Harvard University
Data Science II - Spring 2018
Final Project - Milestone 2 - Scope of Work

Group #11 Members:
Andrew Lund, andrewlund@g.harvard.edu
Nicholas Morgan, nim607@g.harvard.edu
Amay Umradia, amayumradia@gmail.com
Charles Webb, cwebb@college.harvard.edu

Project Statement and Background:

The problem statement for our final project is as follows:

"Can we build a robust movie plot and genre dataset, then apply bag-of-words and word2vec representations of individual plot descriptions to accurately classify movie genres using conventional machine learning algorithms?"

We will primarily be working with The Movie Database (TMDB) and Internet Movie Database (IMDB). Both sites give us the ability to scrape movie genres, plots, and other data for thousands of movies. We will focus initially on building a 1000 movie dataset from TMDB, then supplementing it with data scraped from IMDB.

We will be evaluating a hand-built set of 1000 English-language-only films. Our plan is to apply NLP methods including bag-of-words, and word2vec to transform each movie plot into a vector representation, then apply naive-bayes and SVM models to try and predict genres based on the vectorized plots. We will compare different combinations of datasets, NLP methods, and machine learning algorithms to benchmark the best mix.

Specifically for our word2vec transformation, we will use the Google News pre-trained word2vec model found here: https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit. This model was pre-trained on a Google News dataset containing about 100,000,000,000 words, and it contains 300-dimensional vectors for 3,000,000 words and phrases.

Each TMDB film has at least one, and more often more than one genre. A design decision will be made to address this issue. We have considered 2 methods: transforming the genres into a vector representation, and using scikit-learn to fit a multi-label model.

The success of our project will depend on two primary goals:

1. How well we can build a robust and "smart" dataset. This includes making sound design choices and backing those up with rational explanations.
2. Building traditional machine learning models (naive-bayes and SVM) that achieve reasonable accuracy for movie genre prediction.

Literature Review:

1. A Practical Guide to Support Vector Machines, Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, National Taiwan University, May 2016, https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
2. Distributed Representations of Words and Phrases and their Compositionality, Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, Google, https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
3. Predicting movie genres based on plot summaries, Quan Hoang, University of Massachusetts-Amherst, May 2018, https://arxiv.org/abs/1801.04813

Available Resources/Data:

We will be scraping data from the two sources below.

- ❏ TMDB (https://www.themoviedb.org/?language=en): An open-source dataset of movie information. We have already attained and used an API key to conduct an initial scraping of 1000 english language movies.
- ❏ IMDB (http://www.imdb.com): The standard database for movie information. We are currently adding another layer of IMDB data to our TMDB genres and plot descriptions.
- ❏ We will employ the tmdbsimple and imdb python libraries as well as machine-learning libraries for our analysis.
- ❏ Our group will use a private GitHub repo for collaboration (https://github.com/Nick-Morgan/109b_final)

Tasks, Deadlines, Assignments:

- ❏ Scrape dataset of 1000 movies with genre labels and plot descriptions from TMDB and IMDB. Decide on overall data design for modeling.
    - ❏ Deadline: 04APR18 (complete)
    - ❏ Assignment: Nicholas, Andrew

- ❏ Milestone 3: Submit EDA of genres and plots (counts, categories, years, word counts, rankings, etc) and Revised Project Statement.
    - ❏ Deadline: 20APR18
    - ❏ Assignment: Everyone

- ❏ Apply bag-of-words and word2vec on plot descriptions.
    - ❏ Deadline: 20APR18
    - ❏ Assignment: Amay, Charles

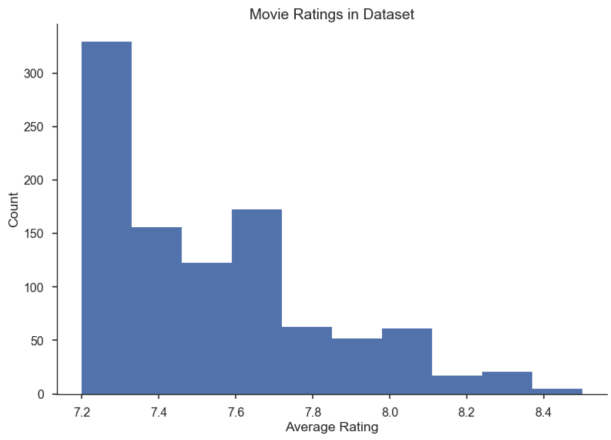
- ❏ Apply naive-bayes genre classifier to bag-of-words and word2vec plot descriptions.
    - ❏ Deadline: 27APR18
    - ❏ Assignment: Amay, Charles

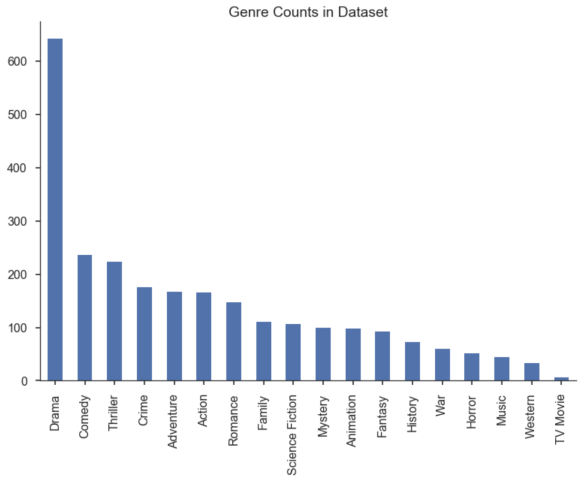
- ❏ Apply SVM genre classifier to bag-of-words and word2vec plot descriptions.
    - ❏ Deadline: 27APR18
    - ❏ Assignment: Nicholas, Andrew

- ❏ Submit final project: notebook, written report, poster, & peer evaluation.
    - ❏ Deadline: 02MAY18
    - ❏ Assignment: Everyone

Preliminary EDA with TODOs:

- ❏ After scraping movie data from TMDB and IMDB (genres/descriptions): visualize movie metrics to get an idea of what our dataset looks like. Is it skewed in any way? Should we adjust it? Do the metrics make sense? Here is a first look at our TMDB dataset:





  You can see from our initial scraping that most of our movies are in the upper-half of TMDB's rating system (1-10), and the distribution of genres skews heavily toward dramas. The genres are not exclusive for each movie (many have more than one associated with each film), so we are not surprised to see so many classified as drama, as many comedies, thrillers, crime, and adventure/actions movies will also fall in the drama category. We plan on our models taking into account more than one genre when training and scoring prediction accuracy.

- ❏ Determine whether or not IMDB data is needed in addition to TMDB: are the IMDB genres and plot descriptions different enough from TMDB to add value to the analysis? Should we train models with TMDB labels and IMDB plots? Which combination will produce the best accuracy?

  This question will be explored further as we learn more from our EDA process. Based on initial analysis, it seems that the plot descriptions differ enough between websites to add value to our model.

- ❏ Consider other interesting questions that can be answered with the data.

Links to other sources:
[1] https://github.com/Spandan-Madan/DeepLearningProject
[2] https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
[3] https://www.aclweb.org/anthology/D14-1162