INFO 5709
Maia Petee

Language Frequency in Movie Synopses

Introduction

Movies are a cultural phenomenon that every individual in the developed world, whether they consider themselves a dedicated moviegoer or not, cannot help but be influenced by. Increasingly, over the last several decades, globally popular films net huge sums that extend far into the billions of dollars. Among these box office winners, however, are there distinguishable subtleties in box office prowess and other individual movie features, such as the language used by fans to describe the movie, or keywords chosen to represent the movie's content? In this study, we will pilot a data-driven approach to this question, merging several datasets with the purpose of addressing disparate aspects of film profitability and how it relates to the language used in film synopses and film genre.

Related Work

The individual concepts and tools that will serve as the theoretical underpinnings of this analysis are not particularly new or untried. Linguistically-based studies, particularly those that utilize online corpora, frequently use wordlists as a basis for linguistic analysis. In one of hundreds of such studies, educational researcher Dongyung Sun performed a contrastive analysis between two wordlists of undergraduate-level vocabulary (Sun, 2017) and a qualitative analysis addressing the philosophical aspects of each wordlist. Well-defined wordlists are considered to be a concrete representation of the lexicogrammar of certain social constructs, including professions (such as the Academic Word List (Cox, 2000)); wordlists can also perform as educational benchmarks in a wide range of subjects (Green & Lambert, 2018) and be used as a metric of psychological wellness or neurological evaluation (Simon, Zarzar, and Settar, 2012). They are also a useful proxy, when dealing with a widespread audience, for the linguistic identity of a subject or group; finally, let us not forget that wordlists are a critical tool for the application of quantitative metrics to language data, enabling it to be analyzed on an objective level and opening up new analytic horizons. Since wordlists have such critical value in representing a subject or interest on a lexical level, their fitness for the task is constantly assessed: new studies proposing minute improvements on existing wordlists (Green, 2019) are not uncommon.

Linguistic analyses that use word frequency as a primary metric are also quite common in corpus-based studies, chiefly because word frequency is another easily quantifiable metric while also being a usefully quantitative measure of a more qualitative datatype. Since word frequency is such a heavily-used metric in linguistic analyses, the criteria by which word frequency is evaluated are always under review, and deficiencies in word frequency as a metric continually being addressed (Brysbaert & New, 2009; Van Heuven *et al.,* 2014). The natural phenomena behind the distribution of lexemes across a language are also analyzed; all natural languages follow Zipf's law (Piantadosi, 2014), which states that words follow an empirical distribution in which the frequency of any given word is proportionate to the inverse of its rank in the frequency table: that is, a word of rank *n* has a frequency *1/n*.

To sum up, word frequency is an established method used in analyses of language data, and it has a robust body of work that has been dedicated to its refinement to legitimize it. In this analysis, we do not build on the immense amount of work that has been accomplished on these tasks, instead attempting to use these already-established principles to construct a new, simple metric for analyzing the language used in movie synopses.

Methods

In the search for an appropriate and relevant set of data, the Movie Synopsis Corpus from Kaggle stood out. This dataset, which includes many thousands of movie synopses (long descriptions of a film's plot) and includes genre tags as part of the data, could likely be used in many types of linguistic analyses, but a computational measure that is easy to calculate is word frequency: a ranking of how frequently each lexeme appears in the source language. Word frequency is usually an ordinal ranking in which words are ranked compared to one another. In general, high-frequency words are simpler and associated with an easier reading experience, while lower-frequency words are generally more obscure, whether due to their difficulty or because they are part of a specialized topic area.

Joined to an already-acquired and cleaned dataset of financial data relating to the top ~800 movies at the box office, the movie synopses comprised a dataset containing 719 rows. The combination of these particular data naturally suggested a research topic: could there be a connection between the frequencies of words found in the synopses and box office performance? Despite the fact that the dataset only contains 720 top-performing films, and not truly obscure films that might have the greatest chance of showing a strong effect, this still remains a question worth exploring.

Data cleaning was a lengthy process involving several steps. First, a way of measuring ordinal word frequencies in English was necessary. A list compiled from Google's trillion-word web corpus comprised over a quarter of a million words, well over the vocabulary of the English language (this exceedingly thorough list contains many possible misspellings of each word and numerous abbreviations to provide sufficient coverage for social media data). The frequency list was a good start, but filtering out simple "function words" — words that have no lexical meaning but serve important grammatical functions, such as "the," "of," or "in" — was also necessary. These words, referred to as stopwords in linguistic circles, have been the subject of many lists with an eye toward data cleaning; of these, the NLTK stopwords list was filtered from Google's English frequency list. The stopwords list consisted of only 127 words, and a vast 333,000 words still remained.

A critical preparatory task for this data analysis was deciding upon the boundaries of the word frequency groups: namely, the schema governing the assignment of frequency ratings. This step involved an immense amount of trial and error. The data table had 720 movies which needed to be divided by some metric into groups that were at least large enough to perform a data analysis upon (i.e. at least 100 movies in each of the two less-frequent groups, and the remainder in the higher-frequency group). However, because none of the target frequency groups was comprised of very low-frequency words, most of the synopses seemed to contain at least one word from each 50-word frequency list. From this problem emerged the task of grouping the filtering logic in such a way that the movies were more evenly distributed, and the frequency groups truly reflected the language contained in the synopses. A schema was chosen containing three frequency groups consisting of fifty words each, and to prevent too much overlap (i.e. to ensure that each group was lexically different from the other groups), the groups were spaced some distance from each other across the unigram frequency list. Group 1 began at word 500; Group 2 was located at 5,000, and Group 3 was at 10,000. Even words at rank 10,000 on a wordlist of length 333,000 are still not necessarily rare; proving this, words in group 3 included "redemption," "profound," "elevator," and "strawberry."

