DATA SCIENCE INSTITUTE

WOOLF

ASSIGNMENT

TEXT MINING AND NATURAL LANGUAGE PROCESSING

Movie Review Analysis

Python

REPORT

Pillar1-e

Prepared By

William C. Phiri

PG Dip in Data Science

Student No. 4295158639

+353-87-3502102

chenechoz@gmail.com

Athlone, IRELAND

BACKGROUND:

The given data contains short and crisp movie reviews by various critics.

QUESTIONS:

1. Import Textdata. Do the essential cleaning of the data.

2. Find top 20 words sorted by frequency.

3. Create a wordcloud using the given data.

4. List the number of lines having sentiments ‘Negative', 'Neutral’ and ‘Positive’.

5. Plot bar graph showing top 15 words by frequency.

GitHub Repository link to Code: Movie Review Analysis with Python

ASSIGNMENT SUMMARY OVERVIEW:

This project implements comprehensive Natural Language Processing (NLP) techniques to analyze movie review text data. The analysis includes text preprocessing, word frequency analysis, sentiment classification, and data visualization using industry-standard Python libraries.

Key Features:

- Text Data Cleaning with regex pattern matching and stopword removal
- Word Frequency Analysis identifying top 20 most common terms
- WordCloud Visualization showing word importance through size
- Sentiment Analysis using TextBlob polarity scoring
- Sentiment Classification into Positive, Negative, and Neutral categories
- Professional Bar Graphs for frequency distribution
- Comprehensive Tutorial for learning and replication

Dataset: **60** lines of professional movie reviews (**8,271** characters) covering multiple films including detailed critiques and analysis.

Analysis Type: Exploratory Text Analysis with focus on sentiment distribution and keyword extraction.

Project Folder Structure:

QUESTION 1: Import Textdata. Do the essential cleaning of the data.

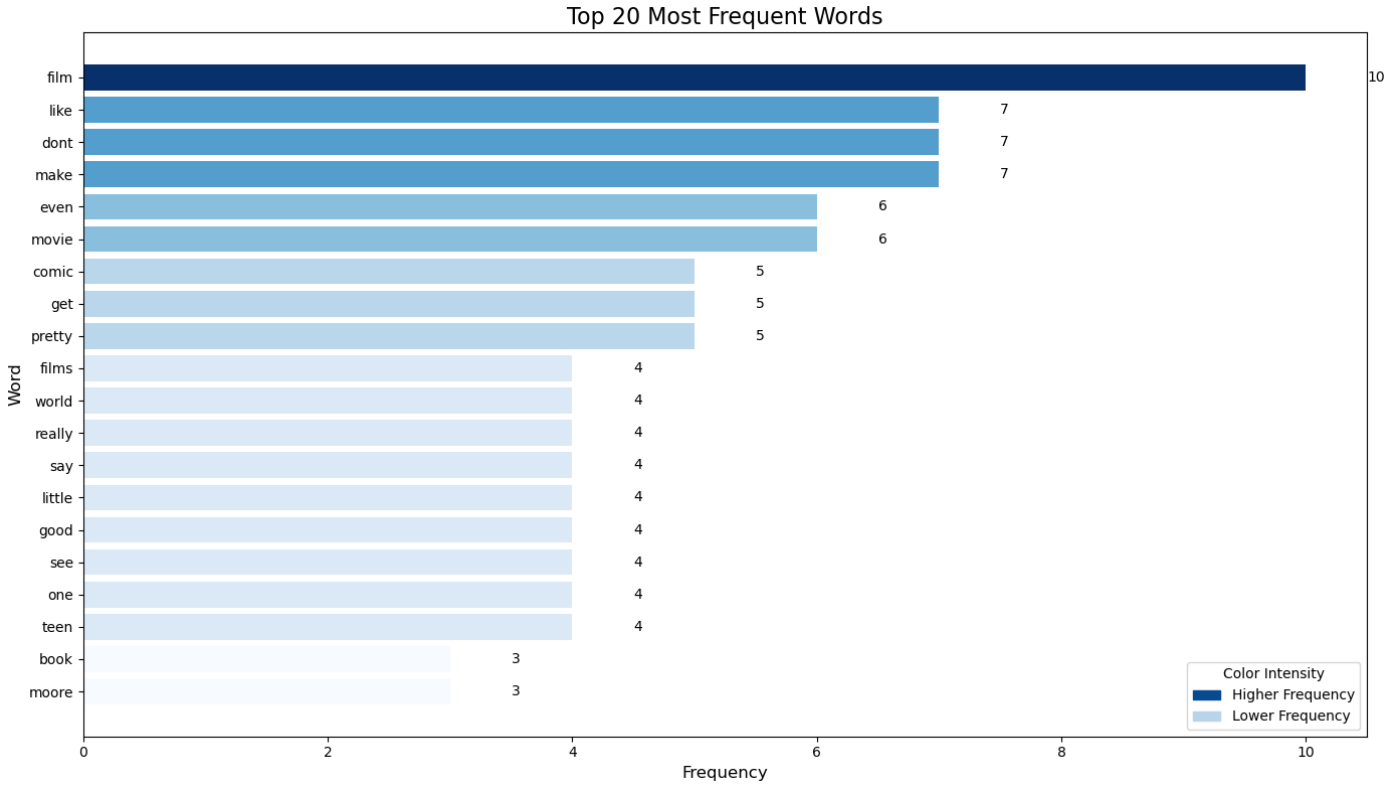
```
STEP 1: Import and Clean the Data
================================================================================

1.1 Reading the text file...
✓ File loaded successfully!
Total characters: 8271
First 200 characters:
films adapted from comic books have had plenty of success , whether they're about superheroes ( batman , superman , spawn ) , or geared

1.2 Splitting text into lines...
✓ Total lines: 60
Sample line: films adapted from comic books have had plenty of success , whether they're about superheroes ( batman , superman , spawn )
```

```
1.3 Creating text cleaning function...
✓ Text cleaned!
Original: films adapted from comic books have had plenty of success , whether they're about superheroes ( batman , superman , spawn ) , o
Cleaned:  films adapted from comic books have had plenty of success whether theyre about superheroes batman superman spawn or geared towa
```

```
1.4 Tokenizing and removing stopwords...
Number of stopwords: 198
Sample stopwords: ['now', "hadn't", 'each', 'll', 'into', 'ain', 'both', 'herself', "it'd", "we'd"]

✓ Tokenization complete!
Total words (after removing stopwords): 727
Sample words: ['films', 'adapted', 'comic', 'books', 'plenty', 'success', 'whether', 'theyre', 'superheroes', 'batman', 'superman', 'spawn', 'geared',
```

QUESTION 2: Find top 20 words sorted by frequency.

```
============================================================
STEP 2: Find Top 20 Words by Frequency
============================================================

Top 20 Most Frequent Words:
--------------------------------------------------
Rank   Word                 Frequency
--------------------------------------------------
1      film                 10
2      like                 7
3      dont                 7
4      make                 7
5      even                 6
6      movie                6
7      comic                5
8      get                  5
9      pretty               5
10     films                4
11     world                4
12     really               4
13     say                  4
14     little               4
15     good                 4
16     see                  4
17     one                  4
18     teen                 4
19     book                 3
20     moore                3

Total unique words: 551
```

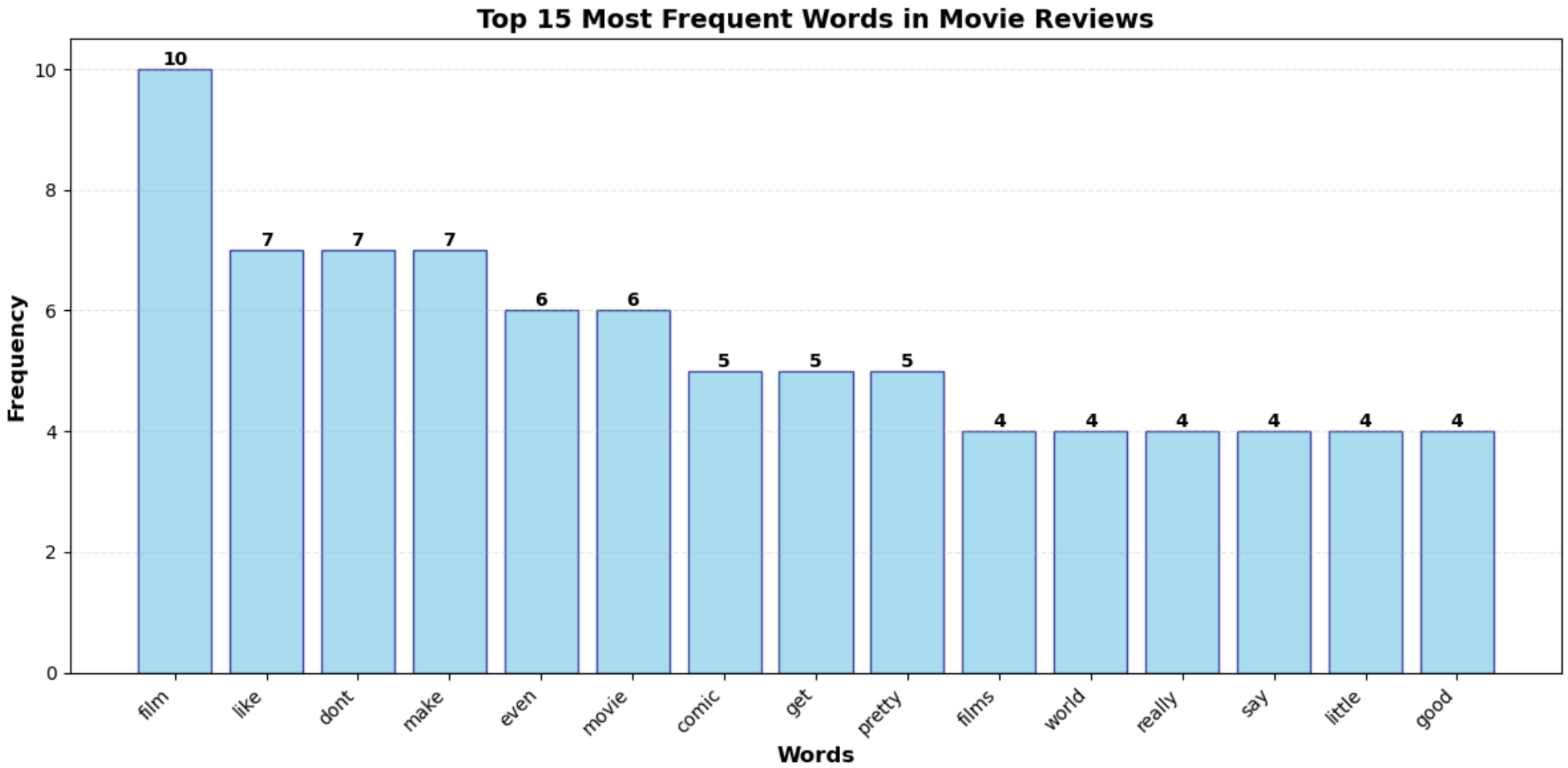


QUESTION 3: Create a wordcloud using the given data.

QUESTION 4: List the number of lines having sentiments ‘Negative', 'Neutral’ and ‘Positive’.

```
================================================================================
STEP 4: Sentiment Analysis - Count Positive, Negative, and Neutral Lines
================================================================================

Analyzing sentiment for each line...

==================================================
SENTIMENT ANALYSIS RESULTS
==================================================
Sentiment       Count      Percentage
--------------------------------------------------
Positive        23         38.3%
Negative        14         23.3%
Neutral         23         38.3%
--------------------------------------------------
Total           60         100.0%

==================================================
Sample Sentiment Classifications:
==================================================

Positive Examples:
  Line 1: films adapted from comic books have had plenty of success , whether they're abou...
  Polarity: 0.175

  Line 2: for starters , it was created by alan moore ( and eddie campbell ) , who brought...
  Polarity: 0.112


Negative Examples:
  Line 3: to say moore and campbell thoroughly researched the subject of jack the ripper w...
  Polarity: -0.130

  Line 5: in other words , don't dismiss this film because of its source .
  Polarity: -0.125


Neutral Examples:
  Line 8: the ghetto in question is , of course , whitechapel in 1888 london's east end .
  Polarity: 0.000

  Line 10: when the first stiff turns up , copper peter godley ( robbie coltrane , the worl...
  Polarity: 0.012

✓ Detailed sentiment analysis saved to: C:\Users\willi\GitHub\TMNLP_movie_review_txt_analysis_PY\reports\sentiment_analysis.csv
```

QUESTION 5: Plot bar graph showing top 15 words by frequency.



FINAL OVERVIEW:

Data Statistics:

- Total lines in file: 60

- Total words (after cleaning): 727

- Unique words: 551

Top 5 Most Common Words:

1. 'film' - 10 times
2. 'like' - 7 times
3. 'dont' - 7 times
4. 'make' - 7 times
5. 'even' - 6 times

Sentiment Distribution:

- Positive: 23 lines (38.3%)

- Negative: 14 lines (23.3%)

- Neutral: 23 lines (38.3%)

Output Files Created:

1. wordcloud.png - Visual representation of word frequencies
2. sentiment_analysis.csv - Detailed sentiment analysis for each line
3. top_15_words_bargraph.png - Bar graph of most frequent words