DATA SCIENCE INSTITUTE

WOOLF

ASSIGNMENT

TEXT MINING AND NATURAL LANGUAGE PROCESSING

Movie Review Analysis

R

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 contain short and crisp movie reviews by various critics.

QUESTIONS:

1. Import **Textdata.** Do the essential cleaning of the data.
2. Find words with minimum frequency 6.
3. List words with at least 0.35 correlation with 'film'.
4. Create a **wordcloud** with words having minimum frequency 4. (Use any palette from **RColorBrewer**)
5. List the number of lines having sentiments 'Sarcasm', 'Very Negative' and 'Very Positive'.
6. Plot graph showing words occurring more than 3 times (Use tidytext package).

GitHub Repository link to Code: Movie Review Analysis with R

ASSIGNMENT SUMMARY OVERVIEW:

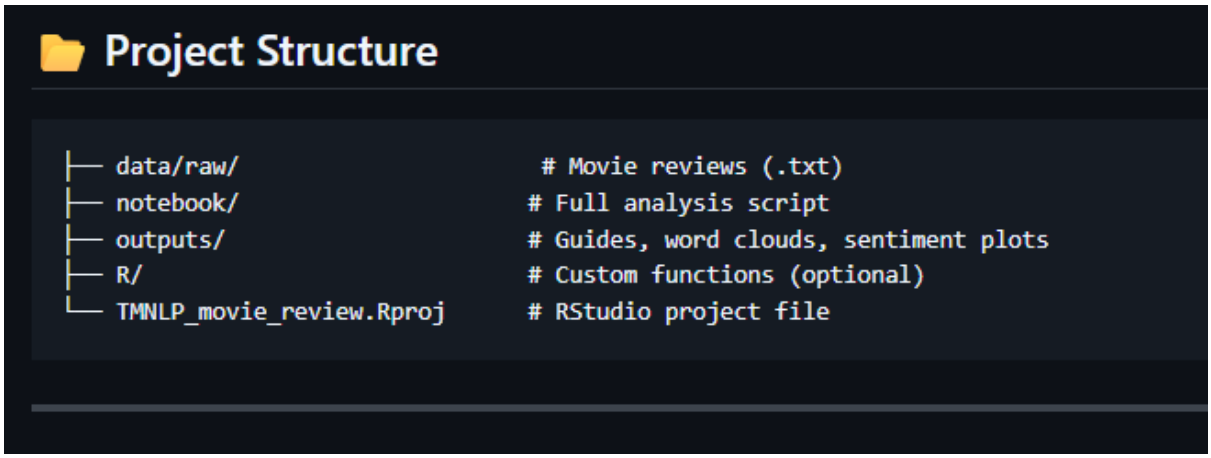
Overview

This project applies **Text Mining and Sentiment Analysis in R** to explore movie reviews. It uses both traditional (tm) and modern (tidytext) NLP techniques to uncover themes, sentiments, frequent words, and emotional tones within the text.

Features:

- **Text Cleaning** – Stopwords, punctuation, numbers, whitespace
- **Document-Term Matrix (DTM)**
- **Word Frequency & Correlation Analysis**
- **Word Cloud Generation**
- **Sentiment Analysis (NRC Emotion Lexicon)**
- **ggplot2 Visualizations**
- **Modular & Reproducible Code**

PROJECT FOLDER STRUCTURE:



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

```
=== QUESTION 1: Importing and Cleaning Data ===
First 3 lines of raw data:
[1] "films adapted from comic books have had plenty of success , whether they're about s
uperheroes ( batman , superman , spawn ) , or geared toward kids ( casper ) or the artho
use crowd ( ghost world ) , but there's never really been a comic book like from hell be
fore . "
[2] "for starters , it was created by alan moore ( and eddie campbell ) , who brought th
e medium to a whole new level in the mid '80s with a 12-part series called the watchmen
. "
[3] "to say moore and campbell thoroughly researched the subject of jack the ripper woul
d be like saying michael jackson is starting to look a little odd . "

Total number of lines: 61

Corpus created with 61 documents
✓ Converted to lowercase
✓ Removed numbers
✓ Removed punctuation
✓ Removed stop words
✓ Removed extra whitespace

Sample of cleaned text:
 [1] "films adapted comic books plenty success whether theyre superheroes batman superma
n spawn geared toward kids casper arthouse crowd ghost world theres never really comic b
ook like hell "
 [2] NA
 [3] NA
 [4] NA
 [5] NA
 [6] NA
 [7] NA
 [8] NA
 [9] NA
[10] NA
```