Pandas was used for data cleaning and manipulation. The synopses were joined to the top sales table using an inner join; in this analysis, we will only be looking at movies with synopses (and therefore frequency groups). Only one of the top 720 movies did not have an available synopsis and was filtered out; the final dataset consisted of 719 movies in total. The dataset was then run through a script that assigned frequency groups to the 719 movies: Group 1 consists of films whose synopses have only words from Group 1 and no words from Groups 2 or 3; Group 2 synopses have words from Groups 1 and 2 and

no words from Group 3, and Group 3 may contain words from all three groups. Groups 2 and 3 were not made mutually exclusive due to the fact that synopses from all three groups are likely to contain words from the much more frequent Group 1; if Group 3 were not allowed to contain words from Groups 1 and 2, it would be very difficult to write a script where Group 3 had any movies at all. This script assigned 241 movies to Group 1, 244 movies to Group 2, and 231 movies to Group 3; while this is not a perfect split, it should be sufficiently balanced for this analysis.

Additional data were compiled on the frequency words that were found in each synopsis. I concatenated all the frequency words that were found in each synopsis, and joined all the keyword cells as a string. As the analysis progressed, additional metrics relating to textual data were gathered: both average word length of each synopsis, and word and character count of each synopsis. With this cleaning done, the data were ready to be loaded into a visualization tool and explored.

Exploratory Data Analysis

Using a library called The WordCloud Python Project (API documentation can be read here), a frequency dictionary of all the words in my keyword string was generated. This dictionary was used to create a word cloud that reflects the distribution of key words in all movie synopses. Higher-frequency

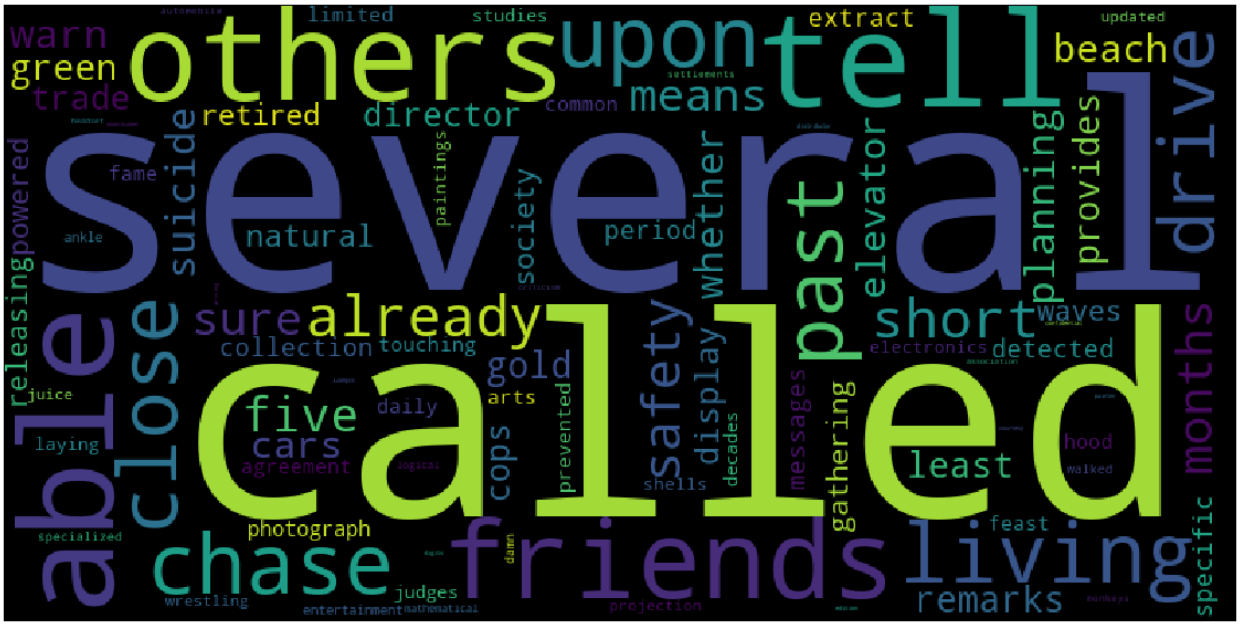

Figure 1: Word Cloud visualizing the frequency of keywords in the movie synopsis corpus

words are those that are generally not genre-specific, such as “several,” “called,” “able,” and “friends,” and could possibly be used to describe movies in any genre. Smaller words might represent common plot elements of specific genres, such as “feast,” “photograph” (thriller or detective film?), “society” (romance?), and “cops” (police procedural or crime drama).

