December 15, 2021

Joyner, Jackson

Topic Modeling Stand up

I have been a comedy fan for as long as I can remember. I am a big proponent of laughter is the best medicine. While being a fan of stand-up comedy, I started listening to their podcasts. One of the common themes of these podcasts is how similar comedians are in nature, but not in their content. I find this idea fascinating and decided to dig a bit deeper into that by performing a cluster analysis of their content. There is a website called scrapsfromtheloft.com that houses over 340 stand up transcripts. Let’s break down what happened.

After cleaning the data up a bit, removing non-English transcripts we ended up with a total of 340 usable transcripts. Once the data had been cleaned up, I decided to get the total token count, total unique token count, average token length, the lexical diversity, and most frequent words. In Table 1, we have some descriptive stats for the transcripts.

|  |  |  |
| --- | --- | --- |
|  | Mean | STD DEV |
| Total Tokens | 2385.54 | 871.68 |
| Unique Tokens | 917.14 | 296.06 |
| Avg Token Length | 5.18 | .18 |
| Lexical Diversity | .41 | .09 |
| Run Time (Minutes) | 67.66 | 14.81 |

Table 1: Descriptive Stats of Comedy Transcripts

As I was analyzing the data, I was bewildered by how small the overall token count was. I was expecting a much higher token count as the average run time for these specials is over an hour long.

After multiple data point gathering iterations, I came to the decision on how to classify the text within the corpora. I ended up choosing Latent Dirchelet Allocation as my classification methodology. It is one of the newest ways of clustering text data and being on the forefront of technology interested me greatly. When running the initial test of the LDA modeling, I had 7 clusters originally, but found that the last 3 clusters only had 2-3 transcripts inside of them, so I decided to drop three overall clusters to a total of 4.

With this LDA topic modeling, once the clusters are built, you get a list of words that are most likely to be apart of each group. In the image below you can see the most likely words for each cluster. Text

Description automatically generated

Image 1: Most probable words for each cluster

As you can see, group one and three have rather coarse language, while group three is seemingly British. What I had to do is label each group based on the words in these lists. I ended up deciding on the labels: Stories, Minority Issues, UK, and Not-Clean.

A picture containing text, meter, device

Description automatically generated

Image 2: Total cluster counts

When all was said of done, the image above is the total count of transcripts in each cluster. What I noticed was that this model doesn’t seem to cluster the transcripts evenly. This made me look into the coherence score which shows the level cohesiveness between words in the model. A good coherence score of .6-.8 is considered a good fit for the LDA model. My model produced a coherence score of .33, which is considered weak. While the model does produce clusters, the clusters could be greatly improved on.

When producing the model, there are numerous “dials” that can be turned, tuned, and fiddled with that could make this model more predictive and accurate. Due to the time constraints, I just had to pick certain settings to run with and this is what I ended up with. If I were to have had more time with the data, I would expect to have a more accurate model. I would like to get to a coherence score above .600 which would indicate a good model was created.

Working on this dataset has been quite a bit of fun for me. Seeing the whole process of data acquisition, data cleaning, and analysis was very interesting, and kept me engaged the entire way. There is something to working on something you can take sole ownership of that made the whole process more rewarding. When I was looking through the final clusters, I noticed that there were some inconsistencies with the groups that I would love to iron out. Overall, LDA was fun to work with and I think there are some practical applications that would utilize LDA fairly well.