Text Analysis: A Star Wars Story

Written and Directed by
Greg Eastman

Cultural Phenomenon

- Despite starting a long long time ago [1977] this movie franchise is still massive.
- There are three major installments, each came out in a different era.
- Lets learn about it
  - What makes it so culturally powerful?
  - Does each trilogy group next to its neighbors?
  - What makes each group of movies different?

Data

- What is the ideal data?
  - Must reflect the characters.
  - Must reflect the setting/world.
  - Should not be bloated with character names.
- I got copies of the Star Wars transcripts.
- I got a dataset of character and location names.

Cleaning

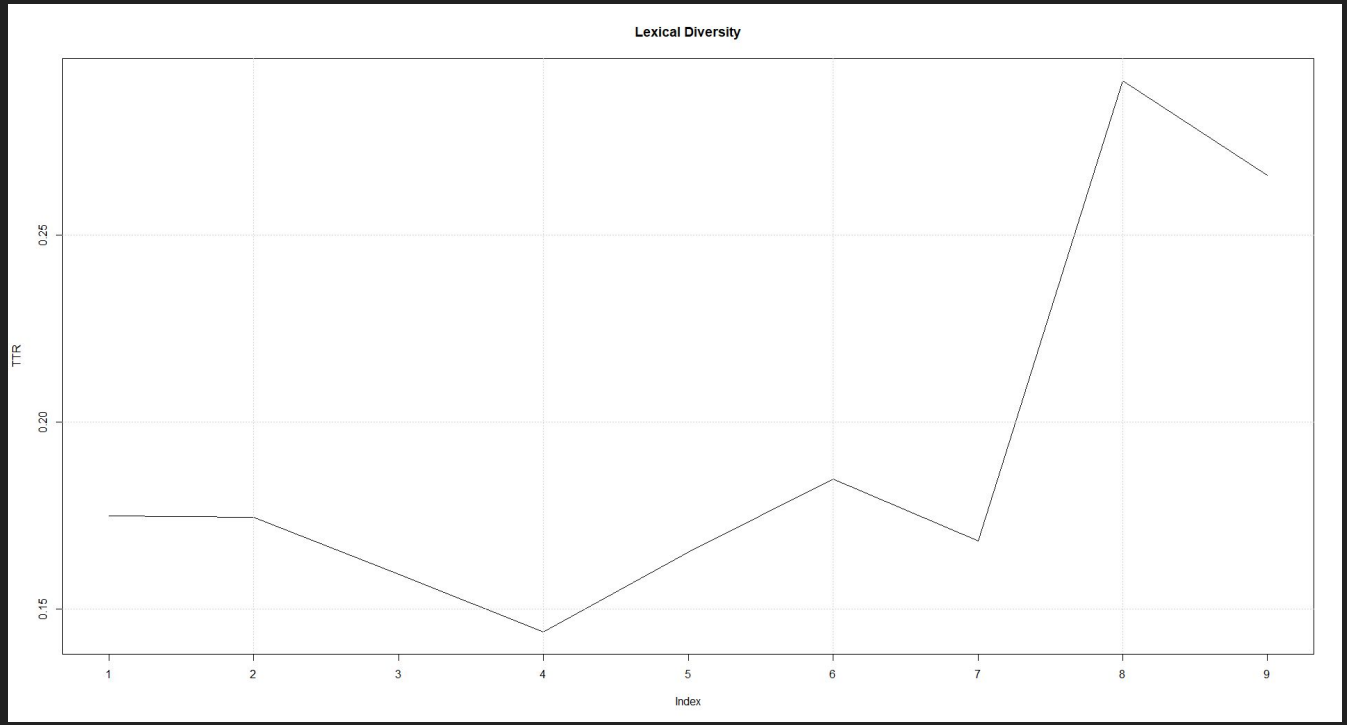
- Made all characters lowercase.
- Removed names from the transcripts.
- Removed stopwords.
- Removed punctuation and text errors.
- Removed numbers.
- Stemmed the text.

Common Words and Themes

- A word cloud to the right:
- Common Terms:
  - Jedi
  - Cockpit
  - Destroy
  - Droid
  - Death
  - Power
  - Space
  - Fighter
  - Senate

