How do *Star Wars* characters speak?

Research Project

Shivani Ishwar
Prof. Michelle McSweeney
Pratt Institute

Audience Statement

This research project is intended for an informal, entertainment-focused publication, such as a blog post on *Medium* or an article on *Polygon* or *Vulture*. It would be especially well-suited for the November-December timeframe, which is typically when new *Star Wars* movies release and when fans are most active and excited about the franchise.

Similar text analysis stories of pop culture have been undertaken before, including these analyses of *The Office*, *Parks and Recreation*, and mystery novels.

Introduction

When your story takes place in a galaxy far, far away, you can take plenty of liberties with how it's told. That's a lesson that George Lucas has taught us all since the first *Star Wars* movie's release back in 1977. From far-off settings to the wide variety of alien characters, *Star Wars* has always had opportunities to make creative choices with its stories. This can range from the set-up of the plot to how the characters act, dress, and talk.

One of the most iconic creative choices the original *Star Wars* franchise made was the decision to include the distinctive speech patterns of Yoda — his peculiar manner of constructing sentences has been a topic of conversation among fans and linguists since the character's debut. While on the surface it may seem that Yoda is the only character who has a distinctive speaking style, the truth is there are subtler patterns to everyone's speech, both in fiction and in real life.

The question I set out to answer is: Are there differences in the ways various *Star Wars* characters speak? And if so, how major are those differences?

With new movies coming out almost every year, plus TV shows and other media that add to the franchise's rich storyline, *Star Wars* is in no danger of going out of fashion. But while new characters, settings, and plotlines add new chapters to the beloved story, most fans agree that the original trilogy of movies — *A New Hope*, *The Empire Strikes Back*, and *Return of the Jedi* — represents the heart and soul of *Star Wars*. In order to capture that original spirit, I focused on those three movies for this research project, hoping to reveal not just trends in each character's dialogue but also the underlying themes for the story as a whole.

Methodology

In order to answer my questions about the *Star Wars* movies, I had to find the right data and the right methods. Kaggle, a resource that provides free datasets of all kinds, has a dataset called "Star Wars Movie Scripts," which is made up of three text files of every line of dialogue from the original trilogy of movies. These text files include a number for each spoken line, the character who says the line, and the dialogue of the line.

Because these text files aren't full scripts of the movie, some context is missing — for example, there is no information about the setting or time in which each line is spoken. I wouldn't be able to answer a question like, "Do characters speak differently on Hoth than on Tatooine?" But because my main question is about how different characters speak, and not about other variables, this limitation didn't hinder my research in a noticeable way.

Another limitation, which I couldn't fully control for, is the tendency of some characters to "speak for" others. Because *Star Wars* has plenty of characters who don't use English to communicate, sometimes other characters compensate for them by rephrasing their meaning in their own dialogue. So, for example, Han Solo's lines might include some of Chewbacca's original meaning, or C-3PO might talk for R2-D2 on occasion. Besides going through each line individually, there was no way to correct this skew. My hope is that those skewed lines weren't a significant proportion of the total lines spoken by each character, so it didn't have too much influence on the outcome of my analysis.

Once I had my dataset, I decided to use text analysis methods in the programming language R. Text analysis encompasses a lot of different types of analyses, but the ones that I focused on were sentiment analysis, word frequency analysis, and topic modeling.

First, I used sentiment analysis on the entirety of the *Star Wars* script in order to find a baseline for the, well, underlying sentiment of the script. Using two different analyzers, called NRC and AFINN, allowed me to draw different conclusions about the *Star Wars* script. First, I used NRC to filter words into different categories, like "joy," "trust," "sadness," and "anger," allowing me to see how many words fit in each emotional category. Then, I used AFINN to give each word a weighted "score": higher points for more positive words, and lower points for more negative words.

After finding the baseline score for the dialogue of the entire *Star Wars* trilogy, I used AFINN to find scores for individual characters: our main trio of Luke Skywalker, Han Solo, and Leia Organa, plus other characters who had 100 lines or more across all three

scripts. My final cast included seven characters: Luke, Han, Leia, Darth Vader, Obi Wan Kenobi, C-3PO, and Lando Calrissian.

Once the sentiment analysis was concluded, I moved on to word frequency analysis. This is a tool that allows you to measure the most commonly-used words out of a group; I removed some common words that are used in typical speech—like “a,” “the,” “we,” and “it,” among many others—and then used word frequency analysis to find the most common words used by each of our seven characters, in order to find any trends in what they talk about the most over the course of the three movies.

Word frequency is just one way to find the main themes in a character’s speech — another is called topic modeling. I removed the same words from each character’s dialogue, then used this method in order to find some cohesive “topics” that the characters talked about. These topics aren’t easily labeled categories like “the Force,” “hope,” or “saving the day” — but they list words that are associated with each topic, so I could still make some connections about what each character’s “topics” were.

Results

In total, across all three of the original *Star Wars* movies, there are 2,523 lines of dialogue, with a total of 25,938 words. That's an average of about 10.3 words per line, and 841 lines for each movie. *Return of the Jedi* had the least number of lines, at 674, and *A New Hope* had the most, with 1,010.

The seven characters with the most lines across all three movies are Luke Skywalker, Han Solo, Leia Organa, Darth Vader, Obi Wan Kenobi, C-3PO, and Lando Calrissian. Here's what their dialogue looks like:

Character	Number of Lines	Number of Words	Words per Line
Luke	494	4,521	9.15
Han	459	4,293	9.35
C-3PO	301	3,594	11.9
Leia	227	1,780	7.84
Vader	140	1,554	11.1
Obi Wan	115	1,850	16.1
Lando	101	1,042	10.3

As we can see, Luke has the most lines and words, while Lando has the least lines and words. Obi Wan has the most words per line, while Leia has the least.

These figures are interesting to look at, but they're just the tip of the iceberg of what we can learn by analyzing this dialogue. Let's dig deeper.

SENTIMENT ANALYSIS

Star Wars is known to be a hopeful, optimistic story. But there's plenty of negativity present, too. The NRC analysis can give us a broad idea of which words fall where on the spectrum of emotions.

But there are also some complications: certain words in typical English have a positive or negative connotation, while in the *Star Wars* universe they have a different one. In my

analysis, I expected to encounter "force" as one of these words: typically we think of that as being pretty negative. But I was surprised to discover that "Obi" and "Wan," the two halves of one of our heroes' names, also counted as negative for NRC. So I took those words out when doing my calculations.