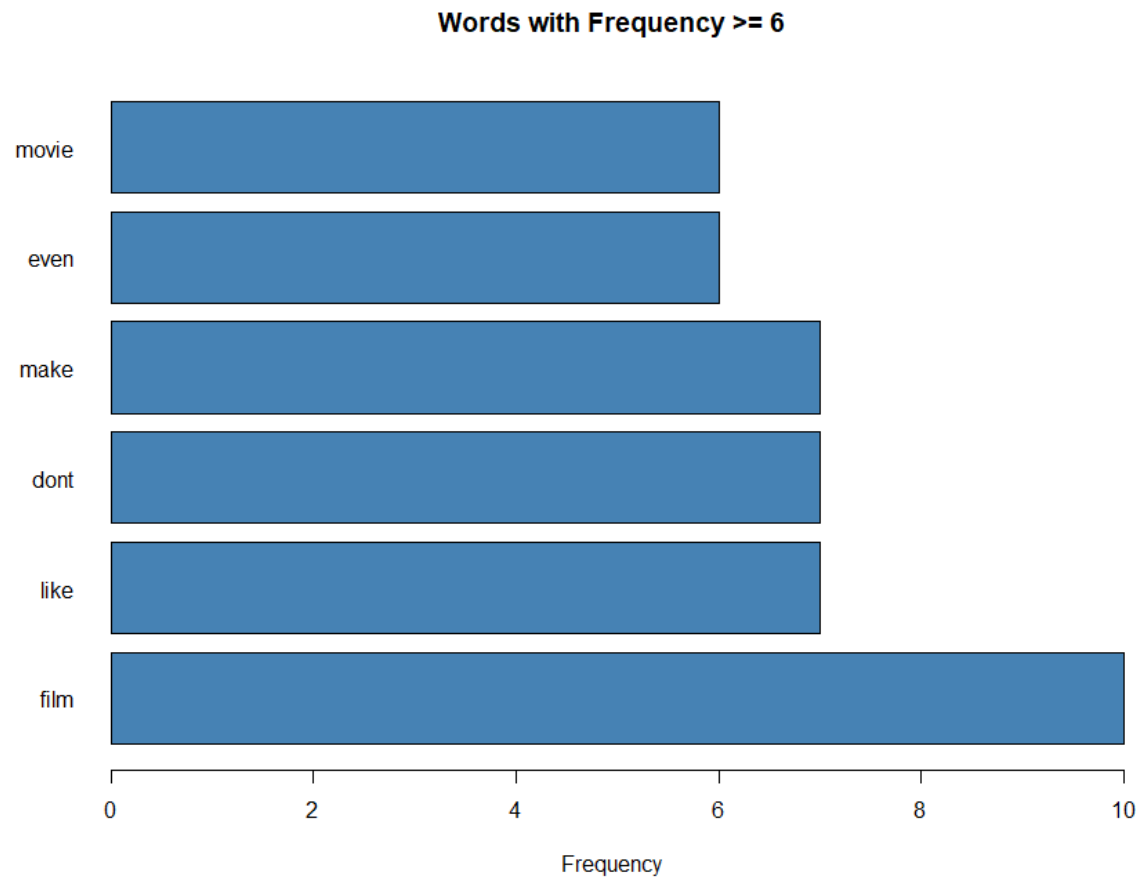
Create the Document Term Matrix(DTM)

```
Document-Term Matrix created:
  Documents (lines): 61
  Terms (unique words): 553
  Sparsity: 97.86 %
```

QUESTION 2: Find words with minimum frequency 6.

