word2vec, GloVe and Tweet2Vec.* Our main techniques can be divided into 3 parts (1) Finding the important or significant events (Defined later in idea & approach section) in a video (e.g. Wicket, Goal and etc). (2) Assigning proper tags to the events. (3) Recommend the events to an interested user (e.g. Highlights of all sixes). There has been work done in the past (more details in Related Work) on parts 1 and 2 of our technique but it has not been used for recommendations. Details of our technique are in the idea & approach section.

4 Related/Background Work

- There is a wide range of literature available for recommendation systems. The most common being on collaborative filtering chapter 9 of the book (Mining Massive Datasets) is very useful in our project. **Using twitter to Detect & Tag Important Events in Live Sports** presents a starting point in event detection in a sports match using tweets. A particular useful resource regarding Natural Language Processing is the play-list of video lectures **Natural Language Processing with Machine Learning (Stanford)**. The existing work first of all has to be manually tested which means it cannot be tested on large datasets and furthermore our work uses tagging in combination with recommendations to provide a better user experience.

5 Your idea and approach

- To find the important events in the video we will use statistical analysis of the number of tweets in a defined amount of time (e.g. 60 seconds). If the number of tweets increases drastically in a certain time period then a significant/important event has happened. The more the number of tweets per time period deviates from the mean the more important the event. After identifying the important events we will tag them using the most frequent words used in that specific time period (e.g. six, bowled and out). Our final part would be to find the sentiment of the tweets (positive/negative) of a user and then recommend the video segments in which he would be interested using collaborative filtering. The limitation of this work could be that if there is a randomized delay between the comments and the video content than it would be difficult to generate tag for the video segments.

6 Data Sets

- We are crawling twitter for the PSL matches tweets. We run a crawler when a match starts and gather the timestamp, user id, tweet text, tweet id. So far we have gathered the data of one match and we have collected $13,700$ tweets (entries) for it. The size on disk for this match is $3mb$. The data will for sure have some noise e.g. people tend to tweet in their native language e.g. Urdu so we will have to filter out those tweets. Also we will need to filter out the stop words.

7 Evaluation Criteria/Metrics

- Our evaluation will be performed differently than the past work, in which the important events had to be manually tagged first by humans and then the similarity with the predicted important events was calculated. For our work we will use the commentary from live websites (e.g. ESPN & TenSport) to tag the important events & this will be used as the ground truth e.g. if in a certain time period there is an important event then the bit vector would contain a $'1'$ then we will calculate our bit vector and compare the hamming similarity of these two. To see if out framework recommends we will have to check that manually.

8 Potential Conference/Publication

- The conference that best suits our work is *ICDM*. The deadline for the submission is 05/06/2018. *ICDM* is well suited for our work because there have been multiple publications in it regarding recommendation systems.