Text Analysis:
A Star Wars Story

Greg Eastman

Introduction

Unsupervised text analysis is an area of research that has been applied to presidents[1], news writers[2], and even the bible[3], but I will instead look at a purely cultural piece, Star Wars. In 1977 the first movie came out, and to this day the films and television shows are still topping the charts. No other fictional universe has ever stayed relevant for so long on the silver screen. To explore this, I got copies of the screenplays, then analyzed them.

To understand what makes each movie a hit, I applied several natural language processing strategies to all the films. First, I look at lexical diversity to look at the prose level. I find that the movies do not use complex language, which may help them connect with a wider audience. Then I analyze the most frequent words, to understand the type of writing and topics of the series. It appears that they focus on visual setting and the world building. After that I use K-means grouping to cluster the films into three sections. Interestingly it appears that the films do not group into the distinct trilogies. Finally, I analyze the sentiment of each cluster to see how the tone of the series changes. With that covered I will review some previous literature wielding these techniques.

Literature Review

Previous researchers have used R to perform a variety of different strategies to understand texts and their authors. Some techniques are frequent word analysis, grouping, and topic analysis. All of which were used in *Unsupervised Learning of Two Bible Books: Proverbs and Psalms* by Wei Hu[3], to investigate the similarities of different books and chapters within the bible. It managed to glean a new perspective on the types of messages. Frequent words can help to understand the over-arching message of each section. Grouping helps to see how the various books interact with each other. These can all be applied to the Star Wars films in a similar way that they were to the bible. Topic analysis reveals the main points of psalms, but it's not as useful for the films. Unfortunately, the large number of made-up terms confuses the algorithm. Thankfully, there is another method that can help to get new insights.

Another strategy that has grown in popularity recently is lexical diversity, which was used to check the authors of Reuters news articles[2]. This helped to analyze the level at which the articles are written. This can help to understand both the audience and the types of topics a body of work or news is aimed at. I am employing these strategies to answer two distinct questions about the Star Wars screen plays. What are the common elements that give them lasting power, and when you group the films how do they differ?

Methodology

In this section I will outline where I am getting the data, and what tools I plan on using. The data will be the transcripts of the screenplays for each film[4]. The licensing for this material allows its use if it is not for profit and proper citations are given[5]. The data is a single text file per

movie script. Each file needed to be cleaned rather extensively. To do this it is important to think about what information matters and what does not. To understand a film, we want to keep text about stage directions, dialogue, and descriptions of scenes. What we do not want are unimportant words, punctuation, numbers, and names. Since every time a character talks, it references the person, so the names bloat the file. Additionally, the films are not cultural hits because of the character names. So, I addressed all the excess information by cleaning it out.

When scrubbing the data, I used a step-by-step process, where order matters, to obtain the final "clean" result. The first step was to make all character's lowercase so that the start of a sentence does not mess up the string matching. The next step was to remove names. Some of the names in Star Wars have numbers and other characters in them so its useful to do it here that way they match up. The database for this came from dplyr's starwars dataset. Additionally, I checked over the data to see if any names were missed, and I added them to the name list and removed them. Next, I got rid of stopwords, or terms that add no meaning. Then removed numbers, white space, and miscoded characters. Finally, I stemmed the data. Stemming is a process of cutting off parts of words so that different tenses and forms of one word all look the same. This is an automated process that comes from the tm[6] package. After all of that I remove sparse terms to get rid of one-off terms that do not really add meaning to the overall corpus. Once all of this was done the text was ready to be analyzed.

