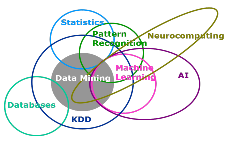
VE488 Data Mining and Machine Learning
Final Report
University of Michigan - Shanghai Jiao Tong University
Joint Institute, Summer 2018



Film Recommendation System Trained with Feature Analysis of Film Reviews

Hu Qiaoyu 5143709060
Xu Dewei 5143709248
Zhang Tianjun 5143709070

Abstract

In this work, we present a movie recommendation system that utilize all the film reviews of a certain film to suggest films of this kind. There have been various studies of recommendation systems. However, most of the systems is based on crowded user preference or user history. For example, these systems identify users that share same film preference or user history preference and then suggest film watched by similar users. We notice that beyond user profile that contain user history preference, user written movie reviews carry substantiation amounts of film related features such as description of genre, characters and plots. Applying natural language processing, it is possible to extract useful features from the review.We then use word2vec that transform a paragraph into a vector which is capable to maneuver with.We make predictions based on the similarity of these feature vectors.Experimental results show that our method can be used to find similar recommendations efficiently.

1 Introduction

Recommender systems are important building blocks in many of today's e-commerce applications including targeted advertising, personalized marketing and information retrieval.[2]

With the development of Internet streaming, thousands of movies become available in a click of button.People now can enjoy movies not only from Hollywood, but also from International cinemas, documentaries, media, etc.[1] With so many films in hand, the consumers faces the choice of which film to choose. At the end of one day, people would like to watch a movie that satisfy his taste and style and this is how the recommendation system could help, suggesting movies according to user's taste and preference. In order to recommend movies, the system need to understand movie first. The better the system can understand movie feature, the better recommendation it can provide.

The online movie databases such as IMDB or MovieLens already have some labels or tags for a certain movie. A common label for a movie could be "romantic","ferocious","heart-breaking"and etc. These labels are often inputs from the user who watched the film. Some recommendation systems that make predictions according to these labels and the result turn out to be quite well. However, such approach is definitely not scalable when the amount of tags become huge.

For the recommendation system to understand film, it needs more detailed information about a film. For example, feature of main characters, the plot, twist and turns of a story mainly determine how a movie develop but that piece of information cannot be informed from tags or titles. Therefore, user written movie reviews is one important source of features. It carries substantial amount of movie related information such as memorable scene description which help the system better understand the inherent contents of the film.

Considering that the users written review is composed of both useful words and useless words(i.e. stopwords), some text-preprocessing work is required before we start to maneuver with the reviews. In our work, we choose to apply the method of Natural Language Processing(NLP) to first eliminate those pause words.

With the handful data after preprocessing, we need to convert the reviews into data that carry the feature of reviews. We select the method of Word2Vec, the detail of the algorithm will be discussed in the following section and the method will ultimately yields a mapping betweenwordsand a fixed length vector that can be used directly as input to other algorithms that compare similarity.

Experiments are performed over a medium size of movie datasets to show that our method is efficient.

2 Review of Prior Work

In this section, we will show here the previous studies on mining user aspects, some studies by using movie reviews for recommendation and showing how are they different from our work.

One common way of recommendation system is built by collaborative filtering [3] which is originated from the idea that if you like an item, then you will also like a "similar" item. Based on the user's history customs, a user profile is generated, which is then used to make suggestions to the user. The user models are presented to the users in terms of the most important features and dimensions in their profile.As the user provides more inputs or takes actions on the recommendations, the engine becomes more and more accurate. However, this method is poor to handle cold start problems. cold start problems of prime concern as it makes the system complex by not containing any prior rating history and involves three cases: recommendation for new user, recommendation for new product and recommendation of new product for new user. For example, if Bob and Wendy liked the same movies as you in the past and they both rated Star Wars highly, you might like it, too. However, recommender systems that employ purely collabo- rative filtering can't recommend an item until sev- eral users have rated it.

There have been some studies about taking into account reviews for recommendation as we did in our approach. It focus on designing systems that produce personalized recommendations in accordance with the available contextual information of users[**Feng**, 4]. Compared to the traditional systems that mainly utilize users' preference history, context-aware recommender systems provide more relevant results to users. As a demonstration, in collaborative filter recommendation, the system is able to figure out the preference trend of crowded users, then use that trend to predict current user preference. However, the method that we use to mining contextual data from textual reviews are different. Overall, there are three significant differences in our work from others including: (1) in our work, we apply word2vec algorithm to vectorize the review data 2) We optimize the method in prediction therefore our approach is less prone to cold start problem within well supervised data and (3) we empirically apply our approach to the first large collected user reviews for recommendation.