There are also many non-textual categories in the dataset of which further exploration is warranted. In Figure 2 below, some common statistical metrics of the dataset are explored. We see that our dataset has 719 movies ranking from 2 to 790 on the top-grossing movies list (the mean of which is around 400, close to the center of the distribution; this indicates that we get a good sampling of movies in the first half of the list as well as the last half. The “Domestic Percentage” and “Overseas Percentage” columns represent the fraction of movie revenue that was made domestically (in the United States) and overseas, respectively. Through an examination of the quartile figures for both domestic and overseas revenue, we can see that the dataset skews slightly toward movies that made a majority of their revenue overseas (the 50% column shows movies that made 41.3 % of revenue domestically and 58.7% of revenue overseas); a previous analysis made clear that this is a function of the growing influence of Western cinema in the developed world, and the sheer number of overseas moviegoers. It is also quite possible, although not provable with this analysis, that moviegoers whose first language is not English might prefer movies with higher-frequency words in the synopsis. These elements of the data are visualized in Figure 3 below. While the small print on the chart axes may be difficult to read, the Overseas chart has a much larger scale, showing that revenue made overseas by some movies is greatly in excess of even the highest domestic outliers. Additionally, the interquartile range of the overseas distribution has a higher bound of more than $100 million greater than the domestic distribution.

	Rank	Worldwide	Domestic	Domestic Percentage	Overseas	Overseas Percentage	Year	Group
count	719.000000	719.000000	718.000000	719.000000	719.000000	719.000000	719.000000	719.000000
mean	400.527121	426.397218	172.131198	0.421975	254.530320	0.578025	2001.115438	1.990264
std	221.108587	272.184065	96.619054	0.124470	192.881069	0.124470	75.388016	0.813305
min	2.000000	200.300000	18.100000	0.000000	40.400000	0.172000	0.000000	1.000000
25%	211.500000	252.700000	110.100000	0.334000	140.700000	0.499000	1998.000000	1.000000
50%	404.000000	336.400000	145.050000	0.413000	200.500000	0.587000	2006.000000	2.000000
75%	587.000000	497.400000	209.375000	0.501000	300.350000	0.666000	2012.000000	3.000000
max	790.000000	2789.700000	760.500000	0.828000	2029.200000	1.000000	2019.000000	3.000000

Figure 2: A statistical description of dataset elements

A heatmap that displays the degree to which all of the datasets variables are correlated with each other can be seen in Figure 4 below. While the “Year” and “Group” rows are not terribly informative due to the ordinal nature of the data, a few interesting correlations do emerge. Domestic Percentage and Overseas Percentage do share a strongly negative correlation (seemingly close to -1),

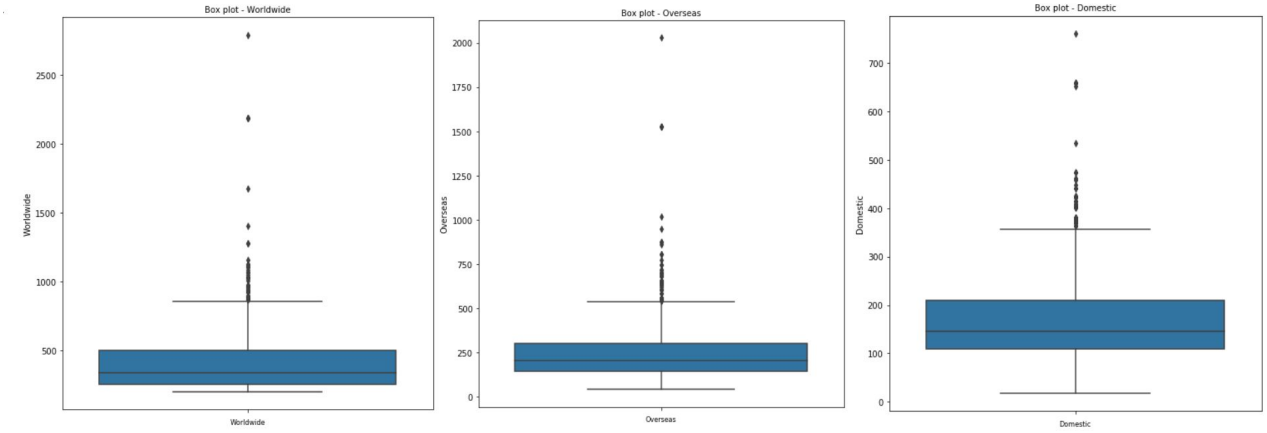

Figure 3 (from left): Box plots showing worldwide, overseas, and domestic revenue

