Computer Science Department

Final Project: Iconic movie Similarities to each other

Course: Introduction to AI (CS156s4)
Course Instructor: Yulia Newton, Ph.D.

Houman Irani

Introduction/Background

The problem here is to find similarities between movies in the same genre. When we think of movies we can clearly understand that Casino is similar to Goodfellas by just watching the both movie, but doing a same thing with AI, requires appropriate Data and skills to do so. This Algorithm will compare 100 different movies, using their plot taken from Imdb and wikipedia. Once done it generates different genres, and then side each genre, similar movies in order of similarity

This Algorithm can be extremely helpful for streaming networks and movie recommending website to suggest people what to watch next based on their watched movie history. It can also help critiques to analyze movies even better than they used to.

Figure 1: Godfather vs Godfather part II

Dataset description

The dataset for this project is summary of plots for 100 well-know movies from both Wikipedia and Imdb. In the dataset, data are separated by 5 columns including Movies Rank, Title, Genre, Wiki_plot and finally Imdb_Plot. There are exactly 100 row which filled the main five columns. In this specific dataset Rank, Title, Genre, Wiki_Plot and Imdb_Plot are independent variables. On the other hand, Movie similarities is considered as dependent variable since it is the goal of the program to find which movie is closer to which. This dataset consists of free text making it a NLP problem to solve, then use K means cluster to complete.

In this particular problem there is no need for training data and test data since the goal is to compare movie plot and found their similarities based on the plots. Using this dataset we separate the data into groups know as clusters. We also needed to combine both Wiki_plot and Imdb_plot into a new column, because sometimes there are different pieces of information provided in each plot. By combining those we try to avoid overhead that is likely to happen. It is important to consider when using Clustering as unsupervised Learning there is no need for training and test data for the model.

Methodology

Algorithm

The algorithm used to classify movie plots will be the Natural Language Processing(NLP) algorithm and K Means Clustering. This algorithm is used to classify free text data and find the similarities by clustering. Since the data of this problem is movie plots taken fro Imdb and Wikipedia, the data from these plots are free text which is why we use the NLP algorithm in this problem.

Figure 2: Lord of the Ring Trilogy. These 3 movies are very close to each others.

Figure 3: Natural Language Processing (NLP)

Application to the project

Once we observed the data in the first place, we combine both plots into a new column and then work on tokenizing , stemming (NLP) and finally combing both together to establish a better meaning. Finally using K means clustering we calculate the similarity distance and draw the results at the end.

Analysis and Results

Test results

In the test results we finally get all the movies sorted in various groups based on the level of similarity in terms of plot they might have. In this case as expected Godfather is most similar to Godfather part II and Goodfellas and Fargo most similar movies to it are Psycho and North By Northwest. In general I think all the 100 movies are sorted properly based on the distance of similarity and results are acceptable.

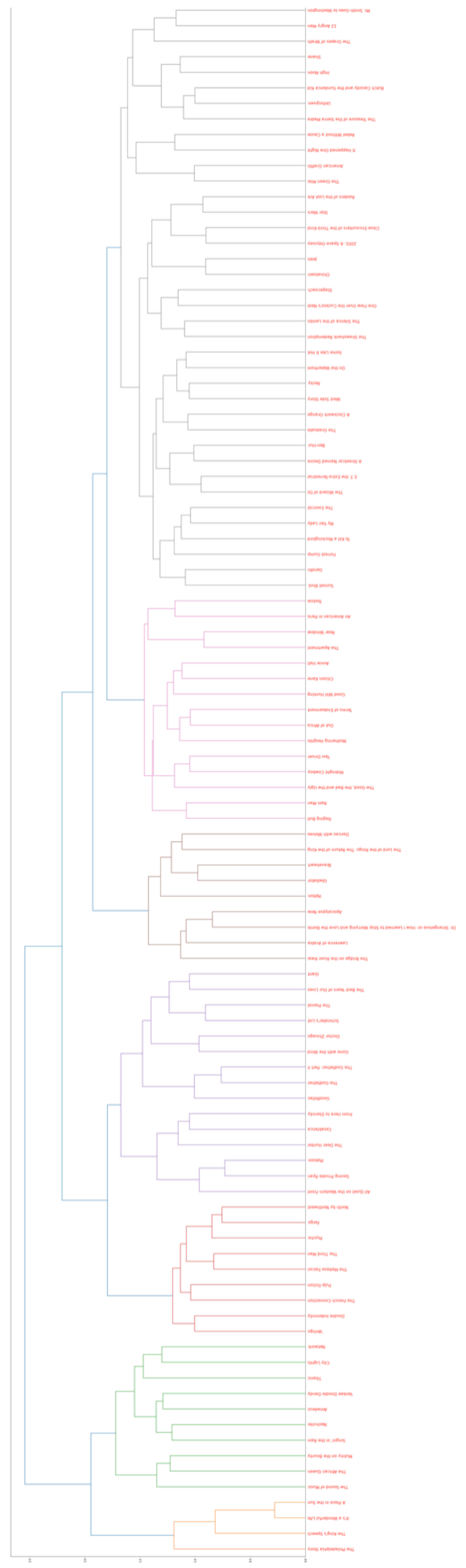

Figure 4: Results based on the similarity

Summary/Conclusions

The final results of this model is very accurate compared to the real word and critics view.
In conclusion this model based on NLP and Means Clustering was able to find similarities among movies listed in the dataset.

I would like to use a bigger dataset to compare more number of movies like Imdb top 250 as well as exploring new algorithms to be able to solve this problem with same accuracy if possible.

Key References

[1] Authors separated by a comma “Article name” in journal name, vol. XX, issue. X, pp. XXX-XXX, 20XX.

www.wikipedia.com
www.imdb.com

Acknowledgements

Any acknowledgements you deem appropriate (e.g. classmates, peers, instructors, family).