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.

We could answer some interesting questions with this data. How do some comics compare to other comics, which comics have the dirtiest language usage, my guess is Jim Jefferies or Bill Burr. We could then ask questions like which years had the most diverse topics in stand-up comedy. What I am planning on doing with this data, is taking the IMDb ratings and reviews, and then predicting some comedy routines ratings by running the token data through a random forest model.

To determine which comic has the dirtiest language, we would loop through each transcript and compare the tokens of the transcript to a list of all swear words to determine the swear word count. We then would cluster the data based around that swear word count and then determine who has the dirtiest mouth.