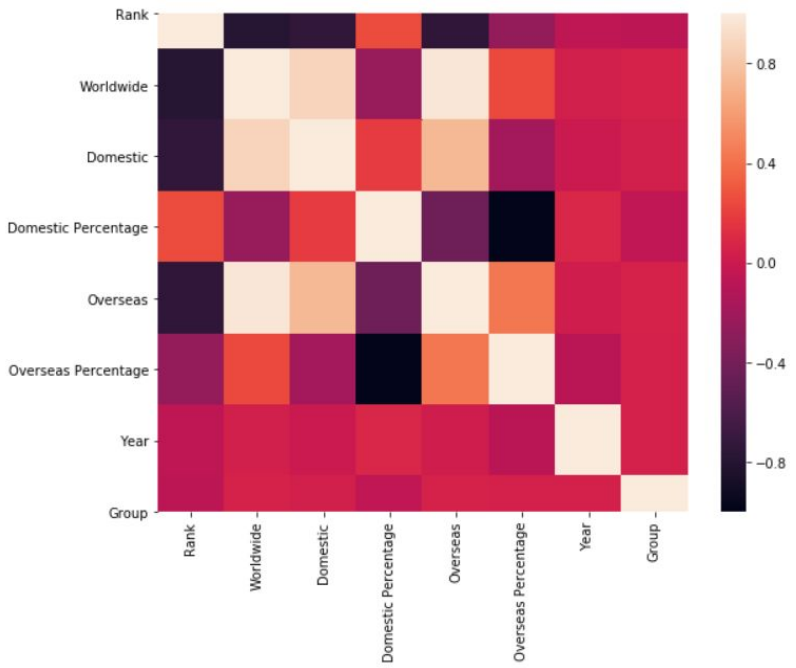

Figure 4: Correlation Heatmap of Dataset Variables

which makes sense when the necessarily inverse nature of this relationship is considered. While worldwide revenue and domestic revenue do have a high correlation, worldwide revenue and overseas revenue share a much higher correlation, indicating that overseas revenue tends to make up more of total revenue than does domestic revenue. What is telling is that percentage of revenue from domestic sales ("Domestic") is slightly positively correlated with rank, while percentage of revenue from overseas sales is slightly negatively correlated with rank. That is, where a movie has made more money domestically than overseas, it tends to be ranked more highly overall. This is not completely commensurate with the data in Figure 3 that shows more revenue coming from overseas sales. Seemingly, this data is more complex than anticipated. Lastly, both domestic and overseas revenue are strongly negatively correlated with rank (up to a straight -1 correlation); this is seemingly unintuitive until one realizes that "higher" rank numbers (780 as opposed to 1), which are in truth more lowly ranked movies, are associated with lower revenue figures.

To explore the linguistic aspect of this dataset further, two additional figures were calculated: average word length of the free text synopsis and word count of this synopsis. While this data will play a small role in this analysis, its calculation is primarily for the benefit of future work that will emphasize a focus on free text analysis by bringing in movie reviews and introducing a sentiment analysis aspect.

Finally, a qualitative exploration of the genre tags was undertaken. To do this, the WordCloud API was used to generate a tag cloud that automatically sizes its tags according to frequency. As can be seen in Figure 5 below, the highest-frequency tags are "violence," "murder," "flashback," "action," and "comedy." "Violence" and "murder" are likely tags that are correlated with the action genres, so it is not surprising to find these three in the top five most frequent tags. Among the least frequent tags are "blaxploitation," "avant-garde," "magical," "queer," and "grindhouse," but these tags only have 1-5 movies each in our dataset, so it might be statistically advisable to choose less frequent but still relatively common tags like "historical," "dramatic," "atmospheric," "melodrama," and "realism" for comparison to the hyper-frequent set. In addition to the figure below, three tag frequency groups were calculated numerically (the top thirty tags, divided into three groups of ten); these back up the intuition gleaned from this chart. This dataset, despite the low ordinal number of frequency groups, will hopefully be sufficient to display a correlative relationship between revenue and word frequency if such exists.

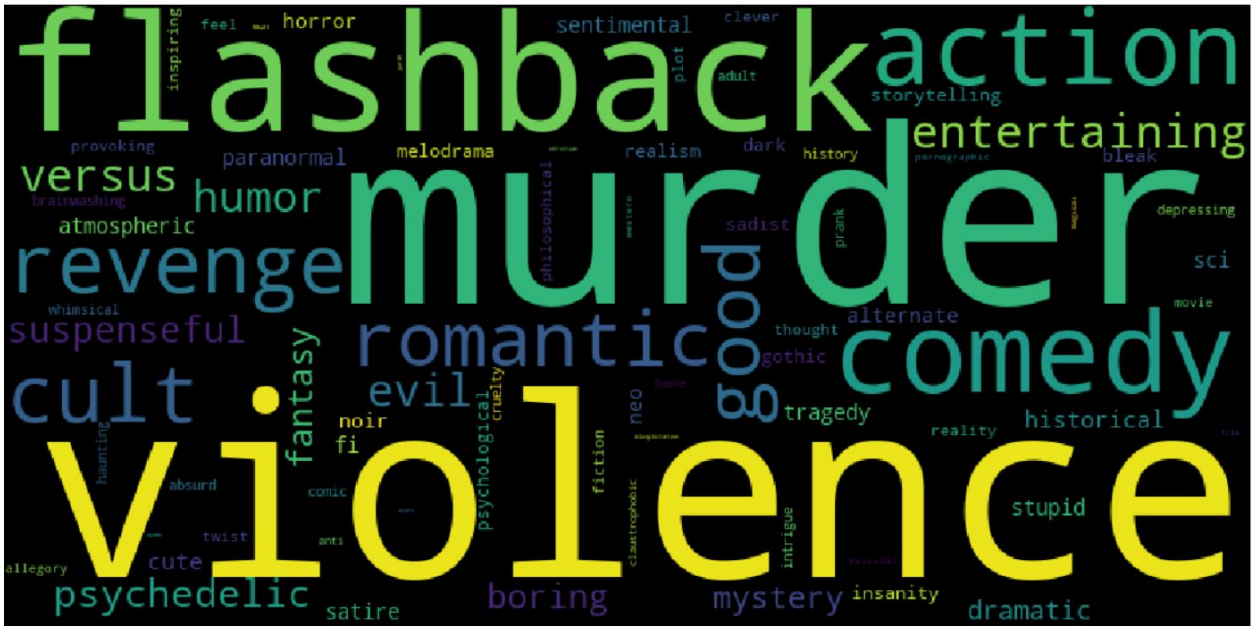

Figure 5: Word Cloud visualizing the frequency of genre tags in the Movie Synopses corpus

Hypotheses

Although our exploratory data analysis has shown some interesting (and perhaps contradictory) trends, we can still venture a few hypotheses about this dataset. First, that keyword frequency (expressed in our dataset as “group”) is positively correlated to box office performance: the higher the frequency group, the higher the average revenue of that group overall. Second, that movie genre (expressed in our dataset as “tags”) displays a relationship with box office performance: genres that are perceived to have more mass appeal such as action and comedy will show the highest average profits, while genres like cult and science fiction will experience overall lower profits. Finally, we hypothesize that movies for which the majority of revenue was made domestically will show higher prevalence in the lowest frequency group (group 3), while movies whose revenue was made primarily overseas will show higher prevalence in Group 1.

