Kevin Luu and Cameron Wolfe
Problem Solving & Software Design
10/21/2020

Assignment 2 Reflection

1. Project Overview

Given the creative freedom, we decided to explore 2 data sources. The first source was Reddit, where we hoped to learn more about what people's thoughts on the stock market and economy. This would enable us to compare and contrast people's sentiments.Using praw, a built-in python tool, we were able to create various for-loops to see different attributes and complete a sentiment analysis. For our second source IMDb, we hoped to become more comfortable extracting data from the web and IMDb was great for pulling information using very similar methodology which allowed us to focus on internalizing the process. Our goal was to easily manipulate the database and challenge ourselves to extract as much information as possible, giving us a larger sample with which to run term frequency analysis.

2. Implementation

For our Reddit analysis, we first began with scraping Reddit's thread titles, comments, and replies to comments in hopes of using word frequency functions to gain a better understanding of what people feel about the stock market and economy. The Reddit tool, praw, allowed us to code with Reddit-specific attributes to find certain data within Reddit. We created individual functions to mine through submissions, comments, and replies to get multiple levels of data. Our word_freq dictionary would continuously be updated through each function until all the words are outputted with their frequencies. To find the most common words, we had a variable, common_words, that held an empty list that iterated through the words we had in our dictionary and outputted the most common words within the empty list. In our other Reddit analysis, we created simple functions to find top moderators of specific subreddits who are known as community leaders. We also looked at titles from 3 different categories (top, new, controversial) within the same topic to compare and contrast the top headlines. Lastly, using the natural language toolkit (NLTK) in python, I was able to perform a sentiment analysis on post titles to

see if stock market and economy sentiment is negative, positive, or neutral. By categorizing all the different titles into the 3 categories mentioned above and having examples of positive and negative titles, we got an overall sense of what people were talking about.

Our second source, IMDb, allowed us a different method by which to extract data on a completely distinct topic. First, we installed and used imdbpie as a means to access the IMDb database. From here, we were able to extract a plethora of information from the IMDb database including top/bottom movies, movie reviews, in-depth information for a film, actor information - which we expanded upon to extract various data points from the lead role in any film. There was such a wealth of data that could have been extracted from this source that we were forced to spend time narrowing down very precisely what data we found most interesting and most appropriate for analysis. I mentioned earlier that our goal with IMDb was to practice extracting data from the web database which is evident through our code as most of the data gathering code (IMDB_data.py) consists of establishing variables and using print functions to display information scrubbed from the database. The first code I wrote was a simple for loop that would display every movie title registered with IMDb that has the given title. This was interesting because I found there were nearly 20 movies dubbed “Inception” but felt as if this wasn’t very interesting to analyze. Next once we decided to gather data based on a single film, we were able to gather information about the film (title, year, rating, directors, cast, review) which prints cleanly. To expand upon this, we decided to narrow our focus by pulling the lead role in the film (Tim Robbins in the case of The Shawshank Redemption) and display their name, birth date, height (in both feet and meters), every film they’ve starred in during their career and a fun fact about the actor. I played around gathering data for a variety of films but decided to focus on one of my favorites, The Shawshank Redemption. In using the lengthy Shawshank review, we set out to conduct a sentiment analysis with the hopes of gathering further insight into how frequently certain words are used. This is complimented by using Term Frequency-Inverse Document Frequency, which aided in allocating proper weights to the words from the text, which is especially useful to consider buffer words like “the” “and” & “to” and properly allocate weight based on their higher frequency in the English language.

3. Results

IMDb Results:

