Coding Steps and details.

We have selected 6 movies to analyse. We chose them for the given reasons……..

1. Get subtitles from Kaggle (get\_subtitles\_from\_kaggle.py)
   1. We download movie subtitle csv files for our selected movies from the Kaggle dataset "mlopssss/subtitles"
   2. We parse the downloaded text, removing timestamps, numbers, and a defined set of stop words.
   3. IMDB id’s are extracted for each movie from the OMDB website using their API.
   4. The cleaned movie subtitles, along with the movie name, year, award (oscar etc), imdb\_id are saved to a csv file (cleaned\_subtitles.csv)
2. Get imdb reviews from Kaggle (get\_imdb\_reviews\_from\_kaggle.py)
   1. We download movie imdb review csv files for our selected movies from the Kaggle dataset “mlopssss/imdb-movie-reviews-grouped-by-ratings”
   2. A custom CSV reader is used to parse the text to extract imdb\_id, user rating, user review.
   3. The cleaned and parsed text is saved to a csv file all\_imdb\_reviews.csv
3. Get screenplay text
   1. We downloaded the screenplay pdf files for each of our movies from the site ????ask Lai????.
   2. We then converted the pdf files to csv.
   3. We parsed the text in the csv files removing user defined stop words.
   4. A consolidated csv is created of the cleaned screenplay text along with movie name, imdb\_id, year, award. (cleaned\_screenplays.csv)
4. Generate Reviews from AI
5. Trigrams (trigram\_analysis.py)