From the *Star Wars* scripts, NRC categorized a total of 319 words as positive and 173 as negative. Including repeats of the same word, positive words appeared 1,401 times, and negative words appeared 1,121 times. I decided to break it down further by specific sentiments:

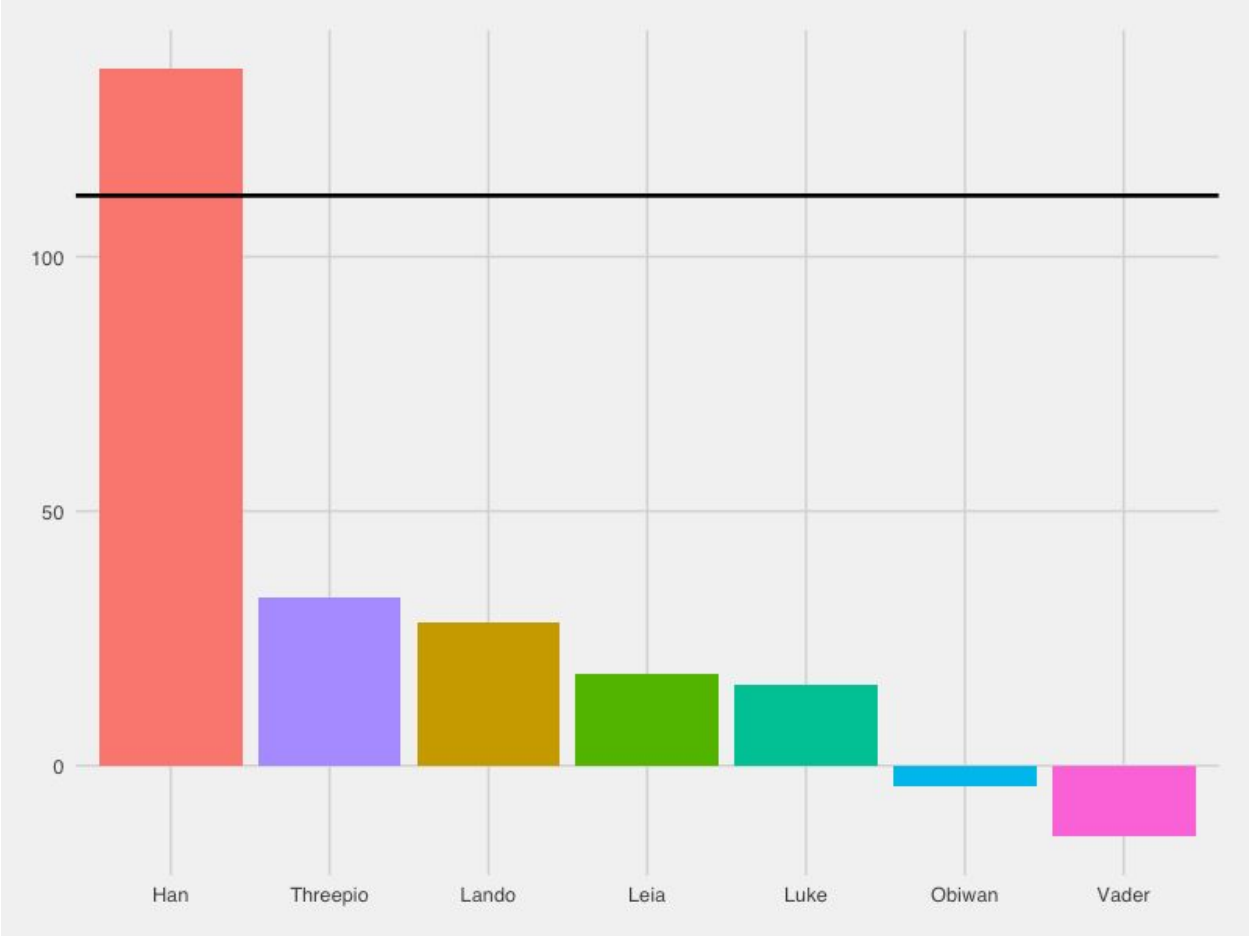
Sentiment	Number of Words	Number of Instances	Most Common Word
joy	101	496	good
sadness	99	472	dark
trust	192	988	sir
anger	75	503	attack
fear	60	703	attack

As we might have expected, the positive emotions outweigh the negative ones. It's also interesting to take a look at which words NRC sorts into which categories. "Sir" is clearly a measure of trust, while "attack" has a lot to do with anger and fear.

AFINN is the other sentiment analyzer I used; it assigns a score between -5 and +5 to words in order to convey their positive or negative meanings. For example, the word "abuse" has a score of -3, and the word "sweet" has a score of +2. By adding together the scores of each word in the dialogue, AFINN gave *Star Wars* a score of 112.

AFINN also assigned a score to each character's individual dialogue, giving a broad idea of how positive and negative they are. Here are the results:

Sentiment Scores by Character

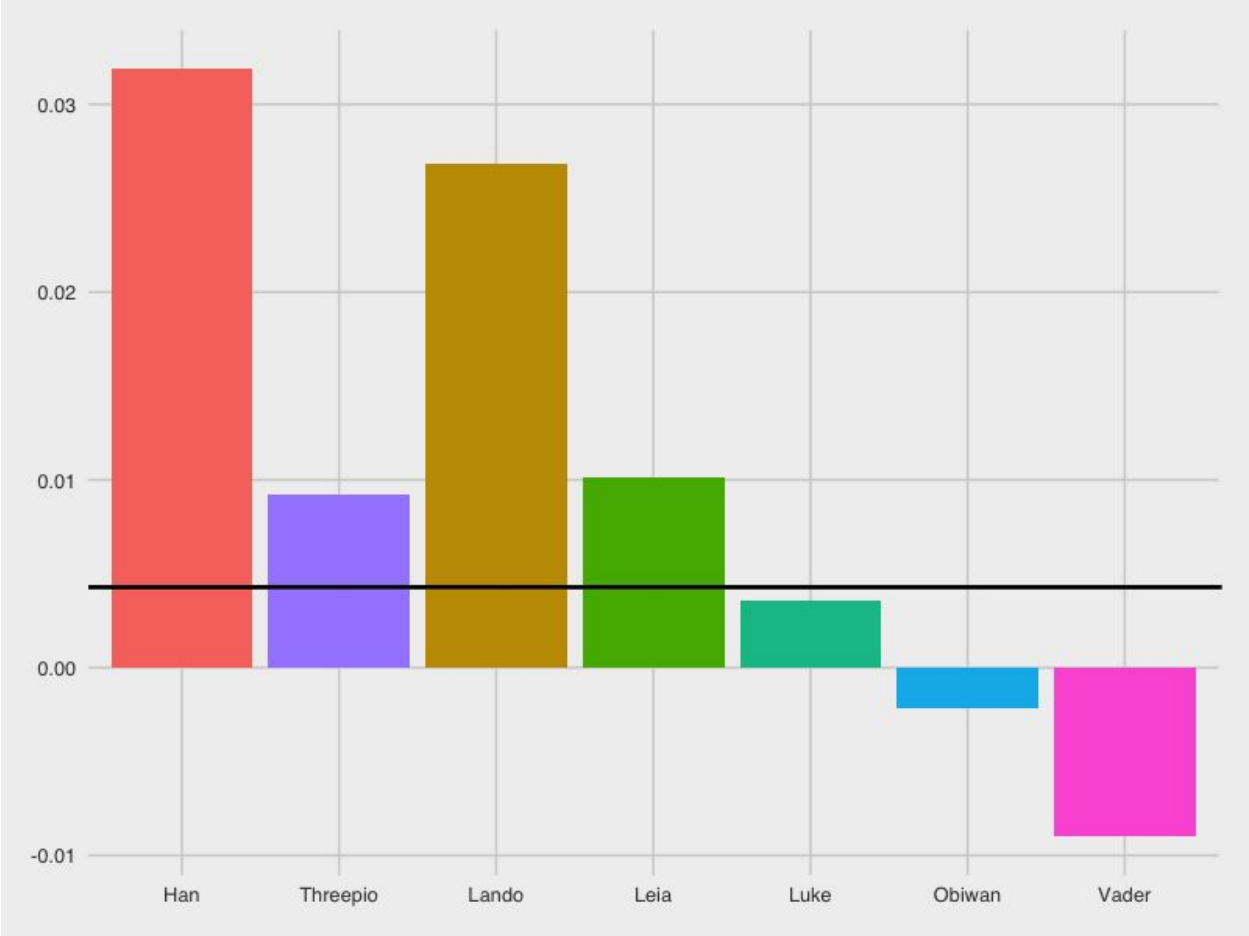


As we can see, Han has the most positive score, with 137, while Darth Vader has the most negative score, with -14. The black line represents *Star Wars*' baseline score of 112.

But of course, this doesn't take into account the whole story, because it doesn't take into account how many words each character says. When we divide each score by the number of words said by that character, we get a different picture.

This time, *Star Wars*' base score is 0.0043. Han still has the highest score at 0.032, and Vader still has the lowest at -0.009, but there are some changes in between:

Sentiment Scores by Character, Normalized by Number of Words



It seems interesting that Han has the most positive score out of all of our characters: considering his reputation as a complaining pessimist, that might seem odd. But Han uses sarcasm a lot in his lines: like "The garbage chute was a really wonderful idea," or the way he teasingly refers to Leia as "Your Worship." Considering that sentiment analyzers have no way to account for sarcasm, it makes much more sense that Han's seemingly-positive words, delivered in a biting tone, might still give him the highest score out of any of the main characters.

WORD FREQUENCY

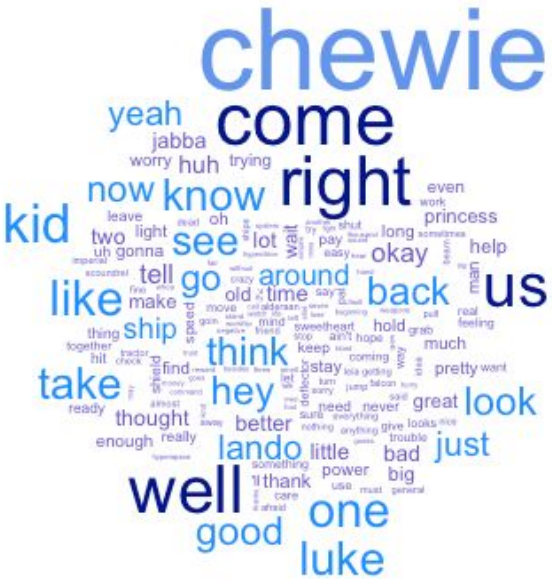
Another way to analyze the dialogue in *Star Wars* is by figuring out the most common words used by each character. After taking out common words in English like "he," "she," "the," "a," and so on, it becomes easy to see trends in what each character is talking about. Here's a breakdown of each character's three most frequent words:

Character	Words
Luke	know, right, come
Han	chewie, come, right
C-3PO	oh, artoo, master
Leia	luke, know, come
Vader	master, now, obi / wan
Obi Wan	luke, force, father
Lando	right, han, vader

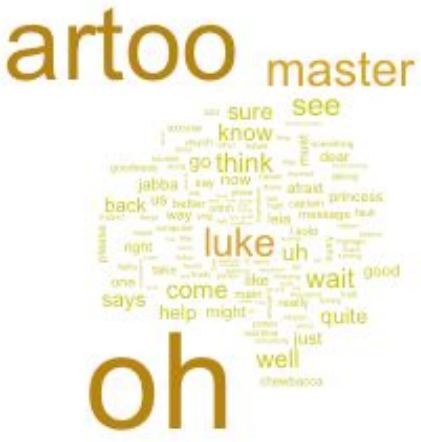
One easy way to visualize the most frequently used words by each character is by using a word cloud. In these visualizations, the larger a word is, the more common it is in the character's dialogue. Here are some word clouds broken down by each character:

Luke

Han



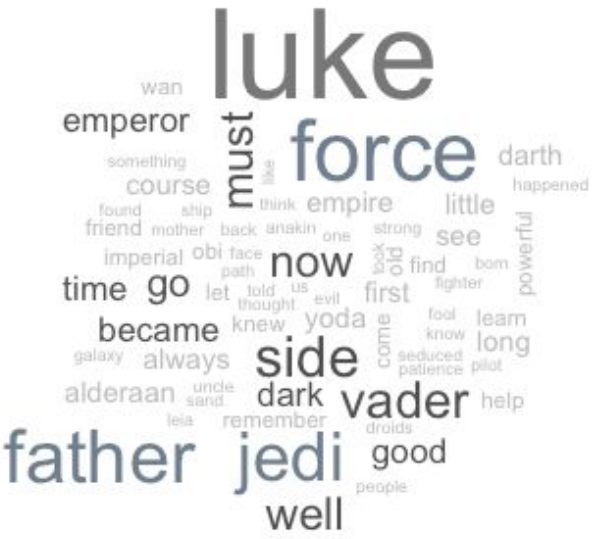
C-3PO



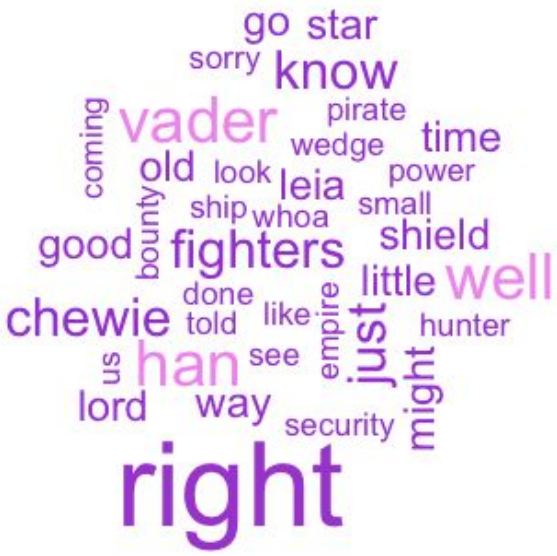
Leia

Vader

Obi Wan



Lando



TOPIC MODELING

While word frequency analysis can give us a quick look at what the characters talk about the most, topic modeling can be even more effective, allowing us to take a closer look at multiple words that are "clustered" together.

