Solutions!

# 1. Top Twenty most Common Words

We examine the Top Twenty Most Common words and show them in a bar graph.

Hide

SC = ScriptsCharacters **%>%**

**select**(id,name,normalized\_text)

SC **%>%**

**unnest\_tokens**(word, normalized\_text) **%>%**

**filter**(**!**word **%in%** stop\_words**$**word) **%>%**

**count**(word,sort = TRUE) **%>%**

**ungroup**() **%>%**

**mutate**(word = **factor**(word, levels = **rev**(**unique**(word)))) **%>%**

**head**(20) **%>%**

**ggplot**(**aes**(x = word,y = n)) **+**

**geom\_bar**(stat='identity',colour="white", fill =fillColor) **+**

**geom\_text**(**aes**(x = word, y = 1, label = **paste0**("(",n,")",sep="")),

hjust=0, vjust=.5, size = 4, colour = 'black',

fontface = 'bold') **+**

**labs**(x = 'Word', y = 'Word Count',

title = 'Top 20 most Common Words') **+**

**coord\_flip**() **+**

**theme\_bw**()

## 2. WordCloud of the Common Words

A word cloud is a graphical representation of frequently used words in the **Normalized** text. The height of each word in this picture is an indication of frequency of occurrence of the word in the entire text.  
**im , dont , hey and homer** are some of the most commonly occuring terms.

Hide

SC **%>%**

**unnest\_tokens**(word, normalized\_text) **%>%**

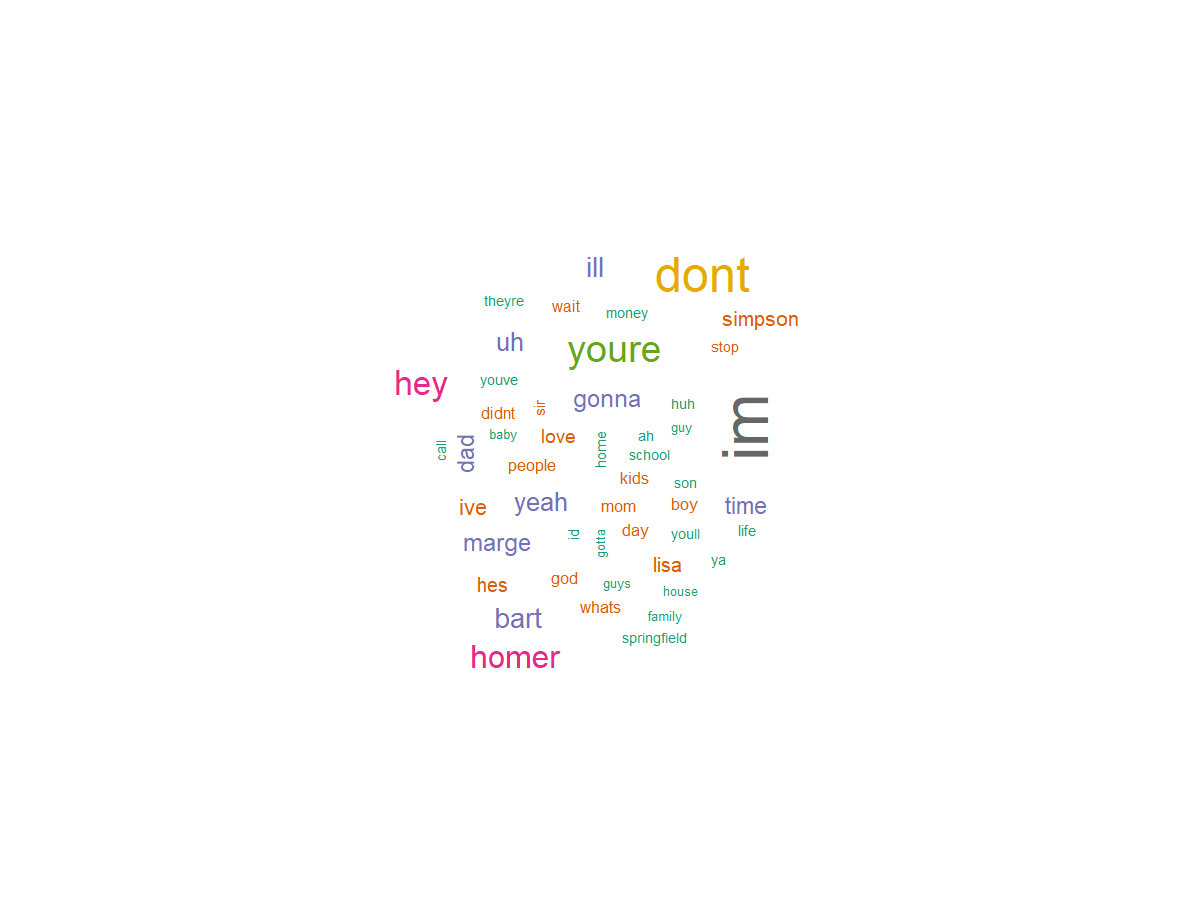
**filter**(**!**word **%in%** stop\_words**$**word) **%>%**

