Philip Tam February 9th, 2019

Project Deliverable 1

1. Dataset

Synopses or summaries are meant to indicate the genre of a show, movie, or book. This project will attempt to do genre classification for movie, book, or series from synopses or summaries using the title and summary. I will be using the CMU Movie Summary Corpus Dataset from Bamman, O'Connor, & A. Smith (2013) because it has all of the necessary components (i.e. genre, summary, title) to do genre classification and over 42,000 movie plot summaries.

2. Methodology

I. Data Preprocessing

Since the movies and the data associated with it are each linked with an id, each genre will have array of ids where the ids associated are movies. The title, and genre will be extracted from the movie. If movies have multiple genres, then they will be in multiple arrays. Each summary will be lowercased, split into sentences, and all punctuations will be removed except periods. Since each summary will have varied sentences, the sentences per genre will be varied. Each genre will have 5000 movies where the training set will be split into 70%, validation 15%, and testing 15%.

II. Machine learning model

This project will try to replicate the Ertugrul's (2018) Bidirectional LNSM model that predicted genres based on a movie's plot summary while also adding an extra input the title of the movie. A Bidirectional LNSM model will be good because it will be able to preserve the information from the past and the future essentially understanding the context better. This will also provide a good challenge as a final project.

III. Final conceptualization

To present this model, I will be demoing a simple web/mobile application that will implement the model by taking as an input a title and summary and then displaying the genres that it has predicted and saving the predictions to a database.

References

Ertugrul, Ali Mert & KARAGOZ, Pinar. (2018). Movie Genre Classification from Plot Summaries Using Bidirectional LSTM. 10.1109/ICSC.2018.00043.

Bamman, O'Connor, & A. Smith. (2013). Learning Latent Personas of Film Characters. ACL 2013, Sofia, Bulgaria.