Book Review Based on Twitter Data

REJUL JAMES A20345321 & DAYANA M SUNNY A20346386

Problem

- People will see contradictory/different ratings for same books in different websites sometimes.
- Collected data from twitter to predict more accurate ratings of the recently published books.
- Geographically plot the users of selected books using the location data extracted from tweets

Approach

- Logistic regression classifier has been used to predict the ratings of the book. A model has been trained on 1500 tweets and predicted rating for 7 books using around 5000 tweets .

- Using Google maps API, we collected the latitude and longitude of these locations plotted these in google static maps.

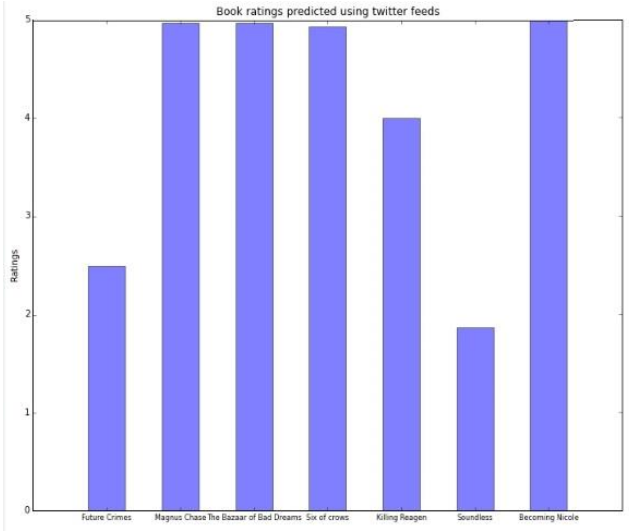
Data

- We collected around 6.7k tweets useful tweets and extracted tweet, User location, name, Retweeted information. Ratings bas been predicted for 7 books
- We have used search api of tweepy.API. Using since_id and max_id, we could overcome the limitations of data collection, we collected maximum tweets (upto 5000) from past two weeks in one single request. Google map API has been used to collect longitudes and latitudes of these location

Name of the Book	Tweets Collected	Fields Extracted twitter API	Fields from Google Maps API
Magnus Chase	2000	Tweet, Name, Retweeted, user location,	Place_id, latitude, Longitude
Six of Crows	2000	Tweet, Name, Retweeted, user location,	Place_id, latitude, Longitude
Future Crimes	500	Tweet, Name, Retweeted, user location,	Place_id, latitude, Longitude
Becoming Nicole	1100	Tweet, Name, Retweeted, user location,	Place_id, latitude, Longitude
Soundless	252	Tweet, Name, Retweeted, user location,	Place_id, latitude, Longitude
Killing Reagen	425	Tweet, Name, Retweeted, user location,	Place_id, latitude, Longitude
The Bazaar of Bad Dreams	860	Tweet, Name, Retweeted, user location,	Place_id, latitude, Longitude

Result

- Naïve Bayes and Logistic regression classifiers were used. Logistic regression model had performed better than Naïve Bayes. Accuracy of the logistic model is around 0.75 on test data. Naïve Bayes was even lower 0.53



Name of the Book	Rating	Category
Magnus Chase	4.97005988	Excellent
Six of Crows	4.93243243	Excellent
Future Crimes	2.5	Average
Becoming Nicole	5	Excellent
Soundless	1.875	Bad
Killing Reagen	4	Excellent
The Bazaar of Bad Dreams	4.96987952	Excellent

- Accuracy has been tested using cross validation k fold method.
- For validation of user location, We took world's 50 famous countries and their cities. Europe and North America contributed towards the tweets the most

Conclusion

- We experimented two classifiers Naïve Bayes and Logistic Regression. Logistic regression performed well compared to Naïve bayes.
- Majority of the reviews we collected are positive, its difficult to get negative reviews/ tweets for books. Large share of one category of data can lead to overfitting of the model
- Google Static Map has limitation that it will allow 50 request per user.