Analysis

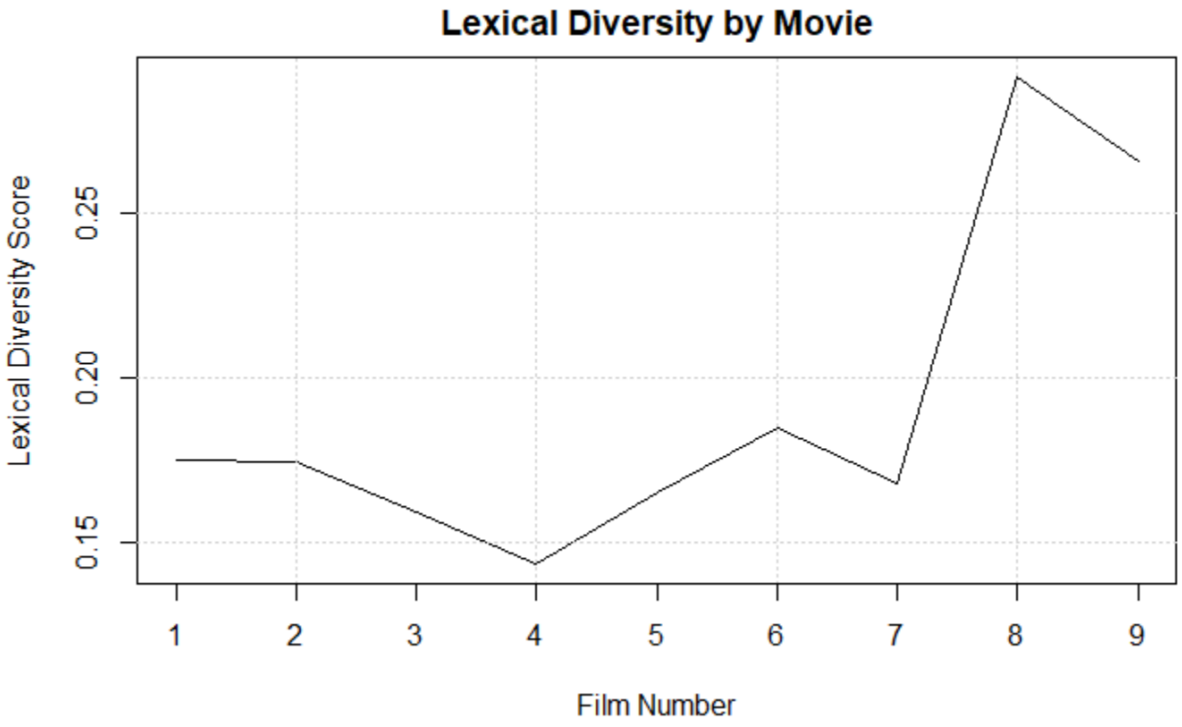
I employed several different strategies to look at the Star Wars transcripts, but first I inspected some of the major themes and ideas present every movie. All the transcripts were

combined into a single corpus. I then used the wordcloud2 package to make a visual of the most common words in the films. The result shown on the next page.

The most common terms are the biggest, and we see that there is a pattern. The largest words are “jedi”, “cockpit”, “droid”, “look”, “stop”, “go”, and “fighter”. These terms are mostly things in the Star Wars universe, like droids and jedi. The next major term category are locations like cockpit and fighter. The latter refers to a spaceship (aka star fighters). The third category is actions like “look”, “stop” and “go”. The final group are terms used in dialogue, like “power” or “death”. They exist in the cloud but are smaller than the others. This has several different implications to the series overall. First, the lore and setting of Star Wars are the most referred to,