The analysis of the IMDb data proved challenging as the text mediums (a common point of analysis for other data sources) were largely based on movie reviews. Initially, I spent the first few days after completing my data-gathering code trying to work with a data set compiled at Stanford of 50 thousand movie reviews, half being positive with the other half being negative. I worked through a sentiment analysis of these reviews to analyze common words and gauge public sentiment by attributing individual words as either positive or negative. This proved to be very challenging because of the nature of the dataset being very compressed making it difficult to read each individual .txt file. Furthermore, after accidentally importing 100,000 .txt files into our repository I decided this dataset was 1) difficult to manipulate and 2) rather underwhelming as far as tangible analysis goes. In other words, I found it rather dull looking at reviews for random movies because it strayed from the core of the project which was to work with something we found interesting and for something we were passionate about. At this late-stage in the course of our project, I transitioned to analyzing a single movie: The Shawshank Redemption. Since this is a very widely regarded film (on IMDb and one of my personal favorites), I found it fascinating to be able to pull more information about the film than I could've imagined. I decided to use the top review I pulled from the film and conducted a sentiment analysis to 1) analyze word frequencies to see if there were any interesting patterns and 2) conduct a word relevancy analysis using Term Frequency-Inverse Document Frequency. TFIDF is a numerical statistic that is intended to reflect how important a word is to our text as a whole, and the tf-idf value increases proportionally to the number of times a word appears in our text and is offset by the number of documents in the corpus that contain the word. This helps to adjust for the fact that some words appear more frequently in general.

Movie Info:

```
MOVIE INFO
The Shawshank Redemption - 1994
rating: 9.3
directors: Frank Darabont
actors: Tim Robbins, Morgan Freeman, Bob Gunton, William Sadler, Clanc
avid Proval, Joseph Ragno, Jude Ciccolella, Paul McCrane, Renee Blaine
Alan R. Kessler, Morgan Lund, Cornell Wallace, Gary Lee Davis, Neil Su
 Haley, Dana Snyder, John D. Craig, Ken Magee, Eugene C. DePasquale, B
as, Charlie Kearns, Rob Reider, Brian Brophy, Paul Kennedy, James Babs
es, Sergio Kato, Michael Lightsey, George Macready, Christopher Page,
name: Tim Robbins
birth date: 1958-10-16
height: 6' 5" (1.96 m)
trivia: Has played in the Heroes of Hockey game at the National Hockey
bio title refs: Bull Durham (1988), Cadillac Man (1990), The Hudsucker
mption (1994), Dead Man Walking (1995), Mystic River (2003), Cradle Wi
(2005), Howard the Duck (1986), Seven Samurai (1954), Top Gun (1986),
{'title': 'The Shawshank Redemption', 'year': '1994', 'imdb_id': 'tt01
hitchcockthelegend
The Shawshank Redemption is written and directed by Frank Darabont. It
```