3 Methodology & Framework

Our proposed approach is applying movie recommender system based on the review of the users with Natural Language Processing. This is a relatively new approach since prior approaches either make movie recommendations based on the movies searched by the users or simply clustering reviews into positive or negative. In this section we will introduce two main framework of our system:

- **Film Prediction System**: we would like to train a model based on a collection of reviews labeled with the film name, and then we will predict the film a new review is talking about based on the trained model.

- **Film Recommendation System**: we will build a film recommendation system which could recommend some new films based on the user input.

In the following part we will introduce details about the two systems.

3.1 Film Prediction System

First, we visit the film website with web request and use regular expression to filter the high-ranking films. Then we visit the websites of these films and download the reviews one by one. The film set we chose is marked as F. The review set for the ith film in F is marked as r_i. The text data is stored into a csv file.

Then we divide the reviews into words and use a library called Nltk to remove these useless or irrelevant words. This is completed by the 'stop-word' trained in the library.

Then we take two methods to deal with the reviews. The first one is to split the reviews into words and train a word-to-vector (call word2vec for short) to transform the words into vectors. Then we average and concatenate the vectors to calculate the representation for the film. The formula is listed as below:

$$V_{review} = \frac{1}{n}\sum_{i=1}^{n} V_{word_i} = \frac{1}{n}\sum_{i=1}^{n} G(word)$$

where G is the word2vec model.

Another idea is to apply the document-to-vector transformation, which can convert some text into a single vector. Therefore, V_{review} can be calculated directly. Here we test the accuracy for both of them and the detailed result will be introduced in the next section.

In terms of word2vec model, the model we use is from Google. Figure 1 shows the detailed model It provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. First, in continuous bag-of-words model (CBOW), there is a projection layer that converts a one-hot representation of 4 words previous to and 4 words after a certain center word to a lower dimension matrix. Averaging them up, one can get a prediction of the center word. Parallel, we adopt the continuous skip-gram model in which a projection layer project the input of a center words to the prediction of 4 words before it and 4 words after it. With the help of these 2 models, we can successfully converts the words to a certain vector representation.

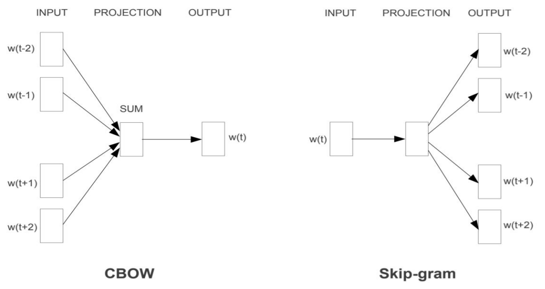
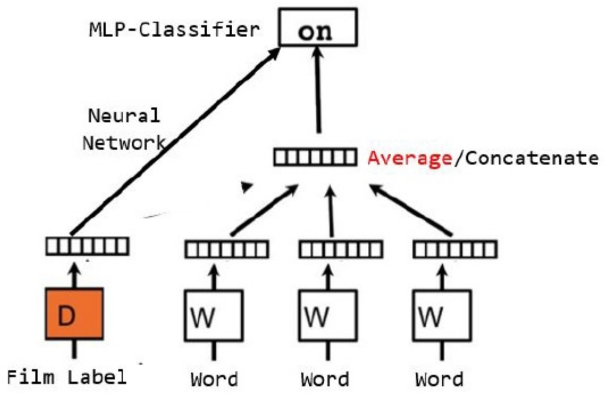


Figure 1: word2vec model architecture

With the vector input, we are able to complete the last step — a multi-layer perceptron (**MLPClassifier**). Here we train the network with flim review vector labeled with the film index. And we test different optimizer like adam or sgd, with different iterations and layers. The training data we choose is some fresh review for the ith film in F but not included in r_i. Therefore, it can prove the efficiency and robustness of our system to detect features in different reviews. And the accuracy is up to 78% in our trained model.



