Anime Recommendation Summarization

CISB Natural Language Processing
Prof. Sohair Zaki
by Joseph Nguyen

For this final project, I will be using a Kaggle dataset Kaggle site: Anime Data

Github Repo: Github Repo

Youtube Demo: Youtube Demo of Anime Rec

Introduction

One of my few passions in life is staying up to date with the latest animes. This final project was exciting in terms of how open ended the actual content would be. I truly appreciate the opportunity to pick a dataset of my choice and be free in terms of what NLP concepts I wanted to apply. I was interested in personalizing this project to one of my personal interests, and creating a recommendation system for anime was one of them. In addition to my love for anime, another motivation in creating this specific project was derived from my own need for more content. I have watched multiple Anime's which is a testament to how invested I am in these unique stories and animation. I often have to grapple with the end of a beloved anime series, and how I could potentially find other shows that are similar to that story. This is the main reason for my interest in creating an anime recommendation system.

Summary

The dataset that I used came from Kaggle, and I was specifically working with the animes.csv file. Each row in this dataset correspsonded to an anime and there were properties of that anime such as release date, score, and more. My First task was to process the data by eliminating unnecessary columns and creating new columns to better fit my needs for exploratory data analysis. For exploratory data analysis, I used a countplot to measure the number of anime sper decade and year. The trend that I noticed from the countplot was a rise in anime's as the year continues. The score column was an indicator of how good an anime was; The lowest score was "Phantomi: Mini Anime" with a score of 1.25, highest score was "Fullmetal Alchemist: Brotherhood" with a score of 9.23, and the average score was 6.36. I was able to utilize a pivot table to express these highs, lows, and averages for each decade. A violinplot was used to visually represent the spread of scores for each decade. In addition a regplot was used in conjunction with the pearson value to dtermine that there is a strong positive correlation between members and score.

The NLP libraries and pretrained models that I used for this final prject included: *SpaCy* , *NLTK*, and *TfidfVectorizer* algorithm. From SpaCy, I wanted to focus on **named entity recognition** which helps identify and categorize specific entities in the text. The entities give important pieces of information. NLTK library is helpful for text processing. I was able to use **tokenization** to split text into individual tokens, **lemmatize** words to break down into root form, and filter out **stopwords**. I also used **Latent Dirichlet Allocation** to identify 12 topics amongst all the anime entries and a **wordcloud** was used to visualize the first two topics. The NLP application that I sought to implement was a recommendation system.

Conclusion

The recommendation system for anime works as I intended. I am able to put in an anime, and the model will output at most 10 animes that are very similar to the input anime. My only concern is that the anime dataset is a bit outdated. I am sure that the number of anime has continued to increase, especially considering the trend that was identified during exploratory data analysis. This model is limited by old data. In terms of how I can improve this model, I wanted to know how I can imcorprate the genre column into deciding animes based on select genres. The genre column was hard to work with considering there were multiple genres, and I had a hard time trying to figure out how each may overlap. The project was difficult and I had a lot of trouble iwth the recommendation system. Implementing it was a struggle, but I am happy that I was able to get a minimum output.