Even if topic modeling doesn't provide us with the complete story of what characters talk about, it's still possible to find some common themes. What topic modeling tells us is the words that are talked about *similarly* within each character's dialogue. That can help us paint a picture of what the characters are thinking when they talk within these various topics.

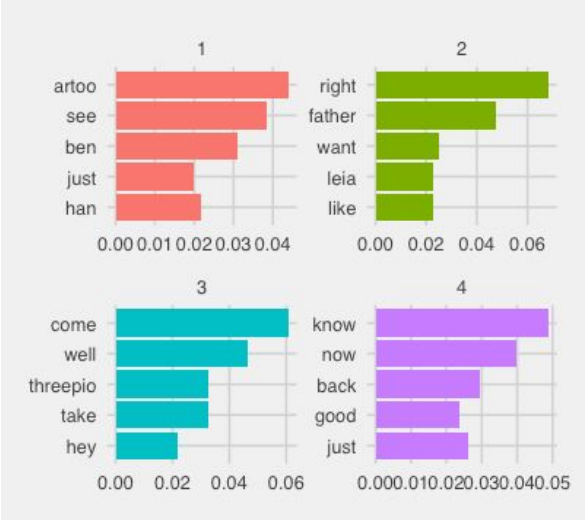
For the purposes of this project, I looked at four topics for each character, which allowed me to see what broad themes each character was talking about. For each topic, the analysis shows the top five words, which can help us reach some conclusions about what each topic is about.

Here are the results of the topic modeling for the whole *Star Wars* script, plus each individual character:

Star Wars

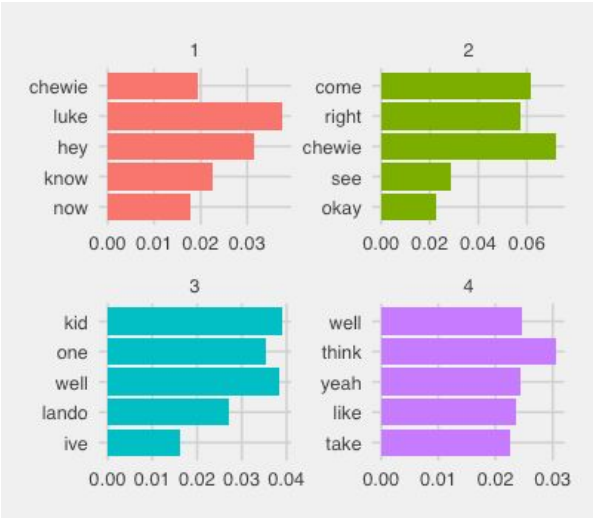
Topic 1: Death Star ("vader," "star")

Luke



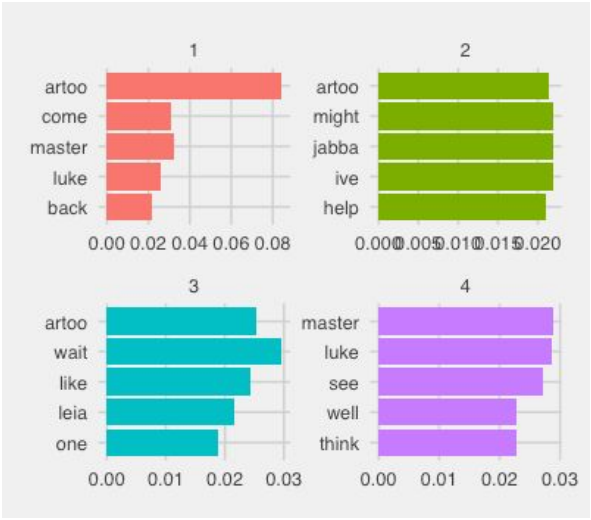
Topic 1: friends ("artoo," "ben," "han")
Topic 2: family ("father," "leia")

Han



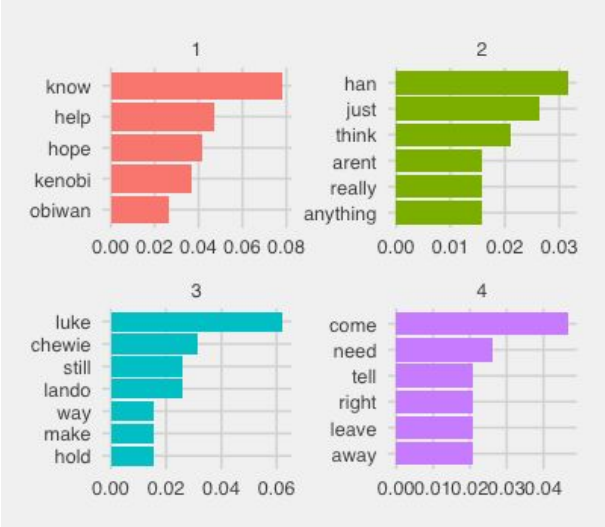
Topic 1: friends ("chewie," "luke")
Topic 3: familiarity ("kid," "lando")

C-3PO



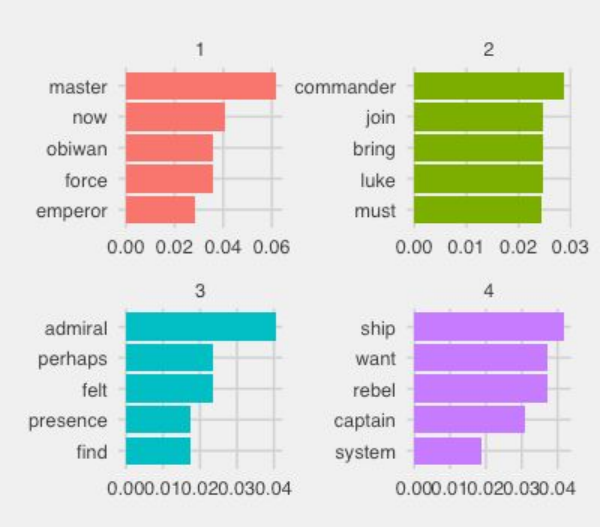
Topic 2: fear ("jabba," "help")
Topic 4: deference ("master," "luke")

