

Twitter is one of the most popular social networking websites in the world. As of the third quarter of 2015, it has around 307 million active users. Every second, on average, around 6,000 tweets are tweeted, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year. The large amount of data, so available, can be and has been used for providing a number of services, like

* Speech analysis-Using twitter to generate a word cloud and from that determine a person’s style of speech. It can also be used to find the topics of interest, etc.

• Successes of a product or campaign-To analyze the popularity and the reviews of a product

• Detection of earthquake

• To predict the social unrest and mass protest

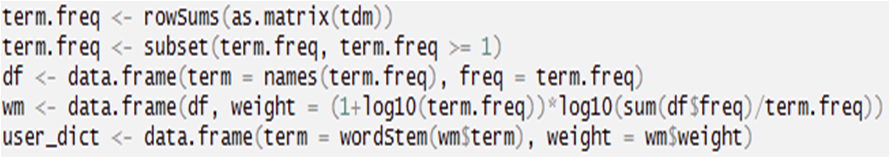
The project was based on understanding the semantics of the tweets made by the user and then recommending movies to the user on the basis of words used.

**Methodology**

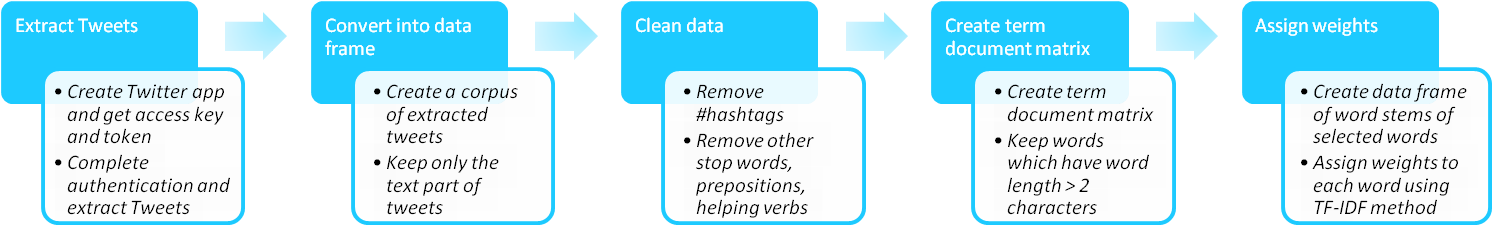
The project was divided into four parts starting from extracting Twitter data and building genre dictionaries; these two steps ran in tandem. This was followed by building movie list from which to recommend movies, and the final step being developing the mapping algorithm.

* **Extracting Twitter Data**

Tweets from a particular Twitter handle are captures using function – userTimeline. Most of the times, tweets captured may not be proper English words, hence the text is cleaned and common words such as helping verbs, prepositions, etc. are removed from the text.



Then, we created a term document matrix of all the words and words which appear at least thrice were kept. Then, these words were reduced to their stems and common stem were added and assigned weights.



For assigning weights, TF-IDF method was used. Term Frequency (TF) captures the raw frequency (count of word in a document) of the word, while Inverse Document Frequency (IDF) captures the relative importance of every word in a document. Final weight of a word would be product of TF score and IDF score.

* **Building genre dictionaries**

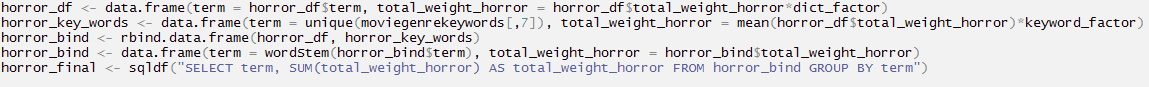
Movie dictionary refers to a list of words that could represent a particular genre of movies. In order to build a movie genre dictionary, we collected data from 4 different sources. The first two included movie description and movie plots from IMDB and Wikipedia. Third source included key/tag words for every movie from themoviedatabase.com(tmdb.com). The final step was manual intervention. A list of words was added to every genre which could represent that genre.



Then, weight to every word was assigned in the same manner as we did for Twitter data.







