

Homework 6: IMDB Text Mining & Sentiment Analysis

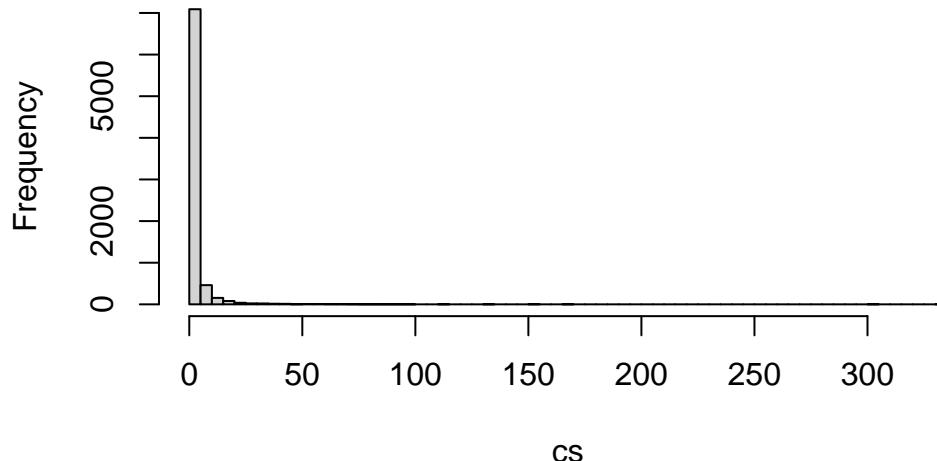
Yiming Cao, Sujan Kakumanu, Angel Sanaa Rutherford

2025-12-12

```
wonderful little production    filming technique    unassuming oldtimeBBC fashion gives co  
<<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

Histogram of cs



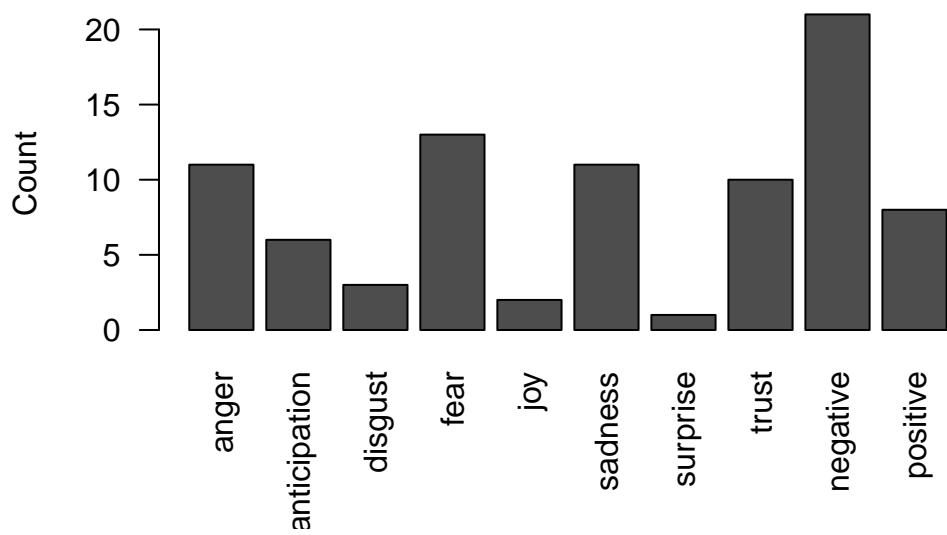
camcorder tack authority hypothetically flourish waybr christopher barrel wwi maltin
 persons substance lived threatens @ highbudget italian limbalsamato meitijs book
 trail pile smoking@ tea skeleton kingdom beginning smarted bonejarringly trini
 interwoven zone brenner troubld dependentunconvincingly anchorman
 armies girls zone fo@ joel heston kickingbrindustry evan observers tarzan
 called haunts foot@ watching countrybr attention
 position affcted neofascists @ utterly ups morty his dutcher carpenter anythingbrcoins accounts secondbr lou
 andor stuck rests@ seniorsequence fulfillment @ clumsyjacobs afterwards
 drawing crime wrong blethys
 think
 andor
 drawing turn@ exciting but sin moody graders route bbr breasts
 mistakebr waiteddaltonforgive matrixormina already
 novaks ship oddlyphysical episodes halfway offendingiraq
 cried driver bleep fibbrush hurriedly lambert sit black cravingboys
 lukewarm grahamie coaster minx staticsoldier atwoods ogrodnikripeil
 irritatinghearted nearly appealing take knocking endednotreviewer ali skin klux
 koslocable touching elbride hasn't degressbr vinson planetbr alvarado adults
 state corrupt stardom cut kidapsdeposedidj guilt associated grandma shake
 stare hes actionlevels tom screwed rooting speak til naff civil decline fishesfests
 bussey gary flower tom rolebrmalevolent swamps ronald vical battles reagan seamless
 stomachbr gaffseagle list involvesii brazilian wander vical wed initially insistsnazi
 shatner music partners talk > may sees larch
 onbr inc professor levels may tanner insomnia
 rex not real nerds bard tour hong shak count tanner insomnia
 hell connected homebr minutes blendedshockill
 rumors fat bug scrub tow displaying pare charis lit contain trulyifscut
 backs doctor legion dogmatically death hero defined sex gallo laughter revealing
 logan conform @ repercussions marshazimbalist world sex dials
 roofromp restlessness words seber raised buts discussing
 hat enabled separate companies geologists didos
 snow ole negative scenarios dresses experimental age
 dimly struggling clear
 third charm expect getting nominated philisophybr clear
 barsi charm @ empathizing target facial hoiy skerritdee
 roots struglightning map binnie knights thumb double feels department
 middle servant werewolf updates spirituality andy double feels department
 referenced stationbr epicsmarmy reaching front greek sailings investigations
 letter aviation emeralde preparationkin furthermore double feels
 aftermath lukephotographed speechless torn jail superb favored intact memory space
 fluid hysterically rubbery gloomy student vcr wishes broadbentcgi
 formulas remarkable boomdeceptive pursued electroshock system they
 calls ian fresh affairbr repeated homoeroticism wifff dynamics explosive baddies

```
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

Sentiment Scores



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 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 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 Word Cloud (Ming)

3.2 Sentiment Analysis (Sujan)

4 Discussion (Sujan)