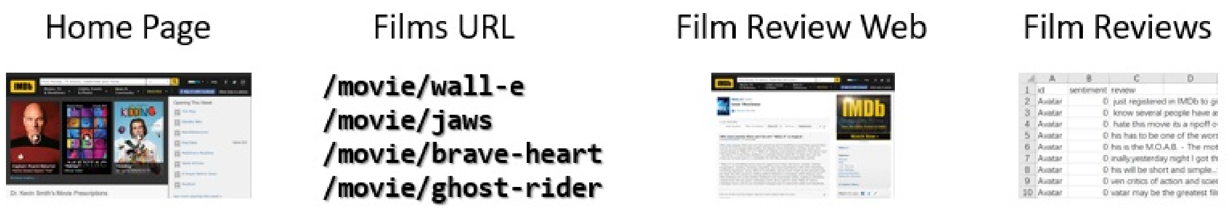


Figure 1: Review Process diagram

Figure 2: The structure of our system

3.2 Film Recommendation System

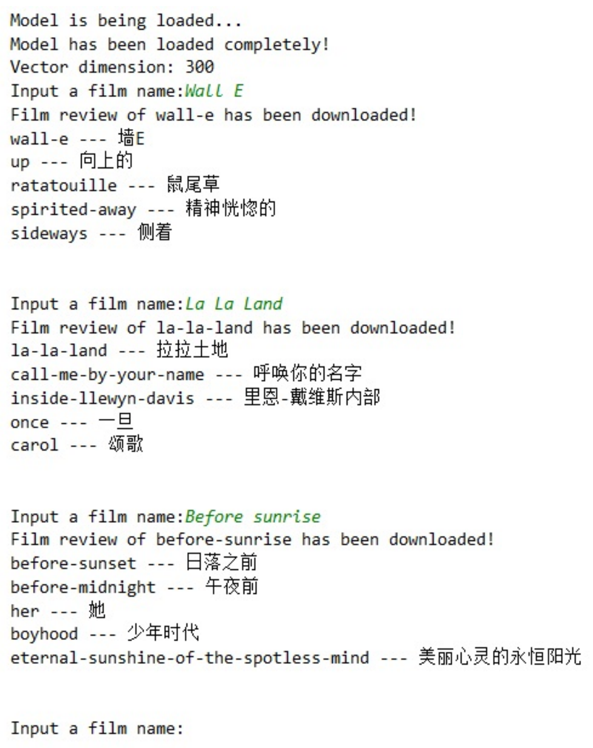
This is the extended application of our system. As we have evaluated before, the review vector can be calculated by averaging all the vector representations of the words. For the same principle, we can represent the film using the average of its review data.

$$v_{film} = \frac{1}{n}\sum_{i=1}^{n} r_i$$

When we input a new film, we find the reviews and combine them into a vector. By comparing the vector distance in high-dimensional space, we can find the closest films to our target film. The vector distance formula is shown below:

$$D_{v_1,v_2} = \frac{v_1 \cdot v_2}{||v_1|| * ||v_2|| + \epsilon}$$

As you can find in the screen-shot, our system performs a good recommendation. For the animated film Wall-E, all the recommendations are animated film with high quality. And for the film Before-Sunrise, the system recommends the trilogy completed by one director. And Before-Sunrise is exactly one of them. The result proves that word2vec model and doc2vec model is a practical solution to the recommendation problem.



```
Model is being loaded...
Model has been loaded completely!
Vector dimension: 300
Input a film name:Wall E
Film review of wall-e has been downloaded!
wall-e --- 墙E
up --- 向上的
ratatouille --- 鼠尾草
spirited-away --- 精神恍惚的
sideways --- 侧着


Input a film name:La La Land
Film review of la-la-land has been downloaded!
la-la-land --- 拉拉土地
call-me-by-your-name --- 呼唤你的名字
inside-llewyn-davis --- 里恩-戴维斯内部
once --- 一旦
carol --- 颂歌


Input a film name:Before sunrise
Film review of before-sunrise has been downloaded!
before-sunset --- 日落之前
before-midnight --- 午夜前
her --- 她
boyhood --- 少年时代
eternal-sunshine-of-the-spotless-mind --- 美丽心灵的永恒阳光


Input a film name:
```

Figure 3: Our recommendation result

3.3 Advantages of our system setup

Comparing to prior approach, this approach mainly improve three aspects. First, previous approaches focus on predicting users' preference of movies based on the search of movies. This would be pretty inaccurate since after watching a certain movie, the user might not like it. Thus, simply predicting preference based on searching results cannot reflect the habit. However, the movie review can accurately reflect those aspects in which a user might give bad remarks on some movies they like. Applying this to movie recommendation will greatly model the taste of users. Second, this algorithm doesn't have a cold start problem. Prior approaches try to model the users' preference based on previous actions. So if a new user enters the system, the algorithm will have a cold start. This algorithm doesn't have such problem since it predicts the preference using other user's data. Thirdly, applying Natural Language Processing technique can greatly improve the accuracy. Modeling the similarity between each word using a word2vec rather than doing simple feature extraction is the key to this approach.