Results

In creating my visualizations, I wanted to explore interactivity across a range of tools. To this end, I set out to create interactive visualizations in both PowerBI and using Python packages for interactive visualization. While PowerBI is excellent at creating communicative and beautiful visualizations with effortless ease, creating a custom Python visualization is almost limitlessly flexible. The first visualization dashboard created using PowerBI explored aspects of the relationship between

frequency group and revenue. Its elements include a grouped bar chart that displays *average* (necessary because of the slight numerical imbalance across groups) worldwide, domestic, and overseas revenue across frequency groups.

Contrary to expectations, average revenue was highest in the third frequency group and lowest in the first. Additionally, box office rank was not positively correlated with average synopsis word length; that is, movies whose synopses had a lower average word length were not necessarily more profitable. This trend was stronger for some genres, while other genres contravened it and did in fact show a strong relationship between these two variables. These genres will be discussed later. In Figure 2 below, the viewer has the option to narrow the results by genre tag; Figure 2 displays results narrowed by "sci-fi" and a highlighted frequency group 3. In the top left, a box plot shows revenue for each frequency group divided into domestic, worldwide, and overseas revenue. The scatterplot in the lower left displays word length (x-axis) and box office rank (y-axis; points physically higher on the chart are of lower rank) and is currently set to show only sci-fi films in group 3. Finally, the pie chart at the right displays percentage of revenue by group based on the filter tags. One can see that in science fiction, group 1 films make the most money on average (~38%), with group 3 coming in second at 33%.

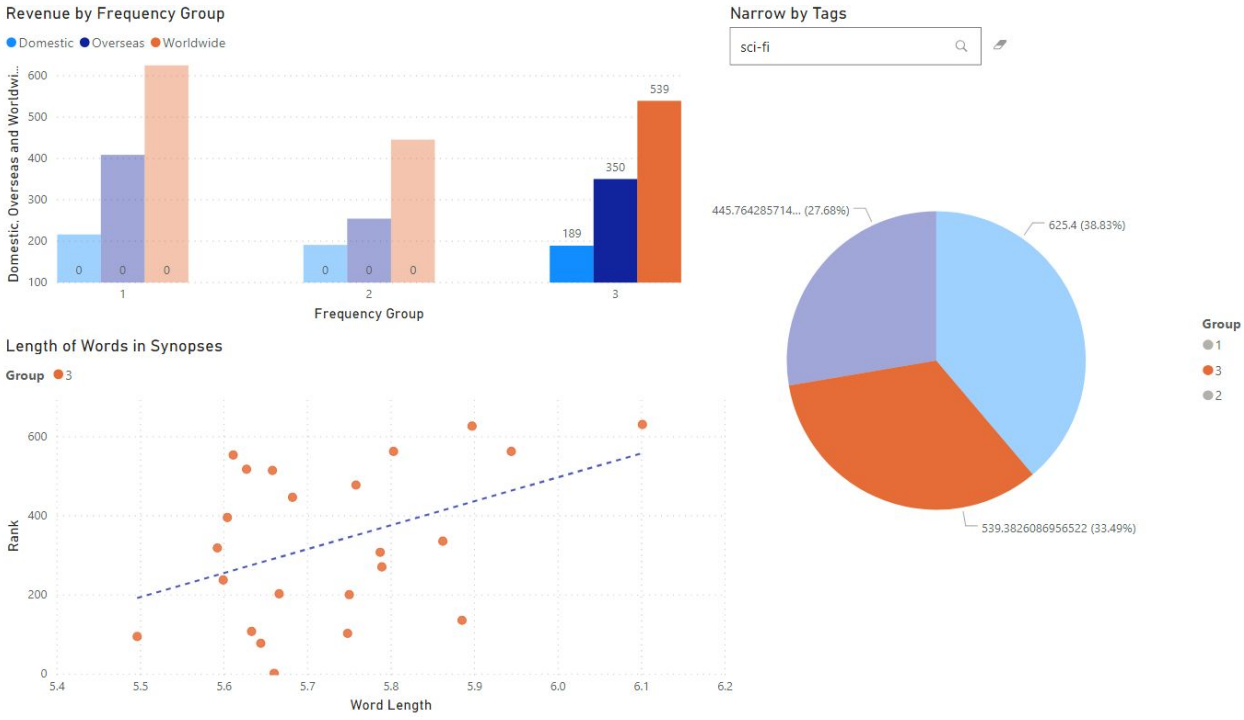

Figure 6: PowerBI visualization highlighting several linguistic traits of synopses, including frequency group and word length.