Leia



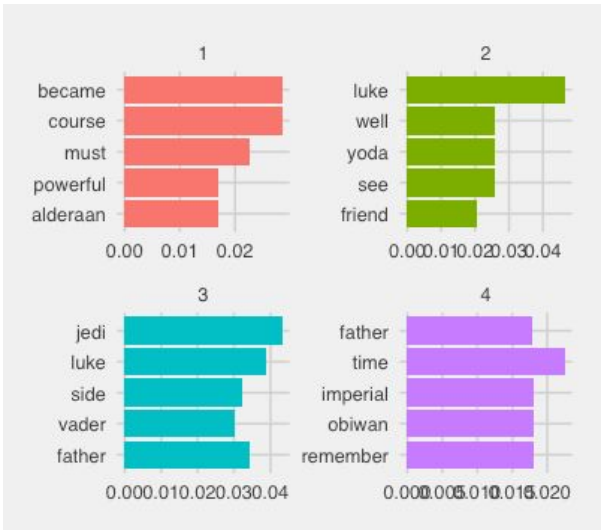
Topic 1: Leia's message ("help," "obiwan," "kenobi," "hope")
Topic 4: command ("come," "leave," "right," "away")

Vader



Topic 1: evil plans ("master," "emperor", "force")
Topic 2: Luke ("luke," "bring", "join")

Obi Wan



Topic 2: Jedi ("luke," "yoda," "friend")
Topic 3: the Dark Side ("jedi," "vader," "father")

Lando

Topic 1: allies ("chewie," "leia," "vader")

As we can see, there are some words that show up in multiple topics: like "Luke" in the general *Star Wars* analysis, or "Chewie" in Han's. The presence of these words indicates that these words, or characters, are often talked about in multiple contexts.

In addition, some characters have more than five words in some of their topics: namely Leia and Lando. That's because these words are tied in frequency, so the topic modeling includes all of them.

While topic modeling doesn't tell us exactly what each "topic" is, a basic knowledge of the *Star Wars* story can help us draw some conclusions. For example, the first topic in the general script has the words "Vader" and "star" in it — that might logically include conversations about the Death Star. Vader's first topic, including the words "master," "emperor," and "force," might consist of his conversations with the Emperor about their evil plans to control the galaxy.

These kinds of insights are difficult to understand without a basic knowledge of the plot of *Star Wars*. Topic modeling is an imperfect method, relying heavily on interpretation in order to draw conclusions. But there are still some interesting themes we can discover by performing this analysis.

In particular, there are a couple of topics that provide interesting insights into the characters. One case is Leia's first topic: it contains "help," "hope," "Obi-Wan," and "Kenobi." If those words sound familiar, it's because they're said together quite a lot in Leia's message to Obi Wan. The full text of the message is: "Help me, Obi Wan Kenobi. You're my only hope." As we can see, those four words from the topic modeling are present in this message. Because it gets played over and over again over the course of the movies, it's no wonder these words show up in the same topic.

Another interesting case is Lando's first topic, containing the words "Chewie," "Vader," and "Leia." This topic indicates that Lando refers to these characters in similar ways, which might seem strange at first, since Chewbacca and Leia are "good" characters, while Vader is an "evil" one. But Lando starts out as a "neutral" character before eventually joining the heroes in the fight against the Empire; in order to help his own people, one of his first acts in the series is to betray the protagonists to Vader's forces. In this context, it makes sense that some of his dialogue may treat Vader similarly to protagonist characters.

Conclusions

Okay, so now we know a lot more about the *Star Wars* original trilogy than we used to. Perhaps we've learned more than we were expecting to. But what's the point of all this?

Well, part of the point, as with any *Star Wars*-related discussion, is because it's fun. The community of *Star Wars* fans hasn't tired, in over 40 years, of finding new things to explore, pick apart, and debate. The language of *Star Wars* is just one tiny facet of this franchise that we can use to find new insights about our favorite space-faring crew.

But even more importantly, there's a reason *Star Wars* has stayed relevant and maintained a solid and even growing fanbase over the past 40-plus years. The stories, characters, and settings all bring people together. People from all walks of life can find something to relate to in the stories of Luke Skywalker, Leia Organa, or Han Solo. *Star Wars* shows us that we have a lot more in common than we think.

Furthermore, *Star Wars*' popularity has had a very real impact on our modern culture. The original franchise's success inspired Hollywood movies to follow its lead and invest more into special effects and stories that appeal to the masses, rather than solely producing dramatic, artful movies for the sake of winning awards. It even inspired real-life science with its sci-fi leanings, encouraging aeronautics to expand our visions of outer space and to reach further for the sake of humanity's exploration of the stars.

Discovering the deeply-embedded trends within the language of the original *Star Wars* movies can help us discover just what has drawn generations of fans to the quirky space epic. In addition to being able to replicate that success in future entertainment, we can find deeper truths about what inspires, impassions, and moves us as human beings.

Future analysis might explore what the rest of the *Star Wars* franchise has to offer: analyzing the speech patterns from novelizations, extended-universe comics, movies, and TV shows can lead to greater insights about the characters we've already explored, as well as adding new ones to the mix. It would also be interesting to see how the tone of the various *Star Wars* movies have changed over the ages: how the original trilogy compares to the prequel trilogy, or to the sequel trilogy set to end this year.

Star Wars has certainly left a lasting imprint on the culture of the whole world, and this project has barely scratched the surface. As the franchise and fandom continue to grow, there will be more and more to learn about just what makes us care so much about this timeless, beloved story.

References

Allen, J. (2018). Text Mining: Every Line from The Office. Retrieved from https://www.jennadallen.com/post/text-analytics-every-line-from-the-office/.

Baert, S. (2018). Happy Galentine's Day! Retrieved from https://suzan.rbind.io/2018/02/happy-galentines-day/.

Cultural impact of Star Wars. (2019). Retrieved from https://en.wikipedia.org/wiki/Cultural_impact_of_Star_Wars.

Kaggle: Your Home for Data Science. (n.d.). Retrieved from https://www.kaggle.com/.

Walsh, B. (2019). Sentiment Analysis of Every Mystery Novel on Gutenberg.org. Retrieved from https://blogs.elon.edu/com329/2019/03/01/sentiment-analysis-of-every-mystery-novel-on-gutenberg-org/.

What is R? (n.d.). Retrieved from https://www.r-project.org/about.html.

Wookieepedia. (n.d.). Retrieved from https://starwars.fandom.com/wiki/.

Xavier, V. (2018). Star Wars Movie Scripts. Retrieved from https://www.kaggle.com/xvivancos/star-wars-movie-scripts.