Word Frequency:


```
{'the': 294, 'shawshank': 263, 'redemption': 239, 'is': 160, 'written': 338, 'and': 12, 'directed': 78, 'by': 47, 'frank': 116, 'darabont': 68, 'it': 161, 'an': 11, 'adaptation': 6, 'of'
: 210, 'stephen': 282, 'king': 166, 'novella': 208, 'rita': 246, 'hayworth': 136, 'starring': 278, 'tim': 306, 'robbins': 247, 'morgan': 186, 'freeman': 117, 'film': 109, 'portrays': 227
, 'story': 284, 'andy': 13, 'dufresne': 85, 'banker': 34, 'who': 329, 'sentenced': 258, 'to': 309, 'two': 312, 'life': 169, 'sentences': 259, 'at': 27, 'state': 280, 'prison': 229, 'for'
: 113, 'apparently': 20, 'murdering': 193, 'his': 141, 'wife': 330, 'her': 139, 'lover': 174, 'finds': 111, 'tough': 311, 'going': 125, 'but': 46, 'solace': 271, 'in': 151, 'friendship':
 118, 'he': 137, 'forms': 114, 'with': 334, 'fellow': 106, 'inmate': 155, 'ellis': 91, 'red': 237, 'redding': 238, 'while': 326, 'things': 301, 'start': 279, 'pick': 223, 'up': 314, 'whe
n': 324, 'warden': 319, 'job': 164, 'more': 185, 'befitting': 37, 'talents': 290, 'as': 25, 'however': 146, 'arrival': 24, 'another': 16, 'vastly': 316, 'change': 52, 'all': 7, 'them': 2
95, 'there': 297, 'was': 322, 'no': 202, 'fanfare': 103, 'or': 216, 'bunting': 45, 'put': 233, 'out': 218, 'release': 242, 'back': 31, '94': 0, 'title': 308, 'that': 293, 'didnt': 77, 'g
ive': 121, 'much': 192, 'inkling': 154, 'anyone': 18, 'about': 1, 'what': 323, 'columbia': 61, 'pictures': 224, 'unsure': 313, 'how': 145, 'market': 179, 'barely': 35, 'registered': 241,
 'box': 42, 'office': 213, 'come': 62, 'academy': 3, 'award': 29, 'time': 307, 'received': 236, 'several': 261, 'nominations': 203, 'although': 9, 'won': 335, 'none': 204, 'stirred': 283
, 'interest': 157, 'its': 162, 'home': 143, 'entertainment': 96, 'rest': 243, 'they': 299, 'say': 254, 'history': 142, 'finally': 110, 'found': 115, 'audience': 28, 'saw': 253, 'propelle
d': 230, 'almost': 8, 'mythical': 195, 'proportions': 231, 'endearing': 95, 'modern': 184, 'day': 69, 'classic': 59, 'something': 273, 'has': 134, 'delighted': 73, 'fans': 104, 'whilst':
 327, 'simultaneously': 267, 'baffling': 33, 'detractors': 74, 'one': 215, 'thing': 300, 'sure': 289, 'though': 305, 'which': 325, 'ever': 98, 'side': 264, 'fence': 107, 'you': 340, 'sit
': 268, 'on': 214, 'continues': 66, 'gather': 120, 'new': 200, 'simply': 266, 'will': 331, 'never': 199, 'go': 124, 'away': 30, 'loose': 172, 'status': 281, 'possibly': 228, 'simplicity'
: 265, 'sends': 257, 'some': 272, 'haters': 135, 'into': 158, 'cinematic': 56, 'spasms': 276, 'implausible': 149, 'plot': 226, 'apparent': 19, 'sentimental': 260, 'edge': 89, 'makes': 17
6, 'nonsense': 205, 'are': 21, 'chief': 54, 'complaints': 63, 'from': 119, 'those': 304, 'dislike': 79, 'passion': 219, 'yet': 339, 'characters': 53, 'this': 302, 'richly': 245, 'drawn':
 84, 'so': 270, 'movingly': 190, 'performed': 221, 'strikes': 285, 'me': 180, 'churlish': 55, 'do': 80, 'down': 82, 'human': 147, 'drama': 83, 'dealing': 71, 'hope': 144, 'faith': 102, '
aspect': 26, 'indeed': 153, 'acts': 5, 'counterpoint': 67, 'suffering': 287, 'degradation': 72, 'shattering': 262, 'soul': 274, 'involving': 159, 'our': 217, 'protagonist': 232, 'need':
197, 'connection': 65, 'needed': 198, 'than': 292, 'during': 86, 'incarceration': 152, 'given': 122, 'quite': 234, 'terrific': 291, 'performances': 220, 'better': 38, 'sublimely': 286, '
making': 177, 'easy': 88, 'easiest': 87, 'world': 337, 'warm': 320, 'support': 288, 'arent': 22, 'faring': 105, 'too': 310, 'bad': 32, 'either': 90, 'bob': 40, 'gunton': 129, 'coiled': 6
0, 'spring': 277, 'smarm': 269, 'norton': 206, 'james': 163, 'whitmore': 328, 'heart': 138, 'achingly': 4, 'great': 128, 'birdman': 39, 'clancy': 58, 'brown': 44, 'menacing': 181, 'antag
onist': 17, 'capt': 51, 'byron': 48, 'hadley': 130, 'william': 332, 'sadler': 250, 'amusing': 10, 'heywood': 140, 'mark': 178, 'rolston': 249, 'impressively': 150, 'vile': 318, 'bogs': 4
1, 'diamond': 76, 'then': 296, 'theres': 298, 'roger': 248, 'deakins': 70, 'lush': 175, 'cinematography': 57, 'camera': 49, 'gracefully': 127, 'glides': 123, 'offering': 211, 'ethereal':
 97, 'music': 194, 'pings': 225, 'conjunction': 64, 'emotional': 93, 'flow': 112, 'movie': 189, 'thomas': 303, 'newmans': 201, 'score': 255, 'mostly': 188, 'piano': 222, 'based': 36, 'do
vetailing': 81, 'neatly': 196, 'andys': 14, 'mind': 182, 'excellently': 100, 'selected': 256, 'soundtrack': 275, 'ranges': 235, 'likes': 171, 'hank': 131, 'williams': 333, 'gorgeous': 12
6, 'le': 168, 'nozze': 209, 'di': 75, 'figaro': 108, 'mozart': 191, 'if': 148, 'love': 173, 'lasts': 167, 'lifetime': 170, 'every': 99, 'viewing': 317, 'brings': 43, 'same': 252, 'array'
: 23, 'emotions': 94, 'anger': 15, 'revilement': 244, 'happiness': 132, 'sadness': 251, 'inspiration': 156, 'warmth': 321, 'can': 50, 'reduce': 240, 'most': 187, 'hardened': 133, 'misty'
: 183, 'eyed': 101, 'wonderment': 336, 'above': 2, 'else': 92, 'offers': 212, 'not': 207, 'just': 165, 'us': 315}
```


