A4 Part 4

Part 1: Cluster Analysis

What we did with K-Means Clustering:

- We used K-Means to study a bunch of movie titles from the IMDB dataset.
- First, we cleaned up the dataset by getting rid of some columns we didn't need.
- We then used TF-IDF to turn the movie titles into a format that the computer can understand and analyze.
- To figure out how many groups (clusters) to divide these titles into, we used the elbow method. This method helps us see at what point adding more clusters doesn't really help us get better information.
- We decided that the best number of groups for our data was 3.

What we did with Hierarchical Clustering:

- We also looked at the movie titles using another method called hierarchical clustering.
- We used three different ways (single, complete, average) to see how the titles can be grouped based on how similar they are.

- We made some tree-like diagrams (dendrograms) to show these groupings.

Part 2: Text Mining

What we did here:

- We took a set of sentences and used two techniques: Count Vectorizer and TF-IDF Vectorizer.
- Count Vectorizer counts how many times each word shows up.
- TF-IDF Vectorizer not only counts words, but also checks how unique these words are across all sentences. Basically, it helps us figure out which words are really important in a sentence.
- We showed these counts and TF-IDF scores in a table and talked about why TF-IDF is useful. It’s great for finding important words in a bunch of text, which can be handy for things like sorting documents or finding what a web page is about.

Conclusion

- By doing all this, we learned a lot about how movie titles are grouped and what words are important in our set of sentences.
- The clustering part (K-Means and Hierarchical) helped us see how movie titles can be grouped based on their names.

- The text mining part helped us understand what words stand out in sentences.
- All this is super useful for making sense of a lot of text data, like figuring out what a bunch of documents are about.