Appendix

The data used for this research was downloaded from Kaggle. The code used to analyze and visualize the data can be found on GitHub, and is noted in full below.

```
library(dplyr)
library(stringr)
library(tidyverse)
library(tidytext)
library(ggplot2)
library(tm)
library(topicmodels)
library(data.table)
library(ggthemes)
library(wordcloud)
library(RColorBrewer)

#basic data cleaning

sw4 <-
  read.csv("~/Desktop/SW_EpisodeIV.csv",stringsAsFactors=FALSE)
sw5 <- read.csv("~/Desktop/SW_EpisodeV.csv",stringsAsFactors=FALSE)
sw6 <-
  read.csv("~/Desktop/SW_EpisodeVI.csv",stringsAsFactors=FALSE)

sw4$X <- "4"
sw5$X <- "5"
sw6$X <- "6"

sw <- rbind(sw4,sw5,sw6)
names(sw) <- c("movie","character","dialogue")

glimpse(sw)
summary(sw)

#sentiment analysis

get_sentiments("afinn")
get_sentiments("nrc")

sw_tidy <- sw %>% ungroup() %>% unnest_tokens(word,dialogue)

nrc_joy <- get_sentiments("nrc") %>% filter(sentiment=="joy")
nrc_sad <- get_sentiments("nrc") %>% filter(sentiment=="sadness")
nrc_anger <- get_sentiments("nrc") %>% filter(sentiment=="anger")
```

```
nrc_trust <- get_sentiments("nrc") %>% filter(sentiment=="trust")
nrc_fear <- get_sentiments("nrc") %>% filter(sentiment=="fear")
nrc_pos <- get_sentiments("nrc") %>% filter(sentiment=="positive")
nrc_neg <- get_sentiments("nrc") %>% filter(sentiment=="negative")

sw_joy <- sw_tidy %>% inner_join(nrc_joy) %>%
dplyr::count(word, sort=TRUE)
#returns 101 instances
sw_sadness <- sw_tidy %>% inner_join(nrc_sad) %>%
dplyr::count(word, sort=TRUE)
#"wan" like in "obi wan" comes up as a sad word!
#not counting those, returns 99 instances
sw_anger <- sw_tidy %>% inner_join(nrc_anger) %>%
dplyr::count(word,sort=TRUE)
#"force" comes up as an angry word! oops!
#not counting those, returns 75 instances
sw_trust <- sw_tidy %>% inner_join(nrc_trust) %>%
dplyr::count(word,sort=TRUE)
#returns 192 instances
sw_fear <- sw_tidy %>% inner_join(nrc_fear) %>%
dplyr::count(word,sort=TRUE)
#"force," "obi," and "wan" come up as fearful words!
#not counting those, returns 60 instances
sw_pos <- sw_tidy %>% inner_join(nrc_pos) %>%
dplyr::count(word,sort=TRUE)
#returns 319 instances
sw_neg <- sw_tidy %>% inner_join(nrc_neg) %>%
dplyr::count(word,sort=TRUE)
#not counting "force," "obi," or "wan," returns 173 instances

sum(sw_joy$n) #returns 496
sum(sw_sadness$n) #returns 472
sum(sw_anger$n) #returns 503
sum(sw_trust$n) #returns 988
sum(sw_fear$n) #returns 703
sum(sw_pos$n) #returns 1401
sum(sw_neg$n) #returns 1121

sw_sent <- sw_tidy %>% inner_join(get_sentiments("afinn"))
#afinn does not count "force," "obi," or "wan" as negative words!
#hooray!
sw_score <- sum(sw_sent$value) #returns 112

#individual characters

luke <- sw[which(sw$character=="LUKE"),]
han <- sw[which(sw$character=="HAN"),]
```

```
leia <- sw[which(sw$character=="LEIA"),]
vader <- sw[which(sw$character=="VADER"),]
threepio <- sw[which(sw$character=="THREEPIO"),]
obiwan <- sw[which(sw$character=="BEN"),]
lando <- sw[which(sw$character=="LANDO"),]

#note: vader is the only bad guy with over 50 lines.
#i chose corpora where the character has over 50 lines.
#funnily, after lando (101 lines), the next is yoda (49 lines).
#so all my corpora have >100 lines.

luke_tidy <- luke %>% ungroup() %>% unnest_tokens(word,dialogue)
han_tidy <- han %>% ungroup() %>% unnest_tokens(word,dialogue)
leia_tidy <- leia %>% ungroup() %>% unnest_tokens(word,dialogue)
vader_tidy <- vader %>% ungroup() %>% unnest_tokens(word,dialogue)
threepio_tidy <- threepio %>% ungroup() %>%
  unnest_tokens(word,dialogue)
obiwan_tidy <- obiwan %>% ungroup() %>%
  unnest_tokens(word,dialogue)
lando_tidy <- lando %>% ungroup() %>% unnest_tokens(word,dialogue)

luke_sent <- luke_tidy %>% inner_join(get_sentiments("afinn"))
luke_score <- sum(luke_sent$value) #returns 16
han_sent <- han_tidy %>% inner_join(get_sentiments("afinn"))
han_score <- sum(han_sent$value) #returns 137
leia_sent <- leia_tidy %>% inner_join(get_sentiments("afinn"))
leia_score <- sum(leia_sent$value) #returns 18
vader_sent <- vader_tidy %>% inner_join(get_sentiments("afinn"))
vader_score <- sum(vader_sent$value) #returns -14
threepio_sent <- threepio_tidy %>%
  inner_join(get_sentiments("afinn"))
threepio_score <- sum(threepio_sent$value) #returns 33
obiwan_sent <- obiwan_tidy %>% inner_join(get_sentiments("afinn"))
obiwan_score <- sum(obiwan_sent$value) #returns -4
lando_sent <- lando_tidy %>% inner_join(get_sentiments("afinn"))
lando_score <- sum(lando_sent$value) #returns 28

#word frequency

sw_stops <-
  c(stopwords("en"),"got","will","dont","will","cant","can","youre"
  ,

  "thats","yes","sir","going","hes","get","ill","theyre","theres",
              "didnt","whats")

luke_freq <- luke_tidy
```

```
luke_freq$word <- luke_freq$word %>% removeWords(sw_stops)
luke_freq <- na.omit(luke_freq)
luke_freq <- luke_freq %>% count(word,sort=TRUE)
luke_freq <- luke_freq[-c(1),]

han_freq <- han_tidy
han_freq$word <- han_freq$word %>% removeWords(sw_stops)
han_freq <- na.omit(han_freq)
han_freq <- han_freq %>% count(word,sort=TRUE)
han_freq <- han_freq[-c(1),]

leia_freq <- leia_tidy
leia_freq$word <- leia_freq$word %>% removeWords(sw_stops)
leia_freq <- na.omit(leia_freq)
leia_freq <- leia_freq %>% count(word,sort=TRUE)
leia_freq <- leia_freq[-c(1),]

vader_freq <- vader_tidy
