Optimization of the reviewing process and assessing popularity of movies

Sohom Ghosh[1], Santanu Modak[2] and Dr. Abhoy Chand Mondal[2]

[1]Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata – 700107, West Bengal, India
[2]Department of Computer Science, University of Burdwan – 713104, West Bengal, India.
{sohom1ghosh, modaksantanu}@gmail.com, abhoy_mondal@yahoo.co.in

Abstract. Opinion analysis has become a flourishing frontier as of late. In this paper, we exhaustively study movie reviews from a popular online database. We randomly sample more than 1000 reviews with titles to train our model. It is capable of suggesting words during the process of appraising a film. It can intelligently anticipate the words that an appraiser is going to use from the title of his opinion. Furthermore, it has the potential to learn. Whenever it finds that it is unable to suggest words, it learns from the critic's opinion. Moreover, this innovative model is able to compute the popularity of a film by examining the opinions. Thus, it simplifies the job of reviewing by making it quicker and effective. It labels a movie as 'super-flop', 'flop', 'cool', 'hit' or 'super-hit' based on what the reviewers opine.

Keywords: Opinion mining, sentiment analysis, natural language processing, polarity computation, recommendation system, machine learning, collaborative filtering, product reviews

1 Introduction

In this paper we discuss about our unique model to propose words to a reviewer while appraising a film. We use the online database of movies, IMDb to build our training and test set. Firstly, we request the appraiser to enter the name of the movie and the title of his opinion. We analyze this title and predict words he is most likely to use while reviewing. Then, we give him a turn to opine. We examine this opinion and compute score from it. Furthermore, we give him a chance to rate the film. We repeat this process for every user. Finally, for each movie we store every appraiser's rating. We evaluate the mean from this and declare whether the movie is 'super-flop', 'flop', 'cool', 'hit' or 'super-hit'.

This model is beneficial as it makes the process of appraising faster and simpler. Users do not have to spend time wondering for words while reviewing a film. They don't need to consider about the score they want to assign. They will receive suggestions at each and every step. Moreover, this model has the ability to learn which enhances its efficiency with usage.

2 Related Works

(1 paper already mentioned by sohom, Refer 4 other papers, append them to references)// 4 papers to be cited by santanu da

3 Definitions and Notations

3.1 Tokenization

The process of splitting a sentence into its constituent words is known as Tokenization. We use split() function to tokenize sentences. Ex: Original Sentence: - "The Sky is Blue"; After Tokenization: ['The','sky','is','blue']

3.2 POS tagger

POS tagger (or parts of speech tagger) is an inbuilt package of NLTK which labels each word to the parts of speech they belong. Example: Original Sentence: "The opening aerial shots"; after using POS tagger: - (S The/NNP opening/VBG aerial/JJ shots/NNS)

3.3 Stemming

Words like 'behaving', 'behave', 'behaved' means the same. They have structural affixes. We convert these words to a single form i.e. 'behave'. This act is referred to as stemming. We use Porter stemmer as a preprocessing step.

3.4 Corpus

Corpus means collection of words. We have built unique corpora separate for positive and negative words. It contains around 6000 words. So, let's look at a sample from it.
Positive words: - [':)', 'absolutely', 'bounty', 'calm', 'meritorious', 'skilled', 'wow', 'zeal']
Negative words:- [':(', 'abysmal', 'hurtful', 'ignore', 'malicious', 'worthless', 'yucky']

3.5 Polarity

Polarity refers to the positivity or negativity of a word. In this paper, we label positive words as '+1' and negative words as '-1'.

3.6 Mapping

Mapping is representation of relations. It refers to a function. For example:

f: X-> Y, denotes that f is a function which maps X to Y. In this paper we use one-to-many and many-to-one mapping.

3.7 Collaborative filtering

For building the recommendation system to suggest words, we use collaborative filtering. It refers to the art of proposing words by gathering interests from the users (collaboration).

3.8 Mean

Mean is a statistical term. It is also referred to as the average. Here, we are finding out the arithmetic mean of some discrete values. The formula for mean is:-

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i$$

Here, A denotes the arithmetic mean, a_i denotes the discrete values, Σ represents summation and n refers to the number of discrete values.

