Question 1:

IMDB Movie Ratings Sentiment Analysis

Movies.csv

The dataset is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset.

The sentiment labels are:

0 - negative

1 - somewhat negative

2 - neutral

3 - somewhat positive

4 – positive

Tasks

1. **Read movies.csv.**
2. **Check if data is clean/clean**
3. **Convert the sentiments from the numerical numbers to their actual label.**
4. **Find the count of positive, negative and neutral sentiments etc.**
5. **Plot a countplot to show the count of sentiments.**
6. **Plot a piechart to show the distribution of the sentiments on the movies**
7. **Perform Sentiment Analysis. Which algorithm will you use? How accurate is it?**
8. **Create a csv for each Sentiment. How many positive and somewhat positive sentiments are there?**

Question2:

NLP ANALYSIS

Find out what this article is about, and generate a wordcloud from the frequent words.

text = '"best, good, better, nice, connectivity, connection, connecting I just wanted to find some really cool new places such as Seattle in November. I've never visited before but no luck here. Some of these suggestions are just terrible… I had to laugh! Most suggestions were just your typical big cities, restaurants and bars. Nothing off the beaten path here. I don't want to go these places

for fun. Totally not worth getting this, This was such a beautiful book. I wasn't even planning any travel when I came across this and just started flipping through the pages. I really like the cover and all the large glossy photographs in this book. John Smith did a wonderful job with the photography. I've found a perfect home for this on my coffee table. I'm planning a trip to Paris and Barcelona soon and I know this will come in handy. In the meantime, it's perfect for assisting this armchair traveler, As a traveler, I really appreciated reading about these great places to visit. The author takes you all over the world. Even with all the free information online these days, I find I'm taking this book with me wherever I go and using it to discover hidden gem, this product was not good, i love this product, this product was not as expected, i like this product was nice to use"'

California Independent Medical Review Dataset

Medical_Reviews.csv

This data is from the California Department of Managed Health Care (DMHC). It contains all decisions from Independent Medical Reviews (IMR) administered by the DMHC since January 1, 2001.

1. Read **Medical_Reviews.cv**
2. Clean Data
3. Drop all duplicates
4. Group data by Diagnosis Category, Diagnosis SubCategory and Patients Gender(Show the count).
5. Plot a barplot to show the relationship between Report year and Patients Gender.
6. **Bar plots** for categorical data distributions (diagnosis categories, treatment categories).

What are the most common diagnosis categories and sub-categories?

- Which diagnosis category appears most frequently?
- What are the top 10 most frequent diagnosis sub-categories?

How are determinations distributed?

- How many cases fall under each determination category?

Perform WebScrapping on the following link and Analyse the text data.

Find the moast common 10 words.

What are the main contents of this Article?

What do you think they are talking about?

Link: https://nation.africa/kenya/life-and-style/dn2/men-enjoy-a-genuine-public-display-of-love-4576356

class_="article-content xxx"

CHATGPT(Sentiment Analysis)

Chatgpt.csv

About Dataset

ChatGPT has been a major talk in the tech world. The tweets about chatgpt were gathered for a month and then the sentiment analysis was made using Natural Language Processing.

1. Perform Sentimental Classification on this chatgpt dataset
2. Read chatgpt.csv