vader_freq$word <- vader_freq$word %>% removeWords(sw_stops)
vader_freq <- na.omit(vader_freq)
vader_freq <- vader_freq %>% count(word,sort=TRUE)
vader_freq <- vader_freq[-c(1),]

threepio_freq <- threepio_tidy
threepio_freq$word <- threepio_freq$word %>% removeWords(sw_stops)
threepio_freq <- na.omit(threepio_freq)
threepio_freq <- threepio_freq %>% count(word,sort=TRUE)
threepio_freq <- threepio_freq[-c(1),]

obiwan_freq <- obiwan_tidy
obiwan_freq$word <- obiwan_freq$word %>% removeWords(sw_stops)
obiwan_freq <- na.omit(obiwan_freq)
obiwan_freq <- obiwan_freq %>% count(word,sort=TRUE)
obiwan_freq <- obiwan_freq[-c(1),]

lando_freq <- lando_tidy
lando_freq$word <- lando_freq$word %>% removeWords(sw_stops)
lando_freq <- na.omit(lando_freq)
lando_freq <- lando_freq %>% count(word,sort=TRUE)
lando_freq <- lando_freq[-c(1),]

#transforming into corpora

sw_corpus <- VCorpus(VectorSource(sw$dialogue))

luke_corpus <- VCorpus(VectorSource(luke$dialogue))
han_corpus <- VCorpus(VectorSource(han$dialogue))
```

```
leia_corpus <- VCorpus(VectorSource(leia$dialogue))
vader_corpus <- VCorpus(VectorSource(vader$dialogue))
threepio_corpus <- VCorpus(VectorSource(threepio$dialogue))
obiwan_corpus <- VCorpus(VectorSource(obiwan$dialogue))
lando_corpus <- VCorpus(VectorSource(lando$dialogue))

#cleaning main corpus

sw_clean <- tm_map(sw_corpus, removeNumbers)
sw_clean <- tm_map(sw_clean, removePunctuation)
sw_clean <- tm_map(sw_clean, stripWhitespace)
sw_clean <- tm_map(sw_clean, content_transformer(tolower))
sw_clean <- tm_map(sw_clean, removeWords, sw_stops)

#topic modeling

sw_dtm <- DocumentTermMatrix(sw_clean)
unique_indexes <- unique(sw_dtm$i)
sw_dtm <- sw_dtm[unique_indexes,]

k <- 4
sw_lda <- LDA(sw_dtm,k=k,control=list(seed=1234))
sw_lda_words <- terms(sw_lda,5)
sw_lda_tidy <- tidy(sw_lda)

sw_top_terms <- sw_lda_tidy %>% group_by(topic) %>% top_n(5,beta)
  %>%
  ungroup() %>% arrange(topic, -beta)

#repeated process for smaller corpora

luke_clean <- tm_map(luke_corpus, removeNumbers)
luke_clean <- tm_map(luke_clean, removePunctuation)
luke_clean <- tm_map(luke_clean, stripWhitespace)
luke_clean <- tm_map(luke_clean, content_transformer(tolower))
luke_clean <- tm_map(luke_clean, removeWords, sw_stops)

luke_dtm <- DocumentTermMatrix(luke_clean)
unique_luke <- unique(luke_dtm$i)
luke_dtm <- luke_dtm[unique_luke,]

luke_lda <- LDA(luke_dtm,k=k,control=list(seed=1234))
luke_lda_words <- terms(luke_lda,5)
luke_lda_tidy <- tidy(luke_lda)

luke_top_terms <- luke_lda_tidy %>% group_by(topic) %>%
  top_n(5,beta) %>%
```

```
  ungroup() %>% arrange(topic, -beta)

han_clean <- tm_map(han_corpus, removeNumbers)
han_clean <- tm_map(han_clean, removePunctuation)
han_clean <- tm_map(han_clean, stripWhitespace)
han_clean <- tm_map(han_clean, content_transformer(tolower))
han_clean <- tm_map(han_clean, removeWords, sw_stops)

han_dtm <- DocumentTermMatrix(han_clean)
unique_han <- unique(han_dtm$i)
han_dtm <- han_dtm[unique_han,]

han_lda <- LDA(han_dtm,k=k,control=list(seed=1234))
han_lda_words <- terms(han_lda,5)
han_lda_tidy <- tidy(han_lda)

han_top_terms <- han_lda_tidy %>% group_by(topic) %>% top_n(5,beta)
  %>%
  ungroup() %>% arrange(topic, -beta)

leia_clean <- tm_map(leia_corpus, removeNumbers)
leia_clean <- tm_map(leia_clean, removePunctuation)
leia_clean <- tm_map(leia_clean, stripWhitespace)
leia_clean <- tm_map(leia_clean, content_transformer(tolower))
leia_clean <- tm_map(leia_clean, removeWords, sw_stops)

leia_dtm <- DocumentTermMatrix(leia_clean)
unique_leia <- unique(leia_dtm$i)
leia_dtm <- leia_dtm[unique_leia,]

leia_lda <- LDA(leia_dtm,k=k,control=list(seed=1234))
leia_lda_words <- terms(leia_lda,5)
leia_lda_tidy <- tidy(leia_lda)

leia_top_terms <- leia_lda_tidy %>% group_by(topic) %>%
  top_n(5,beta) %>%
  ungroup() %>% arrange(topic, -beta)

vader_clean <- tm_map(vader_corpus, removeNumbers)
vader_clean <- tm_map(vader_clean, removePunctuation)
vader_clean <- tm_map(vader_clean, stripWhitespace)
vader_clean <- tm_map(vader_clean, content_transformer(tolower))
vader_clean <- tm_map(vader_clean, removeWords, sw_stops)

vader_dtm <- DocumentTermMatrix(vader_clean)
unique_vader <- unique(vader_dtm$i)
vader_dtm <- vader_dtm[unique_vader,]
```

```
vader_lda <- LDA(vader_dtm,k=k,control=list(seed=1234))
vader_lda_words <- terms(vader_lda,5)
vader_lda_tidy <- tidy(vader_lda)

vader_top_terms <- vader_lda_tidy %>% group_by(topic) %>%
  top_n(5,beta) %>%
  ungroup() %>% arrange(topic, -beta)

threepio_clean <- tm_map(threepio_corpus, removeNumbers)
threepio_clean <- tm_map(threepio_clean, removePunctuation)
threepio_clean <- tm_map(threepio_clean, stripWhitespace)
threepio_clean <- tm_map(threepio_clean,
  content_transformer(tolower))
threepio_clean <- tm_map(threepio_clean, removeWords, sw_stops)

threepio_dtm <- DocumentTermMatrix(threepio_clean)
unique_threepio <- unique(threepio_dtm$i)
threepio_dtm <- threepio_dtm[unique_threepio,]

threepio_lda <- LDA(threepio_dtm,k=k,control=list(seed=1234))
threepio_lda_words <- terms(threepio_lda,5)
threepio_lda_tidy <- tidy(threepio_lda)

threepio_top_terms <- threepio_lda_tidy %>% group_by(topic) %>%
  top_n(5,beta) %>%
  ungroup() %>% arrange(topic, -beta)

obiwan_clean <- tm_map(obiwan_corpus, removeNumbers)
obiwan_clean <- tm_map(obiwan_clean, removePunctuation)