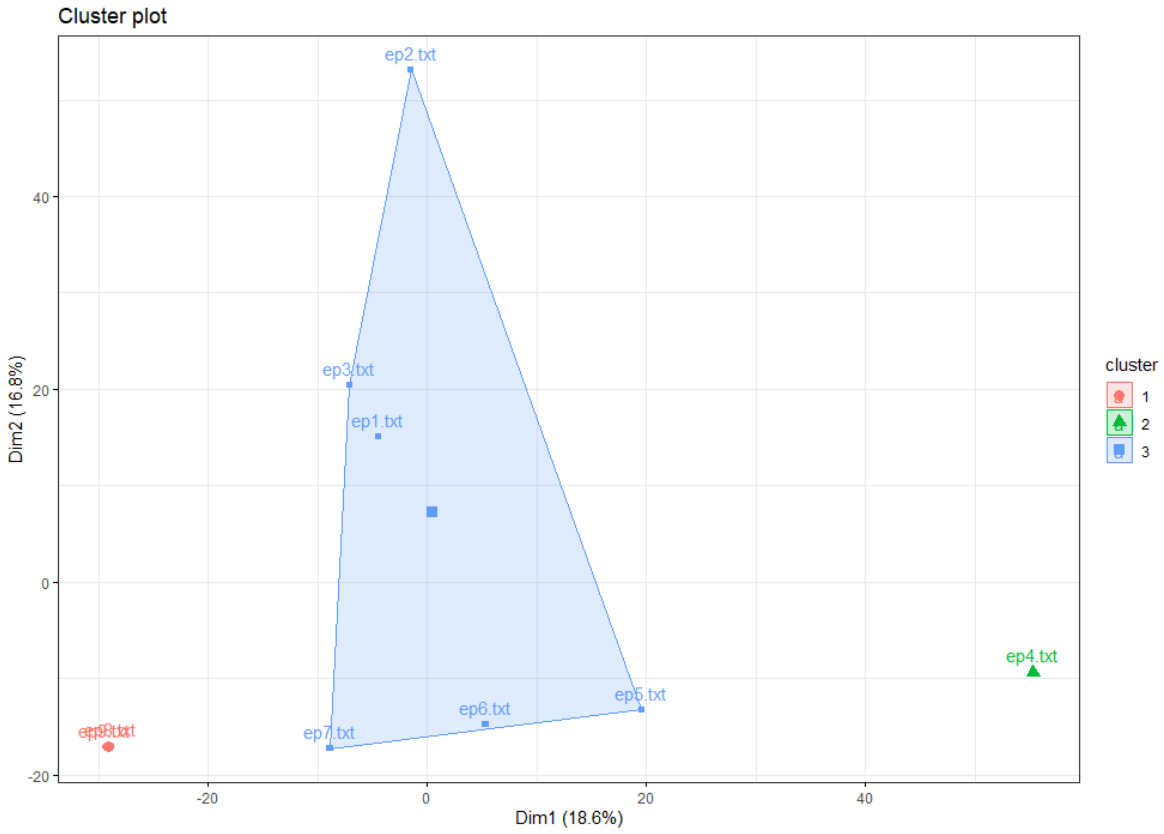
and important parts of the films. This makes sense as the iconic cantina scene in episode 4 is all about the world created by George Lucas. Next the visual terms are most common. Which implies a heavy emphasis on the film's effects and appearance. The action terms being frequent show how important that part of the films are to both the writers and audience They are, after all, science fiction **action**-adventure movies. Now we get to the most interesting implication of the word cloud. There is comparatively little focus put on the dialogue in the films. Were the contrary true we would see more and bigger words that display a sense of emotion, meaning, or feel, but we do not. Therefore, the films all seem to be heavily founded in the atmosphere of the universe George Lucas made, and then about the action of scenes. Now that we know what the focus is, its time to move onto to their language.

Lexical diversity is a measure of the complexity, or language level, of the writing. A higher score indicates that a text is more complex. Texts with high linguistic diversity, like academic news articles, will be close to 1. The below chart shows the LD scores of each film.



None of the language is complex. We can see that the y-axis never goes above .3. This may help to make the movies more approachable to all audiences, since the prose should not be confusing. Also, it is important to look at how each film compares to the others. The last two are the most complex, while the others are significantly less so. Episode 4 is the least lexically diverse by a significant amount. This does not mean that the story or terms of the later films are more "intellectual", but it does imply that a more challenging prose was established. Moving on, George Lucas wrote 1-6, and the 7th movie was intentionally very similar in style to its predecessors. This shows in the lexical diversity chart where Lucas's films use incredibly similar and approachable language. This also implies that the films may not nicely group into three bins, one for each series. To explore this possibility, I used K-means to cluster the movies.