Term Frequency-Inverse Document Frequency Matrix:


```
[[0.02 0.03 0.02 0.02 0.02 0.02 0.02 0.06 0.03 0.02 0.02 0.06 0.26 0.06
  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.03 0.02 0.02 0.02 0.18 0.02 0.03
  0.02 0.02 0.02 0.02 0.02 0.02 0.03 0.02 0.02 0.02 0.05 0.02 0.02 0.02
  0.02 0.02 0.02 0.02 0.06 0.03 0.02 0.02 0.02 0.02 0.02 0.05 0.02 0.02
  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
  0.02 0.02 0.02 0.02 0.02 0.02 0.03 0.02 0.02 0.02 0.02 0.15 0.02 0.05
  0.02 0.16 0.02 0.02 0.02 0.05 0.03 0.03 0.02 0.02 0.02 0.02 0.02 0.03
  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
  0.02 0.03 0.02 0.02 0.05 0.02 0.03 0.03 0.02 0.02 0.02 0.15 0.02 0.02
  0.02 0.03 0.02 0.02 0.03 0.02 0.24 0.15 0.1  0.02 0.02 0.02 0.02 0.02
  0.02 0.05 0.02 0.02 0.02 0.03 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
  0.02 0.02 0.02 0.03 0.02 0.02 0.02 0.03 0.02 0.02 0.02 0.02 0.02 0.03
  0.02 0.02 0.02 0.05 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
  0.29 0.02 0.02 0.02 0.02 0.02 0.03 0.03 0.03 0.02 0.02 0.02 0.02 0.02
  0.02 0.02 0.02 0.02 0.02 0.06 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.03
  0.02 0.05 0.02 0.02 0.03 0.02 0.02 0.02 0.02 0.05 0.02 0.02 0.02 0.02
  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.03 0.02 0.02 0.13 0.02 0.02
  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
  0.03 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.19
  0.63 0.02 0.03 0.03 0.02 0.02 0.03 0.03 0.02 0.02 0.03 0.03 0.02 0.02
  0.02 0.21 0.03 0.02 0.03 0.02 0.03 0.02 0.02 0.02 0.02 0.03 0.02 0.02
  0.03 0.02 0.03 0.02 0.03 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.1  0.02
  0.02 0.03 0.02 0.02 0.03]]
```

It is no surprise that from the word frequency analysis it showed us how the shorter words we use to connect sentences like “the”(294), “us” (315) and “was” (322) appeared higher than more distinct words. What I did not anticipate, however, was the frequency with which words relating to the film appeared. I think because we’re looking at a review it would make sense that these words appear frequently but this was something I had not yet considered. For example, “Shawshank” appeared 263 times and redemption 239. I found it was interesting one fragment of the title would appear so much more but because of the nature of the film, I can see why; Shawshank is commonly referred to as a setting or location while “redemption” is not necessary to describing aspects of the film (the same way “Shawshank” describes the setting). The TF-IDF analysis yielded some interesting results as well, showing that when adding counterweight to adjust for the frequency of filler words like the aforementioned, the frequency of unique words in the text was much more evenly distributed. In other words, by adding weight to the term frequencies, it showed a much more even distribution to display how important each word is.