4 Problem definition

Given the name of a movie and title of an opinion about it, we predict the words an appraiser will probably use. We anticipate the score from what he opines. We have built our unique word corpora separate for positive and negative words. We compare the adjectives and adverbs of the opinion with our corpora to detect the degree of polarity of the film.

5 Experimental Evaluation

Index: SR= Suggestive Rating, UR= User Rating, MP= Movie Popularity, UN= User No. Note: Some of the user reviews have been shortened.

- Case-1
  Movie name: Inception

UN	Title	Suggestion	User Review	SR	UR	MP
1	Too much....WAY too much	average, ordinary, so-so , mediocre	What is going on with the IMDb user reviews lately? It's like the masses can no longer be trusted. In the last month, the	3	3	cool

			users have decreed "Airbender" the worst abomination ever, when in fact it's just an average movie.			
2	Insanely Brilliant! Nolan has outdone himself!!	really, believable, potential, interesting, good, high, high, most, creative, highly	What is the most resilient parasite? An Idea! Yes, Nolan has created something with his unbelievably, incredibly and god- gifted mind which will blow the minds of the audience away.	5	5	hit

- Case 2
  Movie name: Apartment 1303 3D

UN	Title	Suggestion	User Review	SR	UR	MP
1	Don't waste your money	don't, slower, dull, slowly, boredom, never	OK, my Summary basically wraps it up. Remember Eddie Murphy's joke about white people moving into haunted houses? "Oh, nice House." "Get out." "just a few ghosts we can handle that.' This movie is an embarrassment.	3	1	Superflop
2	Waste of time!!	average, ordinary, so-so , mediocre	Where to start. This movie was crap. The acting is horrible. That girl cannot act. She sounds mannish and so mono toned throughout the movie.	1	1	Superflop

- Case-3
  Movie name: The Godfather

UN	Title	Suggestion	User Review	SR	UR	MP
1	"The Godfather" is pretty much flawless, and one of the greatest films	good, acting, spectacular, especially, must-watch	Rather than concentrating on everything that is great about The Godfather, a much easier way for me to judge its quality is on what is bad about it. Almost every film has something that I don't like about it, but I can	3	4	hit

	ever made		honestly say that I wouldn't change			
2	Magnifi-cent por-trait of organized crime	average, ordinary, so-so , medio-cre	This is by far the best movie ever to give a portrait organized crime; this movie goes deep inside and shows it all inside out.	5	5	hit

Here, we can see the popularity of a movie changes. This is because, when each reviewer is rating a movie, the overall score varies. Thus, the popularity of the film is affected.

6 Analysis

$$A=100-(|Q-R|/Q)*100$$

Movie name	IMDb Rating (out of 10)[=I]	IMDb Rating (out of 5)[=Q=I/2]	Rating by our model(out of 5)[=R]	% Accuracy [=A]
Inception	8.8	4.4	4	90.91
Apartment 1303 3D	2.6	1.3	2	46.15
The Godfather	9.2	4.6	4	86.96

Table 1. Percentage Accuracy of predictive scores

Thus, it is obvious that our model is quite accurate for movies with higher ratings. For poorly rated movies our it seems to be inefficient. But, this is not so. For a lowly rated film the denominator 'Q' is comparable to the difference '|Q-R|'. This lowers the magnitude of % Accuracy, 'A'. If we look at the numeric values of Q and R we can easily conclude that they are similar.

7 Conclusions

This ingenious model is quite handy, fast and fit for use. It saves time. There are few scopes of improvement. We have trained it using about 1000 reviews. It's advisable to train it with more reviews to enhance its accuracy. Here, we check the polarity of a word by checking its presence in positive or negative corpora. This is not so efficient. Presence of a negative word within 3 words from another negative word makes the sense of sentence positive. Instead of suggesting too many words, it is recommended to suggest those with higher frequency of occurrences.

8 References

1. Maas, Andrew L. et al “Learning Word Vectors for Sentiment Analysis” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 2011.

[2-5-> 4 references to be added by Santanu Da]