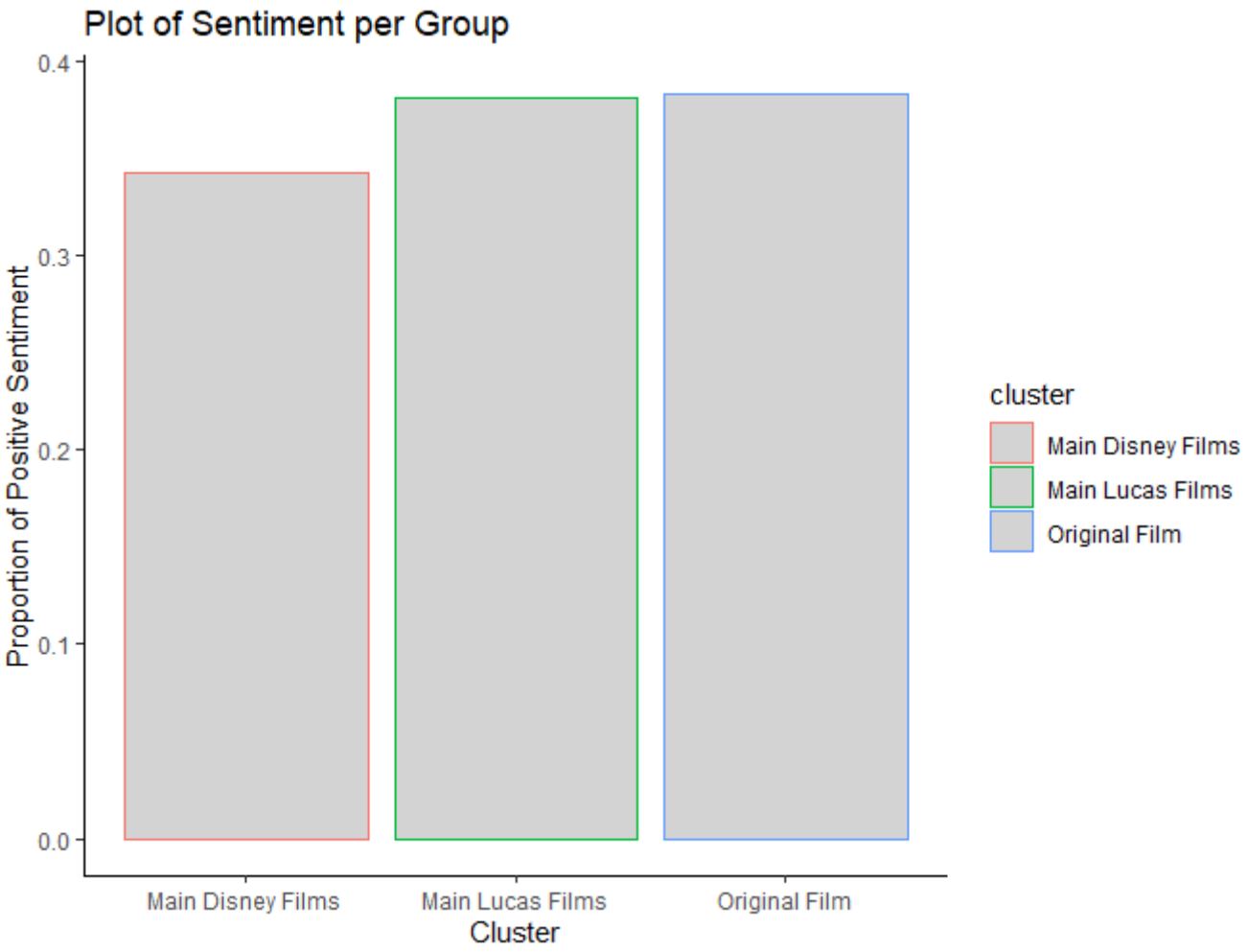
I grouped the films into three sections using Euclidian distance. I found that episodes 8 and 9 were in one cluster, episode 4 was in its own cluster, and the rest formed the third. To visualize this, I used the factoextra[16] package to make the plot below.



We can see the clusters that were described, but this graphic shows additional patterns. First, episodes eight and nine are very similar. They are not identical but are not much different. These were written by Rian Johnson and J.J. Abrams, they also were not made stylistically like the originals. They were more subversive and action oriented. Then we note that three clusters may not be a natural grouping. It looks like episodes 5, 6, and 7 are all similar. This makes sense as the former two were made nearly at the same time under the same circumstances and written by George Lucas. Additionally, episode 7 was made to be stylistically very similar to the original

series, and it seems that attempt was successful. Then the prequels seem to be in their own group. They were all made by Lucas, and there was a lot of funding and excitement for them. Also, special effects were becoming a more forefront part of film making. So, their being similar makes sense. Meanwhile, episode 4 is very different from the rest of the movies. This makes sense, as George Lucas wrote this, but it is the least action and space magic using of the movies. It was also the first one, before the series got more fame and funding. That makes it different, as well as its language was the least complex. Now, to further explore these groups I perform sentiment analysis on each cluster and compare them.