Reddit Results:

After analyzing the data that I discovered through mining Reddit with python, I attempted to compare and contrast differences between the stock market threads and the economy threads, knowing that the stock market hasn’t been a good indicator of how the overall economy was doing. I wanted to see if I could get a sense of the inverse relationship of the two through word frequencies in threads. However, based on the most common words used, there were not enough meaningful words to determine how people were feeling. However, from both outputs, I could see that people on Reddit are money whether its interest rates, loans, paychecks, or income. There are some of the keywords that appeared in both outputs, emphasizing the importance of government support for the economy and the stock market. Looking through economic submissions, I saw that people were talking about a wide variety of topics from Biden’s tax to Jeff Bezos giving employees more money and still being richer than he was before the pandemic. In the stock market submissions, a top thread was about US airlines not needing to be bailed out if they didn’t buy back their stocks and a controversial submission was related to the unpredictable nature of the stock market. Lastly, the sentiment analysis reaffirmed my belief that the stock market is doing unproportionately better than the overall economy. Approx. 23% of the submissions for the economy were found to have negative wording (-1 = negative sentiment | 1 =

positive sentiment) while approx.13% of submissions for the stock market appeared to have negative wording. Some examples of submissions and individual scores are shown above.

Frequent Words: Economy

```
The most common words are:
the      5873
and      5750
to       5225
people   4098
low      3575
a        3487
that     3484
it       3461
rates    3097
of       2735
for      2547
are      2404
interest         2397
will     2330
i        2314
be       2299
loans    2107
would    2077
is       1915
on       1878
income   1852
not      1792
but      1690
you      1607
this     1597
```

Frequent Words: Stock Market

```
no       110653
into     100612
job      98344
minimum          90120
there    86048
americans        86017
or       84624
don't    82690
with     81607
your     78658
for      68898
they     68648
i        66022
live     65545
paycheck         65536
born     65536
on       60140
capitalism       57856
people   55592
how      54389
living   53720
us       51551
out      50638
way      49325
understand       49162
buying   49154
system   49153
```

Economy Headlines:

```
['Slipgrid', 'PostNationalism', 'AutoModerator', 'n0ahbody', 'Tim_The_Enchanter', 'stuffed02', 'ShortIncident']
These are the top submissions!
Title of the submission:Jeff Bezos could give every Amazon employee $105,000 and still be as rich as he was before the pandemic. If that doesn't convince you we need a wealth tax, I'm not sure what will., User name: t3_itdsi3, Upvotes: 24512, Down votes it has:0, Overall score: 245
12, Number of comments: 2192.
Title of the submission:Trump will borrow whopping $4.4 trillion this year. That's $50,000 per family of four. But the same family will get only $1,200! Almost all the new debt goes to bailout corporations, banks & Wall Street, User name: t3_gd19br, Upvotes: 5093, Down votes it has
:0, Overall score: 5093, Number of comments: 423.
Title of the submission:More than 93% of U.S. college students say tuition should be lowered if classes are online, User name: t3_ildzb8, Upvotes: 3528, Down votes it has:0, Overall score: 3528, Number of comments: 260.
--------------------------------------------------
These are the new submissions!
Title of the submission:ニュージーランド: CBDCを発行する計画はないと明らかに, User name: t3_jeh1xb, Upvotes: 1, Down votes it has:0, Overall score: 1, Number of comments: 0.
Title of the submission:Biden's tax plan could create a tax rate of as much 62% for New Yorkers and Californians, studies show, User name: t3_jeg5or, Upvotes: 0, Down votes it has:0, Overall score: 0, Number of comments: 2.
Title of the submission:Flippy, the $30,000 automated robot fast-food cook, is now for sale with 'demand through the roof' — see how it grills burgers and fries onion rings, User name: t3_jeeuqg, Upvotes: 9, Down votes it has:0, Overall score: 9, Number of comments: 2.
--------------------------------------------------
These are the controversial submissions!
Title of the submission:It's Easy to Believe AOC Has an Economics Degree, User name: t3_f8wfx7, Upvotes: 2, Down votes it has:0, Overall score: 2, Number of comments: 42.
Title of the submission:ICELAND FORGIVES ENTIRE POPULATION ITS DEBT. TOTAL US MEDIA BLACKOUT, User name: t3_4gh9d4, Upvotes: 0, Down votes it has:0, Overall score: 0, Number of comments: 11.
Title of the submission:Older Americans would work longer if jobs were flexible, User name: t3_a3nin9, Upvotes: 9, Down votes it has:0, Overall score: 9, Number of comments: 40.
```

