Movie classification using plot descriptions

Problem statement

In this project you are going to make a machine learning pipeline from scratch to solve the problem of movie classification using their plot descriptions. You will not be given any datasets, not even an "unclean" one. You must make your own version of the dataset by scraping which is a common practice in most data science/machine learning projects.

Below are explained some of the data resources you can scrape the data from. Once the dataset is made, starting from the standard bag-of-words representation you can use EDA to identify an appropriate input representation, perform feature engineering on it and find what is a good way to represent movie descriptions. The choice of the model is strongly dependent on the choice of the input representation. So, think carefully about which model fits best with which representation.

To more clearly lay out expectations, in this project you are not expected to apply state of the art deep learning models to obtain a high accuracy. A successful project will not necessarily be one that is highly accurate. The accuracy score of any model depends on the dataset, the exact problem formulation and most importantly, the evaluation metric you use. As you have a free hand to pick all of these, work closely with your TF to make sure you are on track.

As an example, one possible way to design the whole pipeline has been presented in a tutorial by a TF from last year (and the author of this assignment) [1]. Gloss over this tutorial. While you do not need to do so much for this project, at a high level, you need to make your *own* version of approaching the problem of movie classification.

Concrete goals:

1. Scrape a dataset of about 1000 movies
2. Try bag-of-words and word2vec representations for movie descriptions
3. Try a naive-bayes classifier and an SVM classifier for both of these text features.
4. Beyond this, use your creativity to answer interesting questions about movie genres and their plots and make your very own data science project!

Data Recources

We wish to scrape movie plot descriptions and genres. As both of these are high subjective for every movie, it is important to understand the data sources well, and make decisions that you believe will help your project the most. Here, I list 2 sources for movie genres (labels), and 2 sources for movie plots. As this is a project, there is no single correct answer. You must design a dataset as per your preferences, so get creative here!

1. **Data sources for labels (movie genres)-**

    - TMDB: A free, open-source dataset of movie information (https://www.themoviedb.org/?language=en) You will need to create and account to obtain an API key to download information. You can use the library 'tmdbsimple' for making easy API calls.
    - IMDB: The standard database for movie information. You can use the library 'imdb' to get data from imdb. Now that we know how to get information from TMDB, here's how we can get information about the same movie from IMDB. This makes it possible for us to combine more information, and get a richer dataset. Due to the differences between the two datasets, you will have to do some cleaning, however both of these datasets are extremely clean and it will be minimal.

2. **Data sources for movie reviews -**

    - TMDB, IMDB (as above)
    - Wikipedia: Most movies have a wiki page which contains a "plot" section. You can scrape movie description data from here.

High-level project goals

1. Going through the whole data creation process to understand the various design choices that are required.

2. Using conventional machine learning algorithms (not deep nets) to learn a classifier which takes in text as a bag-of-words representation.

3. Exploring other text representations like word2vec [2] and GloVE [3] word embeddings to see how the right representation affects the performance. This is an especially important lesson for non deep models.

4. To be able to take control of the whole machine learning pipeline, and create your own personal version of a project like [1]

References

[1] Spandan Madan. Spandan-Madan/DeepLearningProject: First release of the Deep Learning Project, July 2017.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.