The sentiment analysis I used was a dictionary search using bing encoding for positive and negative terms. Although it is a simple binary it is useful for exploring the overall tone of the films. Additionally, each of the movies explores the basic binary of good vs evil. Therefore, the analysis should reflect this major theme of the series. Below is a plot that was made to look at the proportion of good sentiment in each cluster.



We can see that all the films all share a similar sentiment. They are generally negative, with roughly only 1/3 of the emotional terms in the movies being positive. As to each group, episode 4 was the brightest, and episodes 8 and 9 were the darkest. These films were very disparate in the clustering, so the tone is likely a factor as to why. Although, the magnitude of the difference is rather minimal, less than half of a percent. This implies that the tone of good and bad, and the language of the characters is overall consistent across the groups. To explore what the findings mean, it is time to stop analyzing and start synthesizing.

Conclusion

The Star Wars films all have similar threads that make them what they are, like the level of the language used, the focus, and the sentiment. The writing is simple, which likely helps it be successful for two reasons. It makes them more understandable to all audiences and helps the characters be approachable. The more recent films show a higher prose complexity, but the language is still not complicated. Additionally, all the movies had a focus on the world and action. This may help it be a cinematic experience for the audience that is unique. While many films take place on earth, or in other established settings, Star Wars creates an atmosphere all to its own. This is demonstrated by the terms in the world cloud which focus on the universe Lucas created and the action therein. Additionally, a large part of the atmosphere of a film involves the tone; which like the focus is consistent throughout the series. The movies are generally negative, with over 2/3 of the terms denoting something bad. Although, the films are all unique in their own way, to explore their differences they were clustered and compared.

When grouping the films, it was expected that they would form into three clusters, one for each trilogy; this was not the case. The original Star Wars movie, episode 4, was on its own. This is probably because the budget and reputation increased for future films. The next six consecutive movies were all grouped into one cluster. They all relied on the universe, design, and writing of George Lucas, but they had slightly more complex language, more special effects, more budget, more fame, and a darker focus. The last cluster contained the final two films. These deviated more from the style and themes of the others. This is demonstrated by the notable increase in lexical diversity and stronger emphasis on action. Altogether the Star Wars films grew

in the language complexity over time, and the tone got slightly darker. Along with the emphasis on world building and action these traits are likely what make Star Wars the long-lasting successful franchise it has become.

References

1. Clarke, Isobelle, and Jack Grieve. “Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted between 2009 and 2018.” *Plos One*, vol. 14, no. 9, 2019, doi:10.1371/journal.pone.0222062.
2. Stamatatos, Efstathios. “Authorship Attribution Using Text Distortion.” *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, doi:10.18653/v1/e17-1107.
3. Hu, Wei. “Unsupervised Learning of Two Bible Books: Proverbs and Psalms.” *Sociology Mind*, vol. 02, no. 03, 2012, pp. 325–334., doi:10.4236/sm.2012.23043.
4. “Star Wars Transcripts.” *Transcripts Wiki*, transcripts.fandom.com/wiki/Category:Star_Wars_transcripts.
5. “Copyrights.” *Wookieepedia*, starwars.fandom.com/wiki/Wookieepedia:Copyrights.
6. Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5): 1-54. URL: https://www.jstatsoft.org/v25/i05/.
7. Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.
8. Emil Hvitfeldt (2020). textdata: Download and Load Various Text Datasets. R package version 0.4.1. https://CRAN.R-project.org/package=textdata
9. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
10. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” _Journal of Open Source Software_, *3*(30), 774. doi: 10.21105/joss.00774 (URL: https://doi.org/10.21105/joss.00774), <URL: https://quanteda.io>.
11. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
12. Dawei Lang and Guan-tin Chien (2018). wordcloud2: Create Word Cloud by 'htmlwidget'.R package version 0.2.1. https://CRAN.R-project.org/package=wordcloud2
13. Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr
14. Grün B, Hornik K (2011). “topicmodels: An R Package for Fitting Topic Models.” _Journal of Statistical Software_, *40*(13), 1-30. doi: 10.18637/jss.v040.i13 (URL: https://doi.org/10.18637/jss.v040.i13).
15. Hadley Wickham and Jim Hester (2020). readr: Read Rectangular Text Data. R package version 1.4.0. https://CRAN.R-project.org/package=readr
16. Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra