CS 6140: Data Collection Report

Brian Kimmig, Jesus Zarate

1 How you obtained your data?

We obtained the base of our data from a dataset published on Kaggle. The dataset contained information on ~ 5000 movies. We essentially used this dataset for the movie list and the IMDB IDs. From the IMDB IDs we were able to perform requests to the OMDB API where we compiled the title, plot summary, and genres for all 5000 movies. From there we stored them in a JSON file, with each entry containing the fields ['title', 'plot', 'genres'].

2 Data Size

The general JSON data file is about 2.3 MB. This is just the data for the start, we very well may need/want to add movies to our data set. This is the raw data, we will processing the data to create features that can then be used for classification. Creating the features is a main chunk of the project, but I'd imagine that our matrices will be fairly large in size as they will have as many columns as we have movies (~ 5000), and depending on our k-gram method will most likely have a very large number of rows as well. (We should be able to handle these is a sparse format.)

3 Format/Storage

We are currently storing our data in a JSON file. It contains an entry for each movie and within each entry we have 3 descriptors of the movie – title, plot and genres – essentially giving the file 3 columns. The processing of this data will be a major part of this project, but in general we will always represent the data with matrices. Most likely these will be sparse matrices, due to the large number of paragraphs we will be processing.

4 Processing

As mentioned above, a large portion of our project is the processing of this data. We will be extracting features from the columns 'plot' and 'genre'. The plot features will give us information with which to classify the genre. Both of these items will be represented in matrix form, most likely with sparse matrices. We aim to test a number of processing methods within the k-gram universe, from single words to combinations of words. Most of our matrices will be represented using one-hot encoding. The processing of this data will come as part of our 'intermediate report'.

5 Simulating Similar Data

How would you simulate similar data?