Lindsey Feingold

Natural Language Processing

12/2/19

Process and Findings

The research question I was trying to answer with this project is about the Harry Potter franchise: are the spoken words per main character in the books versus the movies different in any way?

I first found the data I needed to do the project and then decided which natural language processing techniques I wanted to use. Right off the bat I knew I would have to use regex to extract the quotes from the book. For the movie the dataset I used already had the spoken words broken up. The first issue I ran into is that the dataset I wanted to use for the books was in R, so I converted them to csv’s so I could analyze the text in Python. Then, my next issue was that a large chunk of quotes were missing one or both quotation marks. This made my regex code not work, since it relied on the quotations to extract quotes. I had to manually go through and fix the quotations, so I decided to only compare the first book to the first movie, since the process took a really long time to do. After I successfully extracted the quotes, I cleaned up both the movie and book datasets. I then got to the analysis.

I first used tf-idf to compare the words spoken for Dumbledore, Harry, Hagrid, Ron, Hermione, Malfoy, Professor McGonagall, Aunt Petunia, Snape, and Voldemort. My results for the movie: Hagrid, Harry, and Hermione all say Voldemort more than they should. Harry says other people's names more than anything else, and Hermione, Ron, Draco, Hagrid, and Dumbledore both say Harry's name A LOT. Dumbledore also loves to award points. McGonagall references Hermione more than Harry. My results for the book: Dumbledore, Hagrid, and Harry all say Voldemort more than they should (compared to the movie, this is very different). Every character says other people's names a lot less than in the movie as well.

I then used sentence length to compare average sentence measurements. For the movie, Snape uses the most characters on average per sentence at about 58. Ron uses the least at about 27. For the book, Snape still says the most but at 45, whereas Harry says the least at around 20. For the movie, Snape also says the most words per sentence on average at about 10, and Malfoy says the least at about 6. In the book, Snape also says the most at 7, and Aunt Petunia says the least at 1 (however, this is because she has very few lines). If we are ignoring Aunt Petunia, Harry says the least with about 3.6 words on average.

Lastly, I calculated the complexity of each character’s spoken words. Specifically, I calculated character count, word count, sentence count, syllable count, and other scores including the Flesch score (the lower the score, the more difficult the text is to read) and the Dalechall score (it measures a text against a number of words considered familiar to fourth-graders, and the more unfamiliar words used, the higher the reading level will be). For the movie: I found that Dumbledore's sentences are the most complex, according to the Flesch score, and that Hermione's sentences are the most complex according to the Dalechall score. Harry had the least complex sentences according to the Flesch score, and Dumbledore had the least complex sentences according to the Dalechall score. For the book: Dumbledore's sentences are still the most complex (same as the movie), according to the Flesch score, and shockingly Hagrid's sentences are the most complex according to the Dalechall score, which listed Hermione as the most complex in the movie. However, that could be because Hagrid says a lot of words in an accent, which J.K. Rowling emphasizes in how she spells words he says. A list of words a fourth-grader would know definitely wouldn't include "yeh" or "gettin'", it would include "your" and "getting" since that is the correct spelling.

Overall, I found that Harry’s speech is usually more complex and contains a higher variety of words in the movie than in the book, except in the case of the movie Flesch score for him. Snape was captured the most accurately from the book to the movie. The other characters vary depending on the analysis type. There are also some limitations with this data. The quotes in the book with "he said" and "she said" were not extracted, which definitely changes things. Snape was an issue during the vecterizing process ( hewasn't calculated correctly), and vectorizing was also an issue with the words “they've” and “they're” – it created ve and re as separate words, which messed up the analysis for Hermione’s tf-idf result in particular.