Homework 6: IMDB Text Mining & Sentiment Analysis

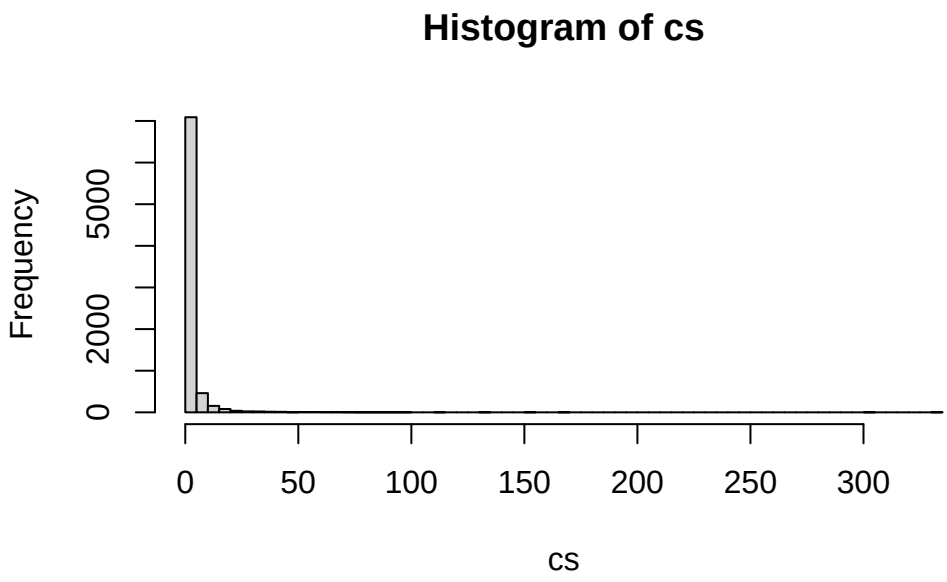
Yiming Cao, Sujan Kakumanu, Angel Sanaa Rutherford

2025-12-12

```
 wonderful little production    filming technique   unassuming  oldtimeBBC fashion  gives  c
```

```
<<DocumentTermMatrix (documents: 199, terms: 7937)>>
Non-/sparse entries: 21018/1558445
Sparsity           : 99%
Maximal term length: 32
Weighting          : term frequency (tf)
Sample             :
     Terms
Docs  even film good just like movie one see story this
  102    5   14    1    1    1     1   0   0     5    3
  157    1    5    0    2    2     0   1   1     1    0
  173    3    0    0    4    5     5   4   3     0    4
  178    4    3    1    1    2     1   3   1     1    0
  192    0    2    0    0    2     0   1   0     2    4
  30     2    4    0    0    0     3   0   0     3    0
  34     1    8    0    1    2     0   5   0     0    1
  49     1    3    0    0    2     1   2   0     2    3
  52     0    0    2    1    3     0   0   0     1    0
  59     0    4    2    0    1     0   2   0     2    1
```

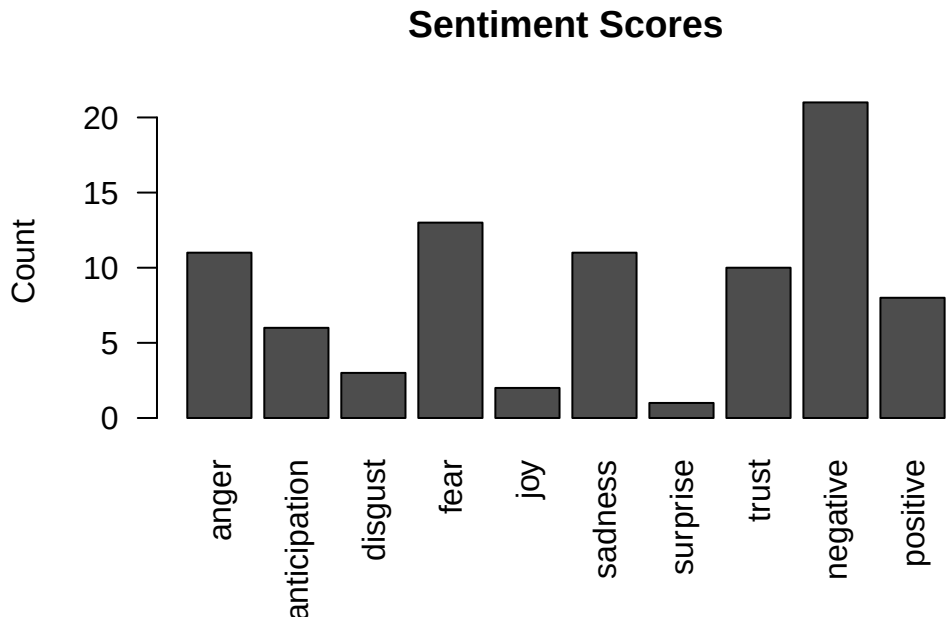
```
[1]  199 7937
```




```
nrc <- syuzhet::get_sentiment_dictionary(dictionary="nrc")
head(nrc, n=20L)
```

```
      lang           word sentiment value
```

1	english	abba	positive	1
2	english	ability	positive	1
3	english	abovementioned	positive	1
4	english	absolute	positive	1
5	english	absolution	positive	1
6	english	absorbed	positive	1
7	english	abundance	positive	1
8	english	abundant	positive	1
9	english	academic	positive	1
10	english	academy	positive	1
11	english	acceptable	positive	1
12	english	acceptance	positive	1
13	english	accessible	positive	1
14	english	accolade	positive	1
15	english	accommodation	positive	1
16	english	accompaniment	positive	1
17	english	accomplish	positive	1
18	english	accomplished	positive	1
19	english	accomplishment	positive	1
20	english	accord	positive	1



1 Introduction (Angel)

In this analysis, we performed text-mining techniques to movie reviews from the Internet Movie Database (IMDb) in order to quantify and visualize word trends and emotional tones across reviews. Text-mining combines data cleaning and language processing techniques, enabling researchers to systematically analyze unstructured text for meaningful patterns such as term frequency and emotional sentiments. This approach combines the ease of decreased manual effort with nuance in understanding narratives and perspectives.

2 Methods

2.1 Data Preprocessing

We began by importing the IMDb dataset csv into R and converting the review column into a corpus object using the `tm` package. The `VectorSource` function was called in order to treat every review as a separate document. The result was a corpus which, in this context, streamlines analysis by serving as a repository of the text documents. The corpus was then preprocessed to ensure uniformity and remove noise. Specifically, all entries were transformed to lowercase, numbers and punctuation were removed, and common English stopwords were excluded. In addition, a small set of self-defined stop and non-english words ("I", "br", "You,", "The", "A", "It") were removed after additional data exploration to further reduce noise.

2.2 Word Cloud Creation (Angel)

After preprocessing we created a document term matrix (DTM) which represents the frequency of terms across all documents. In this matrix, each row corresponds with a document, each column corresponds to a unique term, and the cell values represent the number of times the the term itself appears in a given document.

From this DTM, two visuals were created to represent the frequency of terms: a histogram and a word cloud. In a word cloud visualization, words are displayed in font sizes proportional to the their frequency, allowing words repeated more frequently to be more visible. We used the `wordcloud` function to create our visual which takes a specified threshold for which words to display based on frequency. In analysis, we chose to only display words that appeared more than 500 times to ensure variation but also to reduce noise.

2.3 Sentiment Analysis (Ming)

3 Results

3.1 World Cloud (Ming)

3.2 Sentiment Analysis (Sujan)

4 Discussion (Sujan)