* **Building Movie List**

A list of around 1800 movies was developed. Every movie could fall into 2 to 4 genres out of the 11 major genres selected. A sample list of movies with all the genres is provided below. Binary representation was used to indicate the presence of a movie in a particular genre, either a movie falls into a genre or it doesn’t.



* **Finding User’s preference of genre**

Every word in the user dictionary was matched with every word in all the 11 genre dictionaries. For every word that matched, corresponding scores were multiplied. This gave us score for every genre.



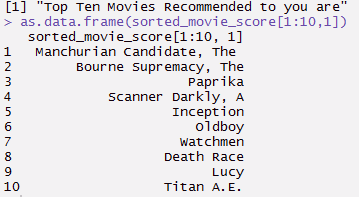
* **Mapping Algorithm**

1. Take dot product of movie matrix (1814 x 11) with the score vector (11 x 1)
2. Get movie score vector (1814 x 1) having score for every movie
3. Recommend movie at the top of the list



**Outcome and Next Steps**

The project's output could be used to identify user's personality in terms of his liking/disliking.

Since the choice of movies for a particular user could be associated with many other things such as books, places he would like to visit, among many others, this could be further build upon.

It can serve a perfect tool for online marketing based on user's social media profile. Some of the improvements that could be done are as follows:

* Improving Semantic and associations of words in User profile
  + There exist different words which imply same meaning. Moreover, even while creating term document matrix, different forms of a same words, or even singular plural are considered different words. We tried to minimize this by using the stem of the word, instead of using the word itself. This improved the accuracy to some extent. The next step to this would be to find the intensity of a word and group same meaning words together and then determine user's personality.
* Increasing number of genres
  + Considering the scope of the project and paucity of time, we considered only 11 genres which could broadly classify most of the movies. Netflix uses around 77,000 micro-genres to classify movies. More the number of genres, better the classification and recommendation will become.
* Find association of Movies
  + So far we tried finding association among various aspects of a user's profile. The same can be done for movies. Each movie could be tagged with few key words and other characteristics such as time of release, cast, actor-director pair, geography, etc. This can then be used to cluster movies and improve the recommendation process.
* Including other aspects of Twitter Profile
  + In the project, as it is, we have considered only the Tweets made by the user. This could be complemented by other aspects such as hashtags, handles that a person is following, retweets, etc. This would enable us in gathering more information about the user.

**Challenges**

* Extracting legible data from Twitter
  + As ironical as it may sound, social media are a store house of data, but only a small fraction of that data can be used to extract useful information.
  + Moreover, the usage of slangs, incorrect and misspelled English words make it difficult to use the extracted data
* Building genre dictionary
  + There are many movies which fall into different categories, and hence a word that represents a particular genre could also fall into different genres. Estimating the weights of such words so as to distinguish their presence in different genres was a challenge. This was taken care of manually adding set of words to every genre.
* Lack of training and test data
  + This problem, being exclusive to the project, was the most difficult to handle. Since, there is no direct literature available related to this project. It was a difficult task to find out whether our algorithm is working fine. We took a sample of few and asked their preferences for movies. Then, we found out movies based on their Twitter handle and compared their preferences. We found out significant overlap between the two lists.

**R Packages Used**

* twitteR
* SnowballC
* sqldf
* tm

**R Functions Used**

* userTimeline
* twListToDF
* regmatches
* TermDocumentMatrix
* tm\_map
* Corpus
* wordStem
* sqldf
* setup\_twitter\_oauth

**Thank You ☺**

**PGDBA Group 08**

Bharathi R ([bharathir2017@email.iimcal.ac.in](mailto:bharathir2017@email.iimcal.ac.in))

Faichali Basumatary ([faichalib2017@email.iimcal.ac.in](mailto:faichalib2017@email.iimcal.ac.in))

Shoorvir Gupta ([shoorvirg2017@email.iimcal.ac.in](mailto:shoorvirg2017@email.iimcal.ac.in))

Vishal Bagla([vishalba2017@email.iimcal.ac.in](mailto:vishalba2017@email.iimcal.ac.in))