In all, the model we used apply a word2vec Natural Language Processing layer before naive bayes model. Applying this on a totally new dataset, we would think that this can greatly improve the performance of movie recommendation system.

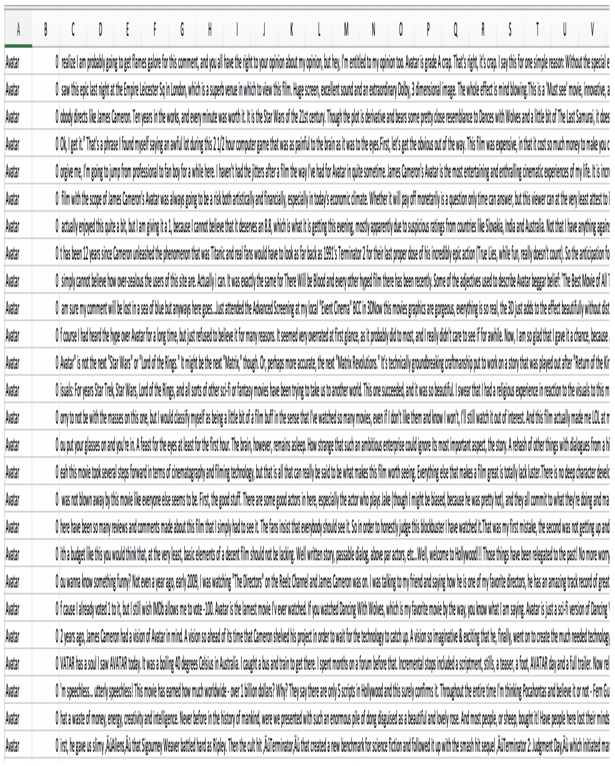
4 Results and Discussion

In this section, we are going to provide the experiment setup and the corresponding test result.

4.1 Experiment Setup

For the testing purpose, since there are currently available dataset for recommending similar movies based on the movie reviews, we start downloading the dataset ourselves. A dataset containing popular movies from IMDB is used for the generation of the training and testing dataset. The dataset called ”$movie_metadata.csv$” contains the movie names, directors, actors, website urls, etc.

We write a python program downloading the first 10 movies who have more than 2000 user reviews. The corresponding movies are: Avatar, The Dark Knight Rises, Batman v Superman: Dawn of Justice, Superman Returns, Man of Steel, King Kong, Titanic, Indiana Jones and the Kingdom of the Crystal Skull, The Dark Knight, Interstellar. Figure 2 shows a sample of the collection of data. First column is the movie name, second is the corresponding integer representation and third is the raw movie reviews.