Stock Market Headlines:

```
['Stockclock-e', 'Fletch71011', 'bigbear0083', 'Sirbinkai', 'lykosen11', 'WittilyFun', 'r_Stockmarket_Bot', 'chiefkul', 'ghostofgbt', 'rStockMarket_Bot']
These are the top submissions!
Title of the submission:US airlines would not need the bailout if they didn't spend their recent enormous profits buying back stock. They could've set up emergency funds. They didn't. The money spent for the stock are gone. Now they are bailed out with taxpayers' money. Mismanageme
nt squared times recklessness = our loss., User name: t3_ful324, Upvotes: 4854, Down votes it has:0, Overall score: 4854, Number of comments: 633.
Title of the submission:Welcome to Black Monday of 2020! Market plunges by the most intraday since the 1987 Crash!, User name: t3_fj17jd, Upvotes: 4071, Down votes it has:0, Overall score: 4071, Number of comments: 408.
Title of the submission:Sen. Kelly Loeffler Dumped Millions in Stock After Coronavirus Briefing, User name: t3_flm21j, Upvotes: 3305, Down votes it has:0, Overall score: 3305, Number of comments: 364.
--------------------------------------------------
These are the new submissions!
Title of the submission:I have seen weird chart patterns before but this is the weirdest $ICSH, User name: t3_jefule, Upvotes: 1, Down votes it has:0, Overall score: 1, Number of comments: 0.
Title of the submission:What do you think about Colabor Group Inc? Im still pretty new in the stock maket so i's like to know if its a good buy., User name: t3_jeczdd, Upvotes: 1, Down votes it has:0, Overall score: 1, Number of comments: 0.
Title of the submission:Tips for a small-safe-long term investment, User name: t3_jecxzx, Upvotes: 1, Down votes it has:0, Overall score: 1, Number of comments: 1.
--------------------------------------------------
These are the controversial submissions!
Title of the submission:Isn't it obvious market going to crash again?, User name: t3_gu0f0y, Upvotes: 2, Down votes it has:0, Overall score: 2, Number of comments: 105.
Title of the submission:What really is happening in stock markets..., User name: t3_g0uo5i, Upvotes: 10, Down votes it has:0, Overall score: 10, Number of comments: 62.
Title of the submission:If you missed out on Weight Watchers rise you might be interested in Nutrisystem which has been growing earnings by 50% annually but still has a trailing PE of under 20., User name: t3_81cBih, Upvotes: 4, Down votes it has:0, Overall score: 4, Number of comm
ents: 26.
```

Sentiment Analysis: Economy:

```
 0    436
 1    256
-1    203
Name: label
```

```
{'compound': -0.3612,
 'headline': 'Pandemic expected to push poorer Americans out of banking system: regulator',
 'neg': 0.2,
 'neu': 0.8,
 'pos': 0.0},
```

Sentiment Analysis: Stock Market:

```
 0    455
 1    200
-1     97
Name: label
```

```
{'compound': -0.1027,
 'headline': '"Robinhood warns day traders to raise their cash buffers on \'widely-held stocks\' '
             'hours before market open "',
 'neg': 0.085,
 'neu': 0.915,
 'pos': 0.0},
```

4. Reflection

From a process perspective, we were able to discuss and allocate work in an efficient way. We planned for me to work on the Reddit analysis and Cameron to work on the IMDb analysis. That way, we would be exposed to more data sources and analysis. Throughout the assignment, we had constant communication and kept each other updated on the progress of our work. We conducted video calls every couple days to go over the codes and any changes we may have made. We worked well together, but it was apparent at times that we both were lost and didn't know what to do, given the amount of new concepts we were learning. Our project was appropriately scoped because we were able to lay out a game plan before we started the assignment and make sure that we were able to challenge ourselves and learn new things. Through trial and error and commenting code out to check if functions are running, we were able to unit test and ensure that our outputs were accurate.

In the end, given that there was a quick turnaround time for this assignment, we were able to learn on the fly. Although it was very challenging and the assignment had a huge learning curve, we definitely learned how to troubleshoot better and be creative. This assignment improved our understanding of foundational concepts that we learned in class and also exposed us to intriguing techniques like text mining and analysis, showing the endless capabilities of python. We wish

we could have had more practice with analyzing different data sources in class, as well as experience with creating visualizations using python. If we knew more about matplotlib/seaborn and its capabilities, our analysis would have been elevated. Overall, it was a good experience to learn and do the assignment simultaneously.