Figure 7 below, created using matplotlib, shows the distribution of worldwide box office revenue across frequency groups. It seems that in the lowest range, i.e. movies that made up to 200 million dollars at the box office, frequency group 1 takes precedence. Similarly, movies in frequency group 3 seem to predominate in a slightly higher revenue group (~ 500 million to 1 billion). Incongruously, group 2 has some of the highest-grossing movies, with some of these films making over 2 billion dollars worldwide. At any rate, the clear relationship between frequency group and revenue that was hoped for is not at all apparent; in many areas, in fact, it is contraindicated. It seems that there is little reason to believe that frequency group as operationalized in this study (the definition of frequency group and the effects this definition has on the results is a valuable topic for discussion), is related to box office revenue; there are several more subtle relevancies which will be explored in the next section.

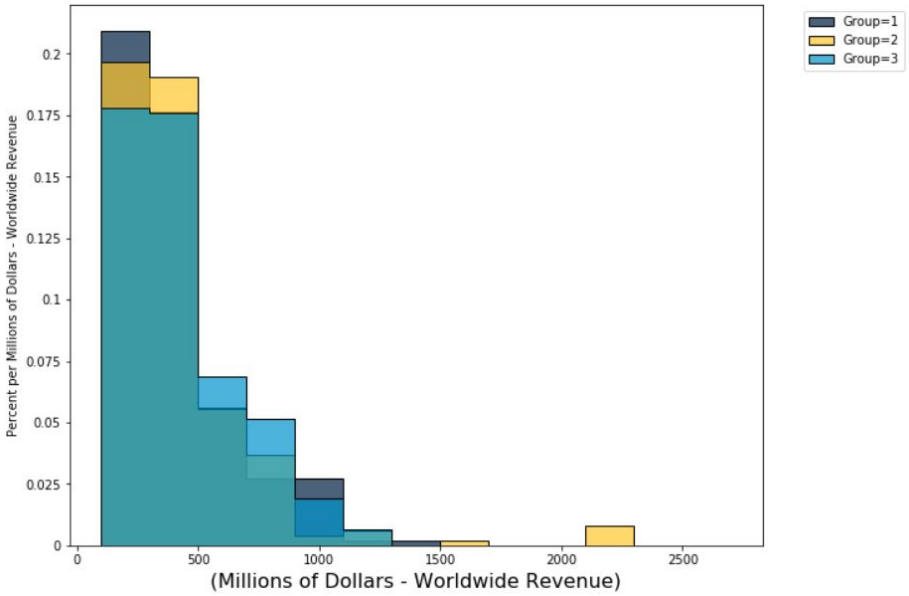

Figure 7: matplotlib histogram showing revenue across frequency groups

To address my third hypothesis: namely, whether high-frequency tags (such as “violence,” “murder,” etc.) were more successful at the box office than low-frequency tags, in addition to how these values were distributed across groups, an interactive Python dashboard was created with the help of the matplotlib and ipywidgets libraries. A function was written that interactively graphs revenue and group data for a selected genre tag, displaying both domestic and overseas revenue for each of the three frequency groups and enabling the exploration of a number of hypothetical factors by genre tag. Below, collected in Figure 8, is a sampling of the plots that are possible with this interactive dataset. The

displayed selection focuses on less-frequent tags, such as “philosophical,” “dark,” “sci-fi,” and “thought-provoking.”

