Jackson Joyner

I have been a lifelong fan of comedy, whether it be watching comedy movies like Step Brothers or Dude Where’s My Car, nothing can heal the soul like laughter. While comedy movies are great, there is nothing better than stand up comedy. There is nothing rawer than a person and a microphone on a stage speaking. I one day aspire to try stand-up comedy, but for now, ill stick to analyzing it. There is a website that houses a bunch of transcripts for tv shows, movies, and stand-up comedy specials called scrapsfromtheloft.com. I decided to scrape all their available stand-up comedy transcripts so I could study some of the greats.

I built a web-scraper that would go through and get all the links from their stand-up page. Then I loop through each link to pull only the transcripts. I then write all of the transcripts to their own csv file. Then I threw the csv’s into a dataframe which also houses data like the title, year the special came out and the descriptive statistics of these transcripts.

There are 300+ transcripts in this dataset with each having an average token length of about 5.5. For most of the transcripts, they have an average token length of around 2.5k tokens. I made sure that I only kept the English specials as non-english specials could be harmful to the dataset. To get the title column and year column it took some finagling with regex, but it ended up getting exactly what we needed. One more weird thing I had to look at is there are stage directions included in some of the transcripts. They look like this: [laughter], [Bert nods his head] that we had to remove, as those aren’t what was said. There is a case to keep them, but I took them out of the data.

Part of this analysis is to get the IMDB ratings and reviews for these specials. We need to get the IMDB id for these specials, that way we can then use those id’s to pull the ratings and reviews from the IMDB API. The dataset shared here, has a total of 2164 unique reviews of these specials. Many of these reviews have a lexical diversity of 1, which means that no word was used twice in these reviews.

You can answer some cool questions like what words indicate that someone is likely to rate the movie highly. I plan on running an ANOVA test on the average rating of these stand-up specials and the clustered data set from my Data Set Share 1 assignment. This will be done with the average rating of the special vs the assigned cluster of the special.