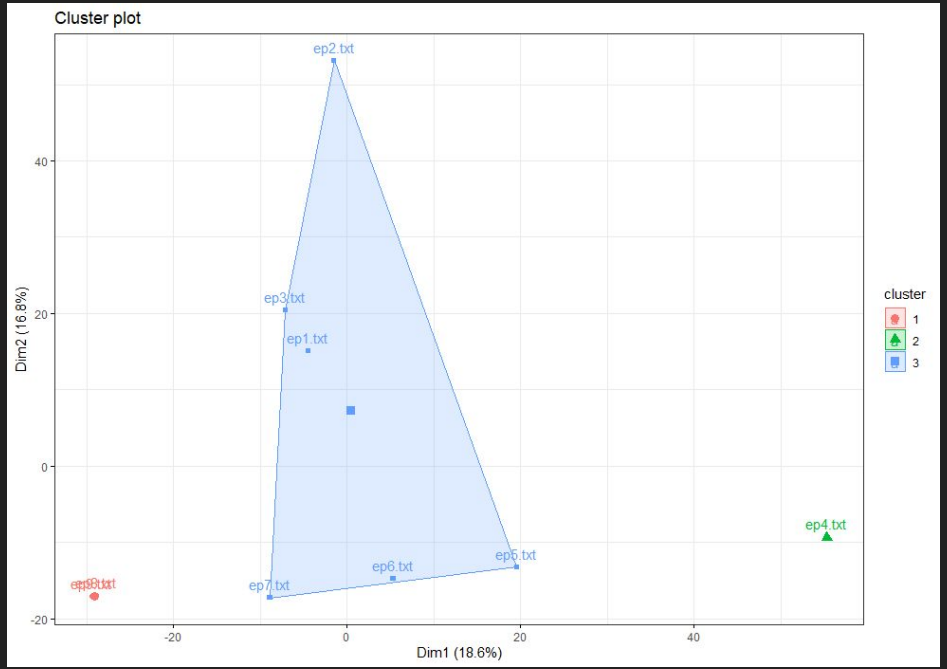
Lexical Diversity

- Measure of the complexity of the language.



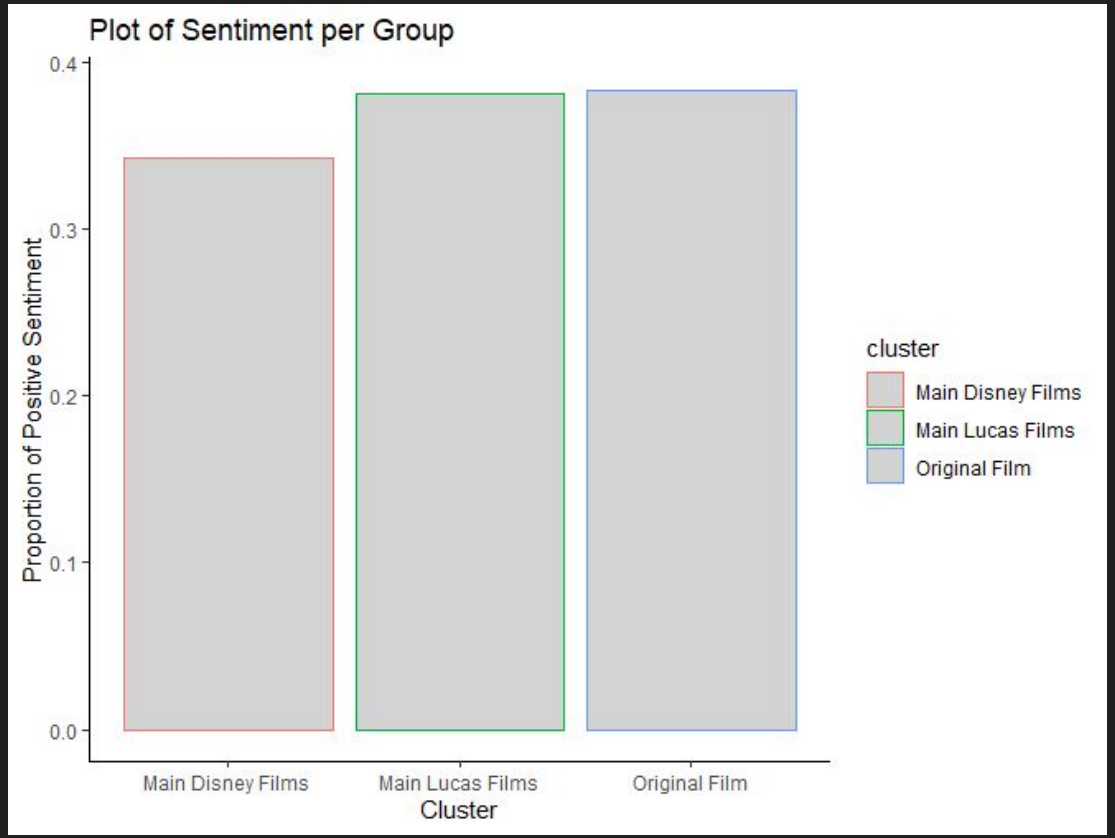
K-Means

- I grouped the films using Euclidean distance into 3 groups.
- I plotted it with the factoextra package.
- The clusters were not expected.



Sentiment Analysis

- Overwhelmingly Negative
- Similar sentiments
  - 4
  - 1, 2, 3, 5, 6, 7



Conclusions

- Similarities across films:
  - Simple language.
  - Generally negative pieces.
  - Uses lots of setting and action over dialogue.
- The groups are not formed into the trilogies.
- Differences:
  - Episodes 8 and 9 use more complex and darker language.
  - Episode 4 is happier and uses the least complex terms.
  - Episodes 5,6,7 are all extremely similar.
  - Episodes 1,2,3 are all extremely similar.

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