obiwan_clean <- tm_map(obiwan_clean, stripWhitespace)
obiwan_clean <- tm_map(obiwan_clean, content_transformer(tolower))
obiwan_clean <- tm_map(obiwan_clean, removeWords, sw_stops)

obiwan_dtm <- DocumentTermMatrix(obiwan_clean)
unique_obiwan <- unique(obiwan_dtm$i)
obiwan_dtm <- obiwan_dtm[unique_obiwan,]

obiwan_lda <- LDA(obiwan_dtm,k=k,control=list(seed=1234))
obiwan_lda_words <- terms(obiwan_lda,5)
obiwan_lda_tidy <- tidy(obiwan_lda)

obiwan_top_terms <- obiwan_lda_tidy %>% group_by(topic) %>%
  top_n(5,beta) %>%
  ungroup() %>% arrange(topic, -beta)

lando_clean <- tm_map(lando_corpus, removeNumbers)
```

```
lando_clean <- tm_map(lando_clean, removePunctuation)
lando_clean <- tm_map(lando_clean, stripWhitespace)
lando_clean <- tm_map(lando_clean, content_transformer(tolower))
lando_clean <- tm_map(lando_clean, removeWords, sw_stops)

lando_dtm <- DocumentTermMatrix(lando_clean)
unique_lando <- unique(lando_dtm$i)
lando_dtm <- lando_dtm[unique_lando,]

lando_lda <- LDA(lando_dtm,k=k,control=list(seed=1234))
lando_lda_words <- terms(lando_lda,5)
lando_lda_tidy <- tidy(lando_lda)

lando_top_terms <- lando_lda_tidy %>% group_by(topic) %>%
  top_n(5,beta) %>%
  ungroup() %>% arrange(topic, -beta)

#sentiment analysis visualization

sent_viz <-
  data.frame("character"=c("Luke","Han","Leia","Vader","Threepio",
                           "Obiwan","Lando"),
                     "score"=c(luke_score, han_score, leia_score,
                               vader_score, threepio_score,
                               obiwan_score, lando_score))

ggplot(sent_viz,aes(x=reorder(character,-score),y=score,fill=charac
  ter)) +
  geom_col(show.legend=FALSE) +
  geom_hline(yintercept=sw_score,col="black",size=1) +
  theme_fivethirtyeight()

sent_viz$score_norm <- c((luke_score/count(luke_tidy)),
                         (han_score/count(han_tidy)),
                         (leia_score/count(leia_tidy)),
                         (vader_score/count(vader_tidy)),
                         (threepio_score/count(threepio_tidy)),
                         (obiwan_score/count(obiwan_tidy)),
                         (lando_score/count(lando_tidy)))

sw_score_norm <- as.numeric(sw_score/count(sw_tidy))

ggplot(sent_viz,aes(x=reorder(character,-score),y=score_norm,fill=c
  haracter)) +
  geom_col(show.legend=FALSE) + theme_fivethirtyeight() #+
  geom_hline(yintercept=sw_score_norm,col="black",size=1)
```

```
#word frequency visualizations

wordcloud(luke_freq$word,luke_freq$n,scale=c(4,0.001),

  colors=c("green","forestgreen","mediumspringgreen","darkgreen"))
wordcloud(han_freq$word,han_freq$n,scale=c(4,0.001),

  colors=c("slateblue","dodgerblue","darkblue","cornflowerblue"))
wordcloud(leia_freq$word,leia_freq$n,scale=c(4,0.001),

  colors=c("palevioletred","rosybrown","maroon","deeppink"))
wordcloud(vader_freq$word,vader_freq$n,scale=c(4,0.001),
          colors=c("red","coral","salmon","darkred"))
wordcloud(threepio_freq$word,threepio_freq$n,scale=c(4,0.001),

  colors=c("yellow3","goldenrod","burlywood3","darkgoldenrod"))
wordcloud(obiwan_freq$word,obiwan_freq$n,scale=c(4,0.001),
          colors=c("gray","gray31","slategray","grey47"))
wordcloud(lando_freq$word,lando_freq$n,scale=c(4,0.001),
          colors=c("orchid","darkviolet","violet","darkorchid"))

#topic modeling visualizations

sw_final_terms <- sw_top_terms %>% mutate(term=reorder(term,beta))
  %>%
  ggplot(aes(term,beta,fill=factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic,scales="free") + coord_flip() +
  theme_fivethirtyeight()
sw_final_terms

luke_final_terms <- luke_top_terms %>%
  mutate(term=reorder(term,beta)) %>%
  ggplot(aes(term,beta,fill=factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic,scales="free") + coord_flip() +
  theme_fivethirtyeight()
luke_final_terms

han_final_terms <- han_top_terms %>%
  mutate(term=reorder(term,beta)) %>%
  ggplot(aes(term,beta,fill=factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic,scales="free") + coord_flip() +
  theme_fivethirtyeight()
han_final_terms
```

```
leia_final_terms <- leia_top_terms %>%
  mutate(term=reorder(term,beta)) %>%
  ggplot(aes(term,beta,fill=factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic,scales="free") + coord_flip() +
  theme_fivethirtyeight()
leia_final_terms

vader_final_terms <- vader_top_terms %>%
  mutate(term=reorder(term,beta)) %>%
  ggplot(aes(term,beta,fill=factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic,scales="free") + coord_flip() +
  theme_fivethirtyeight()
vader_final_terms

threepio_final_terms <- threepio_top_terms %>%
  mutate(term=reorder(term,beta)) %>%
  ggplot(aes(term,beta,fill=factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic,scales="free") + coord_flip() +
  theme_fivethirtyeight()
threepio_final_terms

obiwan_final_terms <- obiwan_top_terms %>%
  mutate(term=reorder(term,beta)) %>%
  ggplot(aes(term,beta,fill=factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic,scales="free") + coord_flip() +
  theme_fivethirtyeight()
obiwan_final_terms

lando_final_terms <- lando_top_terms %>%
  mutate(term=reorder(term,beta)) %>%
  ggplot(aes(term,beta,fill=factor(topic))) +
  geom_col(show.legend=FALSE) +
  facet_wrap(~topic,scales="free") + coord_flip() +
  theme_fivethirtyeight()
lando_final_terms
```