milestone02_yujiaochen_brianho_jonjay_part01

April 12, 2017

0.1 AC209b / CS109b Final Project - Milestone 2 Part 01

Yujiao Chen, Brian Ho, Jonathan Jay // 04/12/2017

We are aware that you have little time this week, due to the midterm. So this milestone is a bit easier to achieve than the others. The goal for this week is to prepare the data for the modeling phase of the project. You should end up with a typical data setup of training data X and data labels Y.

The exact form of X and Y depends on the ideas you had previously. In general though Y should involve the genre of a movie, and X the features you want to include to predict the genre. Remember from the lecture that more features does not necessarily equal better prediction performance. Use your application knowledge and the insight you gathered from your genre pair analysis and additional EDA to design Y. Do you want to include all genres? Are there genres that you assume to be easier to separate than others? Are there genres that could be grouped together? There is no one right answer here. We are looking for your insight, so be sure to describe your decision process in your notebook.

In preparation for the deep learning part we strongly encourage you to have two sets of training data X, one with the metadata and one with the movie posters. Make sure to have a common key, like the movie ID, to be able to link the two sets together. Also be mindful of the data rate when you obtain the posters. Time your requests and choose which poster resolution you need. In most cases w500 should be sufficient, and probably a lower resolution will be fine

The notebook to submit this week should at least include:

- Discussion about the imbalanced nature of the data and how you want to address it
- Description of your data
- What does your choice of Y look like?
- Which features do you choose for X and why?
- How do you sample your data, how many samples, and why?

Important: You do not need to upload the data itself to Canvas.

0.2 Proposal

0.2.1 1. Primary research question: genre prediction

- How well do poster images predict three genres (romance, horror and sci-fi)?
- How well do the text of movie titles and descriptions predict these genres?

Data setup (Y): In Milestone 1 we found that the movie databases contain around 20 genre classes, some of which are much more common than others; that many movies are assigned multiple genre classes; and that a movie's genre assignments can vary across databases. For this project, however, we will sample only movies assigned to one of these three classes (excluding movies assigned to more than one of the three, but including movies which have also been assigned additional genres). Each genre will constitute exactly 1/3 of our sample. These movies are selected based on popularity, based on our assumption that more-popular movies are more representative of their genres, and balanced by year from 1954-present, for reasons of interest in part 2 below. Although the full dataset is over 8000 movies, we will select only the top 2000 movies by popularity, balanced across years, and with 20% reserved for testing. The subsampled size is based on (i) the maximum we believe is computationally feasible for question 1a, and (ii) our a priori belief that less-popular movies from within each genre may be less "pure" representations of the genre. The 80-20 split between training and testing data reflects accepted practice within the discpline. We will account for genre in assigning movies to the testing set, yielding perfectly balanced training and testing sets that will make it easier to detect trends in classifier performance.

Comments: This sampling method eliminates the imbalance problem in the broader database, while allowing us to answer the focused research question of how well CNNs can learn to distinguish among these genres. The particular genres were chosen using the correlation matrix we produced in Milestone 1, finding little overlap among these genres, and based on their sociological relevance. Choosing three classes allows us to consider relative distances among the classes and compare predictor performance in distinguishing among them–e.g., our hypothesis is that among these classes, sci-fi and horror are comparatively closer to each other than to romance, and will therefore be slightly harder to predict accurately.

Data setup (X): We are most interested in movie posters and movie titles & descriptions as predictors. We propose to optimize predictions using primarily (or exclusively) these predictors, rather than attempting to optimize predictions using whatever additional predictor data we might be able to access. Our initial thinking is to run models using these predictors separately (i.e. (a) poster vs. genre and (b) description vs. genre).

Comments: We prefer this approach because it will allow us to consider, in more depth, the relationship between each predictor and the genre classification. These features are all constructed with the intention of conveying information about the movie to prospective viewers (as opposed to, for example, language or director). We expect these features have a true relationship with genre, allowing the possibility of good classification accuracy — our submission this week includes an exploratory modeling exercise to predict horror vs. romance using PCA/SVM with test set accuracy of 87%, demonstrating the feasibility of the general approach. We also believe that the nature of this relationship represents an interesting research question: i.e. how effectively do title/descriptions and posters convey genre, and how well can algorithms learn to detect this relationship? Thus, while including additional features might (or might not) improve classification accuracy, they are not as relevant to the research questions that interest us most.

0.2.2 2. Secondary research question: poster age identification

- Can CNNs predict a movie's release decade based on its poster?

Brief discussion: Time permitting, we would like to set up a distinct classification task in which we sample from one genre (most likely sci-fi) and train a CNN to predict decade (e.g. 1960s/70s/80s/90s/00s/10s). We think this task may be of greatest substantive interest for science fiction movies, where we believe posters are representative of the era's visualizations of

alternative realities. Have these changed over times in ways that a CNN can learn to identify? A priori we think color combinations may be especially predictive of decade.

0.3 Data Collection

Getting TMDB metadata — poster collection is occuring separately, and for reasons of size not uploaded to Canvas.

In [2]:
```
## Some code to get data
## Let's import some libraries!
import imdb
import json
import requests
import pandas as pd
import numpy as np
import time
import matplotlib
%matplotlib inline
```

In [4]:
```
## Get the genre codes from IMDB
payload = {'api_key': '9290a6fe9125b32e7bbe5512036be0d0'}
r = requests.get('https://api.themoviedb.org/3/genre/movie/list', params=pa

genres = pd.DataFrame.from_dict(r.json()["genres"])
genres = genres.set_index("id")
print genres
```

```
                name
id
28            Action
12         Adventure
16         Animation
35            Comedy
80             Crime
99       Documentary
18             Drama
10751         Family
14           Fantasy
36           History
27            Horror
10402          Music
9648         Mystery
10749        Romance
878    Science Fiction
10770        TV Movie
53           Thriller
10752             War
37            Western
```

```
In [5]: genres = genres["name"].to_dict()
        genres

Out[5]: {12: u'Adventure',
         14: u'Fantasy',
         16: u'Animation',
         18: u'Drama',
         27: u'Horror',
         28: u'Action',
         35: u'Comedy',
         36: u'History',
         37: u'Western',
         53: u'Thriller',
         80: u'Crime',
         99: u'Documentary',
         878: u'Science Fiction',
         9648: u'Mystery',
         10402: u'Music',
         10749: u'Romance',
         10751: u'Family',
         10752: u'War',
         10770: u'TV Movie'}

In [6]: ### Queries to TMDB

        # intial API parameters
        def get_movies(years, page_limit, genre):
            # Outer loop for ever year in range
            for i, year in enumerate(years):
                start = time.time()

                # Define initial API parameters for genre and year
                payload = {'api_key': '9290a6fe9125b32e7bbe5512036be0d0',
                           'sort_by':'popularity.desc',
                           'primary_release_year': year,
                           'page': 1,
                           'language':'en-US',
                           'with_genres': genre} #"878|27|10749"

                r = requests.get('https://api.themoviedb.org/3/discover/movie?', pa
                print 'For ', year, ' there are ', r.json()['total_results'], ' tot

                # For first year, create the data frame. Otherwise, add first page
                if i == 0:
                    tmdb_movies = pd.io.json.json_normalize(r.json()['results'])
                else:
                    tmdb_movies = pd.concat([tmdb_movies, pd.io.json.json_normalize
```

```
                # Set max pages to smaller of five or total number
                if r.json()['total_pages'] < page_limit:
                    page_max = r.json()['total_pages']
                else:
                    page_max = page_limit

                # Wait function for polite API querying
                delay = time.time()-start
                if delay < 0.25:
                    time.sleep(0.25-delay)

                if page_max > 1:
                    # Inner loop for every page up to max, startigng with page 2.
                    for page in range(2, page_max+1):
                        start = time.time()

                        payload = {'api_key': '9290a6fe9125b32e7bbe5512036be0d0',
                               'sort_by':'popularity.desc',
                               'primary_release_year': year,
                               'page': page,
                               'language':'en-US',
                               'with_genres': genre}#"878|27|10749"}

                        r = requests.get('https://api.themoviedb.org/3/discover/mov

                        tmdb_movies = pd.concat([tmdb_movies, pd.io.json.json_norma

                        delay = time.time()-start
                        if delay < 0.25:
                            time.sleep(0.25-delay)

            return tmdb_movies
```

In [7]:
```
# Get science fiction movies from 1930
movies_scifi = get_movies(range(1930,2017), 2, 878)
movies_scifi.shape
```

```
For  1930  there are  2  total results across  1  total pages.
For  1931  there are  4  total results across  1  total pages.
For  1932  there are  6  total results across  1  total pages.
For  1933  there are  7  total results across  1  total pages.
For  1934  there are  7  total results across  1  total pages.
For  1935  there are  10  total results across  1  total pages.
For  1936  there are  13  total results across  1  total pages.
For  1937  there are  7  total results across  1  total pages.
For  1938  there are  8  total results across  1  total pages.
For  1939  there are  9  total results across  1  total pages.
For  1940  there are  11  total results across  1  total pages.
```

For 1941 there are 6 total results across 1 total pages.
For 1942 there are 5 total results across 1 total pages.
For 1943 there are 7 total results across 1 total pages.
For 1944 there are 7 total results across 1 total pages.
For 1945 there are 5 total results across 1 total pages.
For 1946 there are 2 total results across 1 total pages.
For 1947 there are 3 total results across 1 total pages.
For 1948 there are 10 total results across 1 total pages.
For 1949 there are 6 total results across 1 total pages.
For 1950 there are 7 total results across 1 total pages.
For 1951 there are 16 total results across 1 total pages.
For 1952 there are 10 total results across 1 total pages.
For 1953 there are 25 total results across 2 total pages.
For 1954 there are 25 total results across 2 total pages.
For 1955 there are 25 total results across 2 total pages.
For 1956 there are 28 total results across 2 total pages.
For 1957 there are 37 total results across 2 total pages.
For 1958 there are 50 total results across 3 total pages.
For 1959 there are 36 total results across 2 total pages.
For 1960 there are 27 total results across 2 total pages.
For 1961 there are 25 total results across 2 total pages.
For 1962 there are 34 total results across 2 total pages.
For 1963 there are 33 total results across 2 total pages.
For 1964 there are 34 total results across 2 total pages.
For 1965 there are 65 total results across 4 total pages.
For 1966 there are 70 total results across 4 total pages.
For 1967 there are 69 total results across 4 total pages.
For 1968 there are 52 total results across 3 total pages.
For 1969 there are 48 total results across 3 total pages.
For 1970 there are 41 total results across 3 total pages.
For 1971 there are 49 total results across 3 total pages.
For 1972 there are 49 total results across 3 total pages.
For 1973 there are 59 total results across 3 total pages.
For 1974 there are 47 total results across 3 total pages.
For 1975 there are 57 total results across 3 total pages.
For 1976 there are 36 total results across 2 total pages.
For 1977 there are 64 total results across 4 total pages.
For 1978 there are 61 total results across 4 total pages.
For 1979 there are 70 total results across 4 total pages.
For 1980 there are 70 total results across 4 total pages.
For 1981 there are 76 total results across 4 total pages.
For 1982 there are 78 total results across 4 total pages.
For 1983 there are 85 total results across 5 total pages.
For 1984 there are 100 total results across 5 total pages.
For 1985 there are 108 total results across 6 total pages.
For 1986 there are 111 total results across 6 total pages.
For 1987 there are 149 total results across 8 total pages.
For 1988 there are 131 total results across 7 total pages.

```
For  1989  there are  131  total results across  7  total pages.
For  1990  there are  109  total results across  6  total pages.
For  1991  there are  103  total results across  6  total pages.
For  1992  there are  79  total results across  4  total pages.
For  1993  there are  94  total results across  5  total pages.
For  1994  there are  117  total results across  6  total pages.
For  1995  there are  128  total results across  7  total pages.
For  1996  there are  126  total results across  7  total pages.
For  1997  there are  128  total results across  7  total pages.
For  1998  there are  128  total results across  7  total pages.
For  1999  there are  128  total results across  7  total pages.
For  2000  there are  118  total results across  6  total pages.
For  2001  there are  139  total results across  7  total pages.
For  2002  there are  128  total results across  7  total pages.
For  2003  there are  150  total results across  8  total pages.
For  2004  there are  158  total results across  8  total pages.
For  2005  there are  169  total results across  9  total pages.
For  2006  there are  188  total results across  10  total pages.
For  2007  there are  190  total results across  10  total pages.
For  2008  there are  210  total results across  11  total pages.
For  2009  there are  266  total results across  14  total pages.
For  2010  there are  253  total results across  13  total pages.
For  2011  there are  270  total results across  14  total pages.
For  2012  there are  280  total results across  14  total pages.
For  2013  there are  339  total results across  17  total pages.
For  2014  there are  357  total results across  18  total pages.
For  2015  there are  421  total results across  22  total pages.
For  2016  there are  331  total results across  17  total pages.
```

Out[7]: (2613, 14)

In [8]:
```
# Get horror movies from 1930
movies_horror = get_movies(range(1930,2017), 2, 27)
movies_horror.shape
```

```
For  1930  there are  2  total results across  1  total pages.
For  1931  there are  10  total results across  1  total pages.
For  1932  there are  23  total results across  2  total pages.
For  1933  there are  19  total results across  1  total pages.
For  1934  there are  10  total results across  1  total pages.
For  1935  there are  16  total results across  1  total pages.
For  1936  there are  15  total results across  1  total pages.
For  1937  there are  4  total results across  1  total pages.
For  1938  there are  4  total results across  1  total pages.
For  1939  there are  15  total results across  1  total pages.
For  1940  there are  17  total results across  1  total pages.
For  1941  there are  12  total results across  1  total pages.
```

For 1942 there are 18 total results across 1 total pages.
For 1943 there are 16 total results across 1 total pages.
For 1944 there are 19 total results across 1 total pages.
For 1945 there are 17 total results across 1 total pages.
For 1946 there are 18 total results across 1 total pages.
For 1947 there are 6 total results across 1 total pages.
For 1948 there are 9 total results across 1 total pages.
For 1949 there are 17 total results across 1 total pages.
For 1950 there are 4 total results across 1 total pages.
For 1951 there are 9 total results across 1 total pages.
For 1952 there are 7 total results across 1 total pages.
For 1953 there are 14 total results across 1 total pages.
For 1954 there are 20 total results across 1 total pages.
For 1955 there are 22 total results across 2 total pages.
For 1956 there are 28 total results across 2 total pages.
For 1957 there are 51 total results across 3 total pages.
For 1958 there are 63 total results across 4 total pages.
For 1959 there are 56 total results across 3 total pages.
For 1960 there are 49 total results across 3 total pages.
For 1961 there are 49 total results across 3 total pages.
For 1962 there are 53 total results across 3 total pages.
For 1963 there are 59 total results across 3 total pages.
For 1964 there are 65 total results across 4 total pages.
For 1965 there are 64 total results across 4 total pages.
For 1966 there are 64 total results across 4 total pages.
For 1967 there are 63 total results across 4 total pages.
For 1968 there are 93 total results across 5 total pages.
For 1969 there are 78 total results across 4 total pages.
For 1970 there are 117 total results across 6 total pages.
For 1971 there are 133 total results across 7 total pages.
For 1972 there are 181 total results across 10 total pages.
For 1973 there are 182 total results across 10 total pages.
For 1974 there are 161 total results across 9 total pages.
For 1975 there are 134 total results across 7 total pages.
For 1976 there are 105 total results across 6 total pages.
For 1977 there are 109 total results across 6 total pages.
For 1978 there are 117 total results across 6 total pages.
For 1979 there are 107 total results across 6 total pages.
For 1980 there are 167 total results across 9 total pages.
For 1981 there are 201 total results across 11 total pages.
For 1982 there are 187 total results across 10 total pages.
For 1983 there are 163 total results across 9 total pages.
For 1984 there are 146 total results across 8 total pages.
For 1985 there are 182 total results across 10 total pages.
For 1986 there are 203 total results across 11 total pages.
For 1987 there are 286 total results across 15 total pages.
For 1988 there are 333 total results across 17 total pages.
For 1989 there are 334 total results across 17 total pages.

```
For  1990  there are  238  total results across  12  total pages.
For  1991  there are  174  total results across  9  total pages.
For  1992  there are  149  total results across  8  total pages.
For  1993  there are  140  total results across  7  total pages.
For  1994  there are  117  total results across  6  total pages.
For  1995  there are  139  total results across  7  total pages.
For  1996  there are  145  total results across  8  total pages.
For  1997  there are  133  total results across  7  total pages.
For  1998  there are  144  total results across  8  total pages.
For  1999  there are  184  total results across  10  total pages.
For  2000  there are  205  total results across  11  total pages.
For  2001  there are  222  total results across  12  total pages.
For  2002  there are  235  total results across  12  total pages.
For  2003  there are  261  total results across  14  total pages.
For  2004  there are  343  total results across  18  total pages.
For  2005  there are  399  total results across  20  total pages.
For  2006  there are  524  total results across  27  total pages.
For  2007  there are  550  total results across  28  total pages.
For  2008  there are  535  total results across  27  total pages.
For  2009  there are  571  total results across  29  total pages.
For  2010  there are  542  total results across  28  total pages.
For  2011  there are  590  total results across  30  total pages.
For  2012  there are  643  total results across  33  total pages.
For  2013  there are  699  total results across  35  total pages.
For  2014  there are  837  total results across  42  total pages.
For  2015  there are  890  total results across  45  total pages.
For  2016  there are  748  total results across  38  total pages.
```

Out[8]: (2771, 14)

In [9]:
```
# Get romance movies from 1930
movies_romance = get_movies(range(1930,2017), 2, 10749)
movies_romance.shape
```

```
For  1930  there are  119  total results across  6  total pages.
For  1931  there are  111  total results across  6  total pages.
For  1932  there are  113  total results across  6  total pages.
For  1933  there are  125  total results across  7  total pages.
For  1934  there are  131  total results across  7  total pages.
For  1935  there are  146  total results across  8  total pages.
For  1936  there are  121  total results across  7  total pages.
For  1937  there are  160  total results across  8  total pages.
For  1938  there are  126  total results across  7  total pages.
For  1939  there are  104  total results across  6  total pages.
For  1940  there are  106  total results across  6  total pages.
For  1941  there are  129  total results across  7  total pages.
For  1942  there are  113  total results across  6  total pages.
```

For 1943 there are 87 total results across 5 total pages.
For 1944 there are 83 total results across 5 total pages.
For 1945 there are 66 total results across 4 total pages.
For 1946 there are 78 total results across 4 total pages.
For 1947 there are 85 total results across 5 total pages.
For 1948 there are 80 total results across 4 total pages.
For 1949 there are 94 total results across 5 total pages.
For 1950 there are 79 total results across 4 total pages.
For 1951 there are 82 total results across 5 total pages.
For 1952 there are 92 total results across 5 total pages.
For 1953 there are 95 total results across 5 total pages.
For 1954 there are 93 total results across 5 total pages.
For 1955 there are 123 total results across 7 total pages.
For 1956 there are 101 total results across 6 total pages.
For 1957 there are 99 total results across 5 total pages.
For 1958 there are 99 total results across 5 total pages.
For 1959 there are 79 total results across 4 total pages.
For 1960 there are 103 total results across 6 total pages.
For 1961 there are 83 total results across 5 total pages.
For 1962 there are 86 total results across 5 total pages.
For 1963 there are 91 total results across 5 total pages.
For 1964 there are 84 total results across 5 total pages.
For 1965 there are 93 total results across 5 total pages.
For 1966 there are 89 total results across 5 total pages.
For 1967 there are 101 total results across 6 total pages.
For 1968 there are 102 total results across 6 total pages.
For 1969 there are 99 total results across 5 total pages.
For 1970 there are 104 total results across 6 total pages.
For 1971 there are 99 total results across 5 total pages.
For 1972 there are 84 total results across 5 total pages.
For 1973 there are 86 total results across 5 total pages.
For 1974 there are 95 total results across 5 total pages.
For 1975 there are 101 total results across 6 total pages.
For 1976 there are 92 total results across 5 total pages.
For 1977 there are 99 total results across 5 total pages.
For 1978 there are 129 total results across 7 total pages.
For 1979 there are 119 total results across 6 total pages.
For 1980 there are 128 total results across 7 total pages.
For 1981 there are 123 total results across 7 total pages.
For 1982 there are 130 total results across 7 total pages.
For 1983 there are 131 total results across 7 total pages.
For 1984 there are 159 total results across 8 total pages.
For 1985 there are 171 total results across 9 total pages.
For 1986 there are 151 total results across 8 total pages.
For 1987 there are 161 total results across 9 total pages.
For 1988 there are 177 total results across 9 total pages.
For 1989 there are 157 total results across 8 total pages.
For 1990 there are 139 total results across 7 total pages.

```
For  1991  there are  153  total results across  8  total pages.
For  1992  there are  164  total results across  9  total pages.
For  1993  there are  188  total results across  10  total pages.
For  1994  there are  190  total results across  10  total pages.
For  1995  there are  230  total results across  12  total pages.
For  1996  there are  230  total results across  12  total pages.
For  1997  there are  285  total results across  15  total pages.
For  1998  there are  280  total results across  14  total pages.
For  1999  there are  319  total results across  16  total pages.
For  2000  there are  334  total results across  17  total pages.
For  2001  there are  373  total results across  19  total pages.
For  2002  there are  400  total results across  20  total pages.
For  2003  there are  433  total results across  22  total pages.
For  2004  there are  412  total results across  21  total pages.
For  2005  there are  484  total results across  25  total pages.
For  2006  there are  510  total results across  26  total pages.
For  2007  there are  525  total results across  27  total pages.
For  2008  there are  556  total results across  28  total pages.
For  2009  there are  615  total results across  31  total pages.
For  2010  there are  626  total results across  32  total pages.
For  2011  there are  581  total results across  30  total pages.
For  2012  there are  599  total results across  30  total pages.
For  2013  there are  682  total results across  35  total pages.
For  2014  there are  780  total results across  39  total pages.
For  2015  there are  768  total results across  39  total pages.
For  2016  there are  788  total results across  40  total pages.
```

Out[9]: (3480, 14)

In [10]:
```
# Join our individual genre data frames
movies = pd.concat([movies_scifi, movies_horror, movies_romance])
print movies.shape
movies.head()

# To balance classes, let's only use data from years where there are at le
movies["release_date"] = pd.to_datetime(movies["release_date"])
movies = movies[movies["release_date"] >= "1954"]
movies.tail()
```

```
(8864, 14)
```

Out[10]:

	adult	backdrop_path	genre_ids	id	\
15	False	/elSS1VQHKZF1j2ZxOnVdA2VhZzU.jpg	[18, 10402, 10749]	382399	
16	False	/xA959uQHOMCV1as2ZbUfm06dyXw.jpg	[35, 18, 10749]	291328	
17	False	/2fgz6p16JYI1OnhwFIuVCWCFayG.jpg	[10749, 18]	315880	
18	False	/faccYsNO4Oxf5nS8gVloYRp5J7Y.jpg	[10749, 35]	390669	
19	False	/2pqYl9JHfzBuZbnwf3m59s4tvDT.jpg	[18, 9648, 10749]	316021	

```
   original_language     original_title  \
15                en        High Strung
16                en  Rules Don't Apply
17                en  La corrispondenza
18                it           Un bacio
19                en       Frank & Lola

                                              overview  popularity  \
15  When a hip hop violinist busking in the New Yo...    1.832925
16  It's Hollywood, 1958. Small town beauty queen ...    1.815620
17  The relationship between Ed, a married astrono...    1.674114
18  Lorenzo, Blue and Antonio have a lot in common...    1.647663
19  A psychosexual noir love story, set in Las Veg...    1.596720

                         poster_path release_date              title  vid
15  /vBpiQ4urEIlSOntCaNNMkb1SBNR.jpg   2016-04-08        High Strung  Fal
16  /lCdgGqZ0ibzmtUoqdgGOcXAoA59.jpg   2016-11-23  Rules Don't Apply  Fal
17  /ghuK137xSNdtZWxXqG96GUFLeYR.jpg   2016-01-14  The Correspondence  Fal
18  /eQw1JGdeMzaAADZVZPbFm6OKowR.jpg   2016-03-31           One Kiss  Fal
19  /9ZdlnjePpUhxRWz69KjqhIe7H5U.jpg   2016-12-09       Frank & Lola  Fal

    vote_average  vote_count
15           7.0         148
16           6.0          39
17           6.0          66
18           7.1          37
19           6.1          25
```

```
In [10]: # Export dataframe
         movies.to_csv("movies_from_1954.csv", encoding = "utf-8")
```

0.4 Word Cloud and PCA Prep

Here we accomplished several tasks: 1) Collecting horror movies and romance movies from top 10000 movie data base, analyzing the contents of the movie title and movie overview, creating the wordclouds that show the most common words in two different genres; 2) Creating a corpus from the above-mentioned movies, filter the most frequent words, using a long boolean vector to indicate each word's appearance in each movie's title or overview; 3) Conducting PCA and choosing first PCs that explain 80% of variance in the data, cleaning the data format and outputting to .csv files for further PCA and SVM study in R.

See milestone02_yujiaochen_brianho_jonjay_part02.ipynb

0.5 Transition to R for PCA and SVM

See attached Rmd file — in R we accomplished several tasks: 1) From the dataframe of 527 Horror movies and 570 Romance movies prepared by python codes (please refer to Milestone_2_Part_1.pdf), we randomly chose 300 movies to be the test set, and the rest of them to

be the training set; 2) Using PCA to extract first 150 PCs that explain 80% of the variance in data, and projecting the training data and testing data to get PC score in each sets; 3) Employing SVM with radial basis function to classify the horror and romance movies based on their PC scores. The parameter (gamma and cost) has been found through tuning; 4) The final predicting accuracy on test set using this model is 87%, which is a satisfactory result.

See milestone02_yujiaochen_brianho_jonjay_part03.rmd