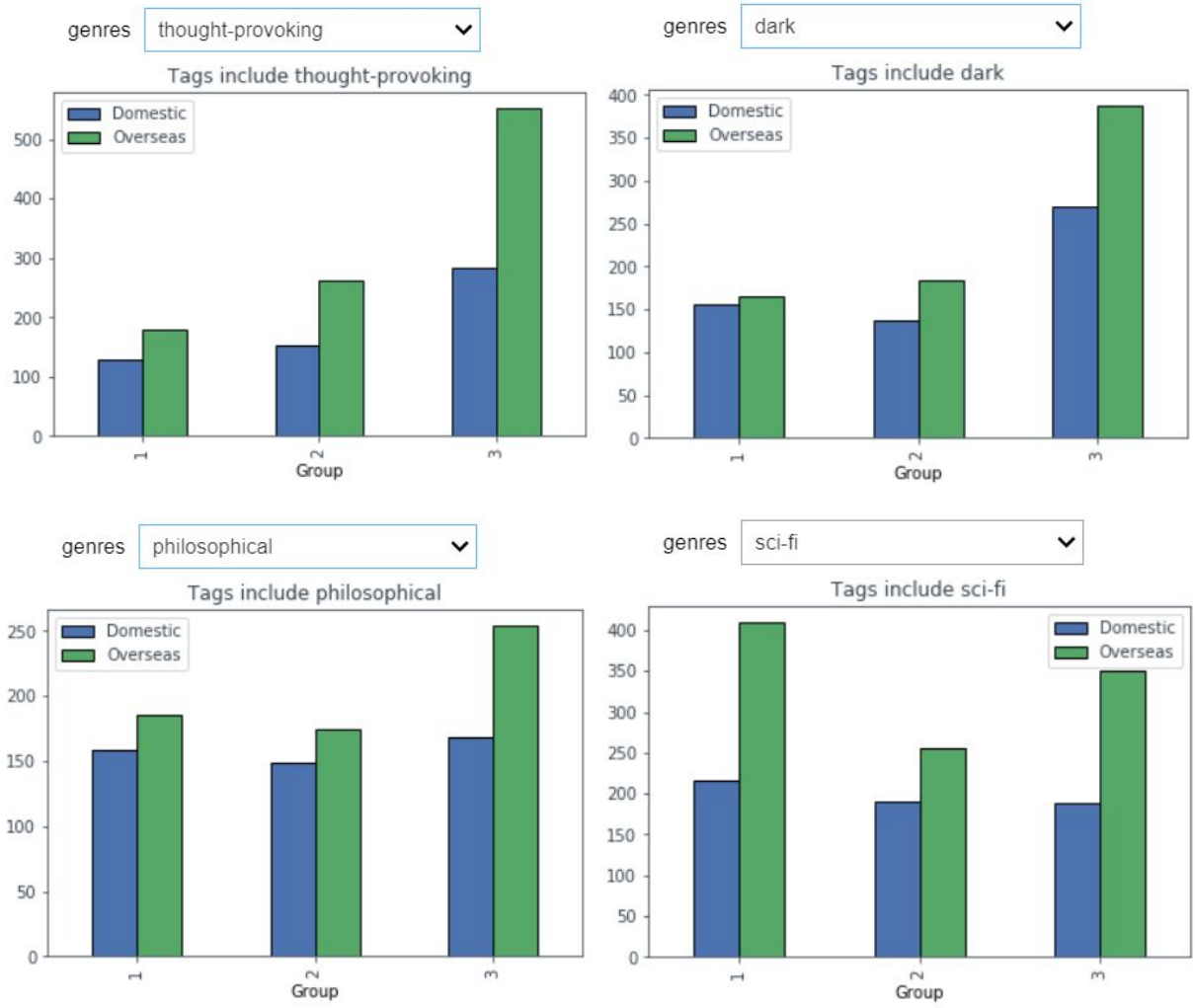

Figure 8: Four samples from the interactive genre tag dashboard, focusing on less-frequent tags

While these genres do show higher revenue figures across group 3, the lowest-frequency group, it does not seem as though the hypothesis that more obscure tags would be more popular *both* in lower frequency groups and domestically is borne out. There are, once again, subtler effects that we see when examining revenue both overseas and domestically across tag groups, but not a general correlation.

Discussion

Out of the three initial alternative hypotheses, not one of them was found to be true. The blanket relationships that were expected were not present, and in some cases the data seem to

contradict the hypotheses directly. The first hypothesis, which stated that box office revenue would be higher on average for films in the first frequency group (group 1) than in the second or third groups, was proven false: there was actually a slight inverse relationship, as films in group 1 made 407 million dollars on average worldwide, whereas for films in group 3, this figure was 444 million. Group 2 films fell in the middle at 429 million dollars. It could be possible that genre plays a larger role in this relationship than initially anticipated (e.g. that the hoped-for correlation would indeed be present with more popular genres); there are some intriguing relationships when certain genres are examined individually. More-frequent genres, largely considered to be those with the highest mass appeal, do not display the

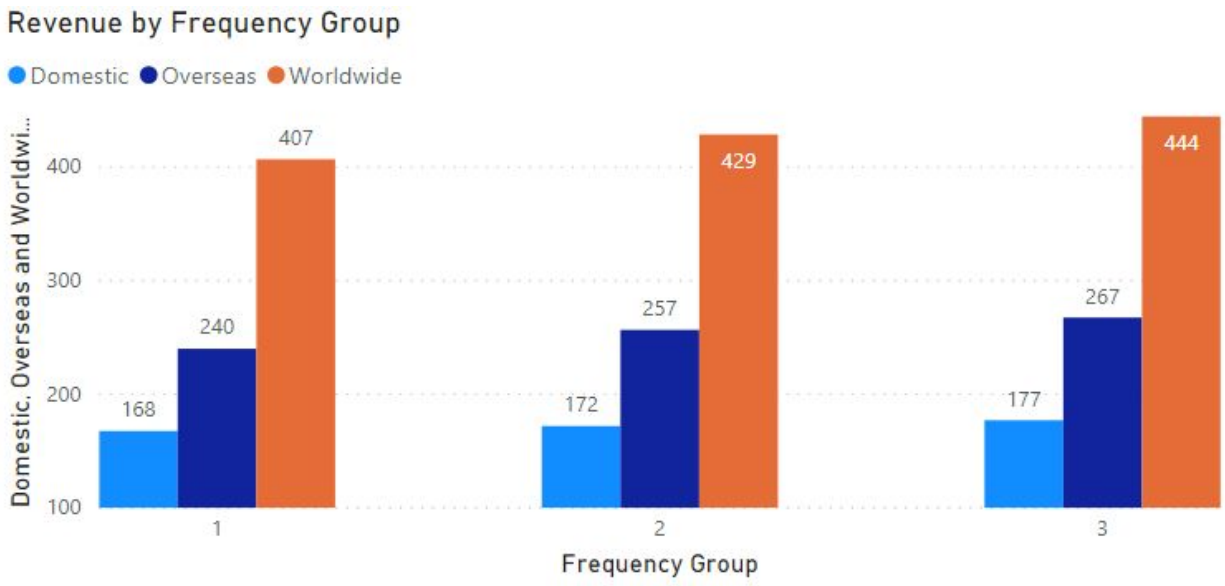

Figure 9: Evidence contravening the positive relationship between frequency group and revenue

same inverse relationship seen in Figure 9 above. Group 1 and group 3 action films, for example, made almost identical amounts on average, and comedy films decreased in box office popularity as their word frequency declined. However, this is not the case for all the genres in the list of ten most frequent tags; films tagged “violence” and “murder” did not display the same trends as their popular brethren. There are also intriguing trends to be discovered in less-frequent genre tags: “philosophical” films in group 3 were over $100 million more popular on average than films with the same tag in either Group 1 or Group 2, and the same trend holds true for films tagged “psychedelic.” Yet other genres in the same tag frequency group display different trends: sci-fi films, for example, seem to be wildly more lucrative in Group 1 than in either other group (this singular result is likely due to the blockbuster Star Wars films being tagged as sci-fi, and quite possibly does not extend to more cerebral exemplars of the genre).