A	B	C
Avatar	0	realize I am probably going to get flames galore for this comment, and you all have the right to your opinion about my opinion, but hey, I'm entitled to my opinion too. Avatar is grade A crap. That's right, it's crap. I say this for one simple reason: Without the special e
Avatar	0	saw this epic last night at the Empire Leicester Sq in London, which is a superb venue in which to view this film. Huge screen, excellent sound and an extraordinary Dolby, 3 dimensional image. The whole effect is mind blowing.This is a 'Must see' movie, innovative, a
Avatar	0	obody directs like James Cameron. Ten years in the works, and every minute was worth it. It is the Star Wars of the 21st century. Though the plot is derivative and bears some pretty close resemblance to Dances with Wolves and a little bit of The Last Samurai, it does
Avatar	0	Ok, I get it." That's a phrase I found myself saying an awful lot during this 2 1/2 hour computer game that was as painful to the brain as it was to the eyes.First, let's get the obvious out of the way. This film was expensive, in that it cost so much money to make you c
Avatar	0	orgive me, I'm going to jump from professional to fan boy for a while here. I haven't had the jitters after a film the way I've had for Avatar in quite sometime. James Cameron's Avatar is the most entertaining and enthralling cinematic experiences of my life. It is incr
Avatar	0	film with the scope of James Cameron's Avatar was always going to be a risk both artistically and financially, especially in today's economic climate. Whether it will pay off monetarily is a question only time can answer, but this viewer can at the very least attest to i
Avatar	0	actually enjoyed this quite a bit, but I am giving it a 1, because I cannot believe that it deserves an 8.8, which is what it is getting this evening, mostly apparently due to suspicious ratings from countries like Slovakia, India and Australia. Not that I have anything agains
Avatar	0	t has been 12 years since Cameron unleashed the phenomenon that was Titanic and real fans would have to look as far back as 1991's Terminator 2 for their last proper dose of his incredibly epic action (True Lies, while fun, really doesn't count). So the anticipation fo
Avatar	0	simply cannot believe how over-zealous the users of this site are. Actually i can. It was exactly the same for There Will be Blood and every other hyped film there has been recently. Some of the adjectives used to describe Avatar beggar belief: 'The Best Movie of All T
Avatar	0	am sure my comment will be lost in a sea of blue but anyways here goes...Just attended the Advanced Screening at my local "Event Cinema" BCC in 3DNow this movies graphics are gorgeous, everything is so real, the 3D just adds to the effect beautifully without dist
Avatar	0	f course I had heard the hype over Avatar for a long time, but just refused to believe it for many reasons. It seemed very overrated at first glance, as it probably did to most, and I really didn't care to see if for awhile. Now, I am so glad that I gave it a chance, because .
Avatar	0	Avatar" is not the next "Star Wars" or "Lord of the Rings." It might be the next "Matrix," though. Or, perhaps more accurate, the next "Matrix Revolutions." It's technically groundbreaking craftmanship put to work on a story that was played out after "Return of the Kir
Avatar	0	isuals: For years Star Trek, Star Wars, Lord of the Rings, and all sorts of other sci-fi or fantasy movies have been trying to take us to another world. This one succeeded, and it was so beautiful. I swear that I had a religious experience in reaction to the visuals to this m
Avatar	0	orry to not be with the masses on this one, but I would classify myself as being a little bit of a film buff in the sense that I've watched so many movies, even if I don't like them and know I won't, I'll still watch it out of interest. And this film actually made me LOL at m
Avatar	0	ou put your glasses on and you're in. A feast for the eyes at least for the first hour. The brain, however, remains asleep. How strange that such an ambitious enterprise could ignore its most important aspect, the story. A rehash of other things with dialogues from a hi
Avatar	0	eah this movie took several steps forward in terms of cinematography and filming technology, but that is all that can really be said to be what makes this film worth seeing. Everything else that makes a film great is totally lack luster.There is no deep character develo
Avatar	0	was not blown away by this movie like everyone else seems to be. First, the good stuff. There are some good actors in here, especially the actor who plays Jake (though I might be biased, because he was pretty hot), and they all commit to what they're doing and ma
Avatar	0	here have been so many reviews and comments made about this film that I simply had to see it. The fans insist that everybody should see it. So in order to honestly judge this blockbuster I have watched it.That was my first mistake, the second was not getting up and
Avatar	0	ith a budget like this you would think that, at the very least, basic elements of a decent film should not be lacking. Well written story, passable dialog, above par actors, etc...Well, welcome to Hollywood!!! Those things have been relegated to the past! No more wor
Avatar	0	ou wanna know something funny? Not even a year ago, early 2009, I was watching "The Directors" on the Reelz Channel and James Cameron was on. I was talking to my friend and saying how he is one of my favorite directors, he has an amazing track record of great
Avatar	0	f cause I already voted 1 to it, but I still wish IMDb allows me to vote -100. Avatar is the lamest movie I'v ever watched. If you watched Dancing With Wolves, which is my favorite movie by the way, you know what I am saying. Avatar is just a sci-fi version of Dancing
Avatar	0	2 years ago, James Cameron had a vision of Avatar in mind. A vision so ahead of its time that Cameron shelved his project in order to wait for the technology to catch up. A vision so imaginative & exciting that he, finally, went on to create the much needed technology
Avatar	0	VATAR has a soul I saw AVATAR today. It was a boiling 40 degrees Celsius in Australia. I caught a bus and train to get there. I spent months on a forum before that. Incremental stops included a scriptment, stills, a teaser, a foot, AVATAR day and a full trailer. Now rel
Avatar	0	'm speechless... utterly speechless! This movie has earned how much worldwide - over 1 billion dollars? Why? They say there are only 5 scripts in Hollywood and this surely confirms it. Throughout the entire time I'm thinking Pocahontas and believe it or not - Fern Gu
Avatar	0	hat a waste of money, energy, creativity and intelligence. Never before in the history of mankind, were we presented with such an enormous pile of dong disguised as a beautiful and lovely rose. And most people, or sheep, bought it! Have people here lost their minds
Avatar	0	irst, he gave us slimy ‚ÄúAliens‚Äù that Sigourney Weaver battled hard as Ripley. Then the cult hit ‚ÄúTerminator‚Äù that created a new benchmark for science fiction and followed it up with the smash hit sequel ‚ÄúTerminator 2: Judgment Day‚Äù which initiated mar

Figure 4: Sample Dataset Overview

4.2 Text Processing

For text processing, we use NLTK library. First, we iterate over the full text and tokenize each file, allowing us to analyze the file at text level. We will remove the common punctuation and stopwords - grammatical words which are usually ignored as they do not provide any useful information, i.e, $other, there, of, the, are$. Using the NLTK's default library, we are able to identify the stopwords that are commonly used in the review.