Words appearing at least 6 times:

Word	No. of Times Appearing
film	10
like	7
don't	7
make	7
even	6
movie	6

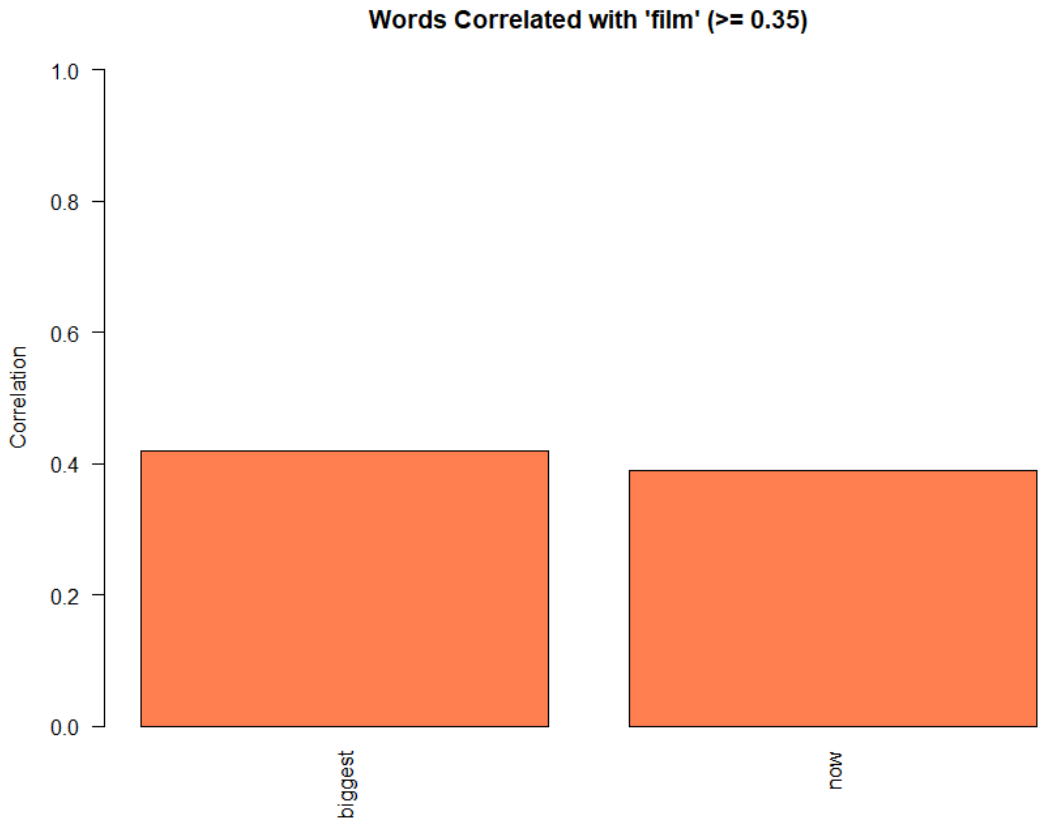


QUESTION 3: List words with at least 0.35 correlation with ‘film’.

Words with correlation >= 0.35 with 'film':