While there are some clear trends, they do not seem to be able to be isolated solely by examining the frequency of genre tags; in addition, the dataset is small enough (~250 films in each word frequency group) that one extremely popular film in any genre could quite possibly skew the results. However, collecting a language-based dataset larger than this could be difficult; while the original movie synopsis dataset contained far more movies than the 720 we work with in this analysis, all films without both synopses and available financial data were filtered out. Genre, when not examining profits with respect to word frequency groups, could be a more telling indicator of financial success, but a cursory investigation into this correspondence does not bear this intuition out. Box office averages, not sums, were considered, due to the lack of balance in the number of films in the dataset tagged with a particular tag. While the *sum* of profits made by films tagged with "violence," "action," or "murder" is well over twice that of films tagged as "sci-fi" or "philosophical," those trends do not persist when considering averages, which indicates simply that there are *more* films in the top 800 that qualify for these tags, not that any one of these films is especially profitable. Quite to the contrary, some more "popular" genres made less money on average than did "less-popular" genres.

Our third and final hypothesis postulated that movies for which the majority of revenue was made domestically would show higher prevalence in the third frequency group, while films that predominated overseas would show higher prevalence in the first frequency group. As it turned out, across all frequency groups, and regardless of frequency group, films made more revenue overseas than domestically. This is likely due to undoubtedly higher numbers of ticket sales, given the much higher overseas population. Despite this ubiquitous trend, there were some genres and groups for which this gap was narrowed, and very occasionally disappeared; one hard-to-example was for historical films in frequency group 1, and again for blaxploitation films in frequency group 1 (these last actually made much more money domestically, perhaps because they capitalize on a singularly American cultural subgroup). While there are individual data points that beg further exploration, a trend that supports the hypothesis is not evident. Statistically, it would be worthwhile to control for the population difference between the domestic and overseas groups, but the data necessary to do so is not available.

Future Work

The specific and tentative relationships discovered in this analysis could be clarified and expanded upon by creating a far more robust measurement of word frequency. Defining frequency

groups and assigning the films in the available dataset to one of these frequency groups was a difficult and imperfect process: since the synopses were long, in many cases hovering at 1000 words, larger (and therefore more comprehensive) slices of the word frequency list would have resulted in positives for all the films in the dataset, making a division impossible. This is in fact what happened in the first several iterations of the frequency group assignment. I believe that, should the word frequency connection continue to be pursued, a much more complex and comprehensive metric needs to be developed to measure the *overall* frequency of the words in a synopsis, not merely the presence or absence of a specific and limited set of keywords. Such a metric would need to be mathematically derived and proven to bear a strong relationship to the reality of word frequency; it is possible that such a salient measurement already exists in the literature. Additionally, such a relatively complex frequency calculation would have the benefit of being distributed continuously rather than ordinally, and would therefore enable a wider and more salient range of analyses.

Additionally, the linguistic connections made here could be greatly expanded upon by the addition of another free text measure — possibly movie reviews — and analysis of such, including a sentiment analysis that explores the relationship between review text and box office performance. An analysis of language data that are likely to show a close quantitative relationship (e.g. if reviews of a movie are favorable, it would be likely to perform well at the box office) would hopefully shed greater light upon the connection between the language around films and the success of those films.

References

Brysbaert, M., and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41:4, 977-990

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly; 34,* 213 - 238. DOI: 10.2307/3587951

English Word Frequency. (2017, September 6). Retrieved November 2, 2019, from https://www.kaggle.com/rtatman/english-word-frequency

Green, C., and Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects. *Journal of English for Academic Purposes, 35*, 105-115.

Green, C. (2019). Enriching the academic wordlist and Secondary Vocabulary Lists with lexicogrammar: Toward a pattern grammar of academic vocabulary. *System*, *87:* 188-198.

MPST: Movie Plot Synopses with Tags. (2019, April 16). Retrieved November 2, 2019, from https://www.kaggle.com/cryptexcode/mpst-movie-plot-synopses-with-tags

NLTK's List of English Stopwords. (2010, August 27). Retrieved November 2, 2019, from https://gist.github.com/sebleier/554280
Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*; 21:5, 1112-30.
Simon, M., Zarzar, K., and Settar, C. (2012). PRM118 Methodology for Neuropsychological Assessment Wordlist Adaptation. *Value in Health, 15,* A842.
Sun, D. (2017). A Contrastive Analysis between English Vocabulary Profile and College English Wordlist. *Theory and Practice in Language Studies*; 7, 729-736. DOI:10.17507/tpls.0709.04
Van Heuven, W., Mandera, P., *et al.* (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67:6, 1176-1190.
Wordcloud. (latest version 2019, May 7). Retrieved November 2, 2019, from https://amueller.github.io/word_cloud/index.html