4.3 Word2vec/Doc2vec and MLP Model Training

For the purpose of converting words and paragraphs to vectors, we explored 2 commonly used techniques. For Word2vec, we used the pre-trained

Table 1: Test Result

Optimizer	Layer	Iteration	Random	Alpha	Word2vec Accuracy	Doc2vec Accuracy
adam	150,100,60,30,10	100000	0	0.00001	0.51	0.658
adam	100,50,10	100000	0	0.00001	0.49	0.625
adam	80,20	100000	0	0.00001	0.47	0.662
adam	60,20,6	100000	0	0.00001	0.42	0.614
adam	100,	100000	0	0.00001	0.47	0.652
sgd	100,50,10	100000	0	0.00001	0.53	0.607
sgd	80,20	100000	0	0.00001	0.46	0.662
sgd	150,100,60,30,10	100000	0	0.00001	0.5	0.617
sgd	60,20,6	100000	0	0.00001	0.17	0.599
sgd	100,	100000	0	0.00001	0.6	0.717

model Google trained on Wiki dataset. After that, for a paragraph, we took the average of all the unique words to generate a 300 dimension vector. For Doc2vec, we also used the pre-trained model trained on English Wiki dataset. Also, a 300 dimension vector representing the whole paragraph is automatically generated.

The last training step, we applied the MLP algorithm on the generated vectors and fit that to a corresponding movie label.

4.4 Evaluation

For evaluation purpose, we split our dataset into a training set and a testing set. We randomly select 100 movie reviews from each movie and evaluate the trained model on them to see whether they correctly predict the movie name or not.

4.5 Results

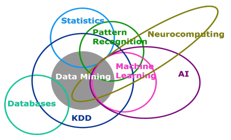
For verification, we compare the 2 different models on different MLP architecture. We mainly change the hidden layer and its size of the MLP architecture and hope to find the best result. As a result, our best model shows the accuracy of 71.7%. Table 1 shows the result of our test.

From there, we can see that the doc2vec has a slightly better performance than the word2vec model. From the data, we can see that our model basically have 60% to 70% accuracy. So we can conclude that the model is pretty good at predicting similar movies. From various models, we can see that the most simple MLP model with only one hidden layer of units 100 performs the best. This is potentially because our model is very simple with the output layer size of 10, so learning the feature of such model does not need a very deep MLP model.

However, there is still some works to do. First, the test dataset is only based on the new reviews for existing movies. We will still don't know how the model is performing on totally new movies. Second, we can train our own doc2vec model labeling with its own movie tag. We would think this can further improve the accuracy of our system. Third, we can further expand the movie size of the system to see how it perform. Currently, the movie size is 10, that is we are recommending 1 out of 10 movies based on the review. Trying to expand the movie size would be a good way to generalize our system.

5 Conclusion

In general, we implemented a system based on NLP training algorithm to recommend to the user the movies they might like based on their reviews to a certain movie. Evaluating such system on the self-generated dataset, we got a performance of 60% to 70% accuracy.



References

[1] SUVIR BHARGAV. “Efficient Features for Movie Recommendation Systems”. MA thesis. 2014.

[2] Robin Burke Negar Hariri Bamshad Mobasher and Yong Zheng. “Context-Aware Recommendation Based On Review Mining”. maththesis. Uppsala: DePaul University, 2016.

[3] James Salter and Nick Antonopoulos. “CinemaScreen Recommender Agent: Combining Collaborative and Content-Based Filtering”. In: *The Journal of Narrative Technique* (). Excerpt in Roger Matuz, ed. *Contemporary Literary Criticism*. Vol. 61. Detroit: Gale, 1990, pp. 204–208.

[4] Seong-Bae Park Xuan-Son Vu. “Mining User/Movie Preferred Features Based on Reviews for Video Recommendation System”. MA thesis. Cambridge, Mass.: Kyungpook National University, 2017.