I. biggest: **0.42**

II. now: **0.39**

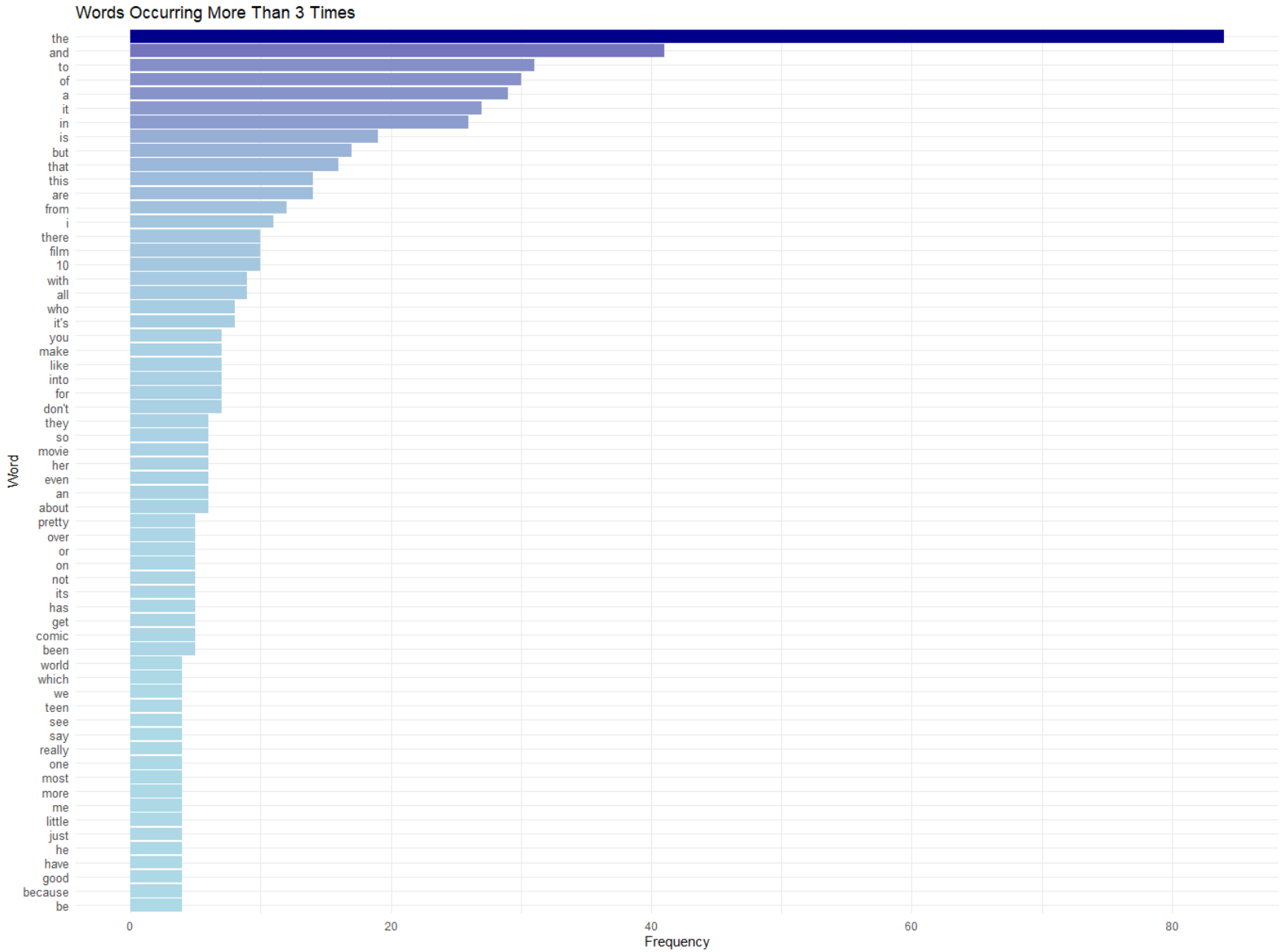


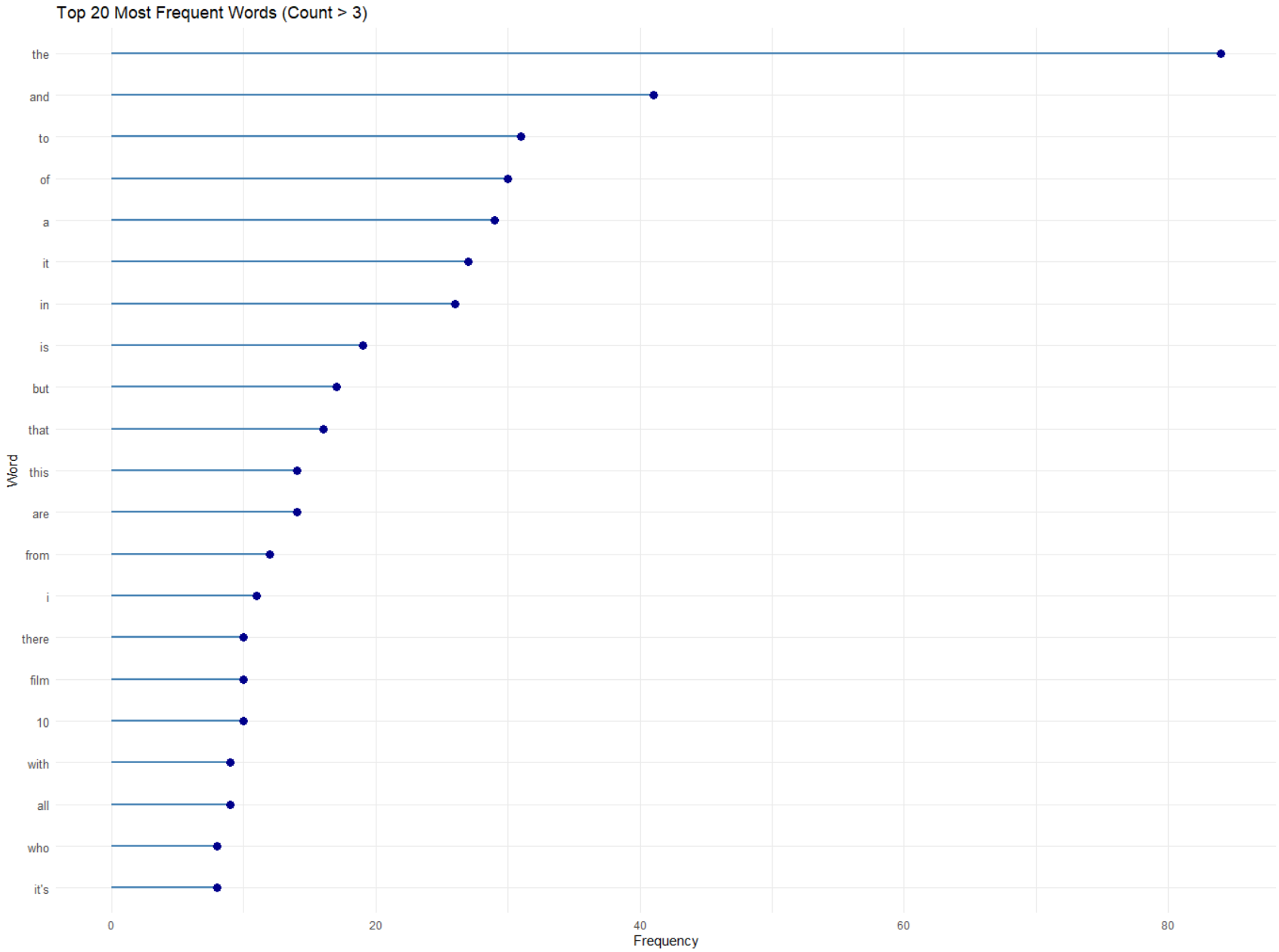
QUESTION 4: Create a **wordcloud** with words having minimum frequency 4. (Use any palette from **RColorBrewer**)

QUESTION 5: List the number of lines having sentiments ‘Sarcasm’, ‘Very Negative’ and ‘Very Positive’.

```
=== QUESTION 5: Sentiment Analysis ===
Total words in tidy format: 1402
NRC lexicon loaded with 13872 word-sentiment pairs

Available sentiments in NRC:
 [1] "anticipation" "joy"          "positive"     "trust"        "fear"
 [6] "anger"        "disgust"      "negative"     "sadness"      "surprise"

=== SENTIMENT COUNTS BY LINE ===
Lines with NEGATIVE sentiment: 31
Lines with POSITIVE sentiment: 36
Lines with strong negative emotions (anger/disgust/fear/sadness): 32
Lines with strong positive emotions (joy/trust/anticipation): 40

Note: NRC lexicon doesn't detect 'Sarcasm' directly
      Sarcasm detection requires more advanced NLP techniques

=== OVERALL SENTIMENT DISTRIBUTION ===
      sentiment  n
1      positive 63
2      negative 56
3  anticipation 42
4          fear 35
5       sadness 33
6         trust 32
7           joy 31
8         anger 22
9      surprise 22
10      disgust 21
```

1. **QUESTION 6:** Plot graph showing words occurring more than 3 times (Use tidytext package).




Top 20 Most Frequent Words (Count > 3)
the
and
to
of
a
it
in
is
but
that
this
are
from
i
there
film
10
with
all
who
it's
Word
0
20
40
60
80
Frequency