**count**(word,sort = TRUE) **%>%**

**ungroup**() **%>%**

**head**(50) **%>%**

**with**(**wordcloud**(word, n, max.words = 50,colors=**brewer.pal**(8, "Dark2")))



## 3. Sentiment analysis

visualize\_sentiments <- **function**(SCWords) {

SCWords\_sentiments <- SCWords **%>%**

**inner\_join**(**get\_sentiments**("afinn"), by = "word") **%>%**

**group\_by**(name) **%>%**

**summarize**(score = **sum**(score **\*** n) **/** **sum**(n)) **%>%**

**arrange**(**desc**(score))

SCWords\_sentiments **%>%**

**mutate**(name = **reorder**(name, score)) **%>%**

**ggplot**(**aes**(name, score, fill = score **>** 0)) **+**

**geom\_col**(show.legend = TRUE) **+**

**coord\_flip**() **+**

**ylab**("Average sentiment score") **+** **theme\_bw**()

}

TopCharacters = ScriptsCharacters %>%

group\_by(name) %>%

tally(sort = TRUE)

Top20Characters = **head**(TopCharacters,20)**$**name

SCWordsTop20Characters <- SC **%>%**

**unnest\_tokens**(word, normalized\_text) **%>%**

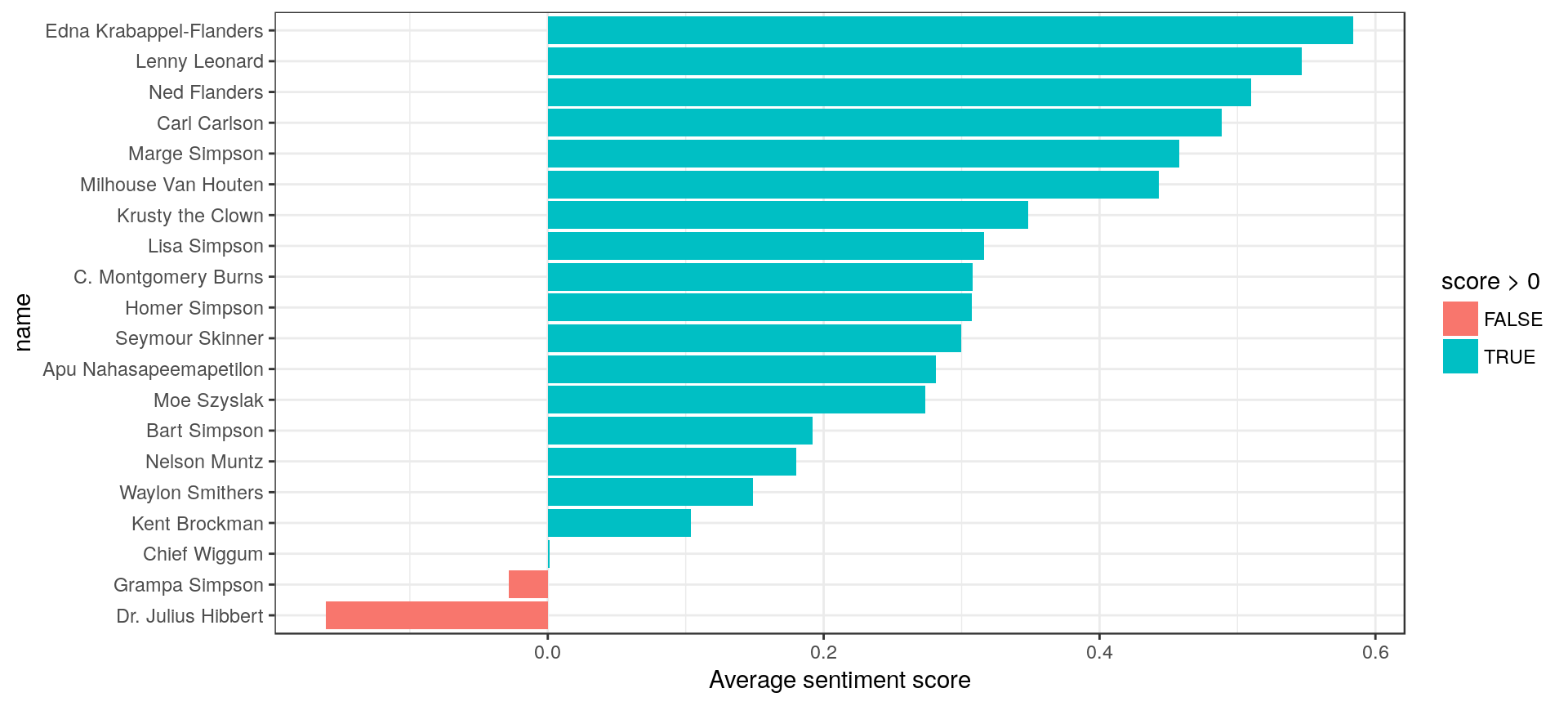
**filter**(name **!=** "NA") **%>%**

**filter**( name **%in%** Top20Characters) **%>%**

**count**(name, word, sort = TRUE) **%>%**

**ungroup**()

**visualize\_sentiments**(SCWordsTop20Characters)



# 4. Topic Modelling

SC = ScriptsCharacters **%>%**

**select**(id,name,normalized\_text)

corpus = **Corpus**(**VectorSource**(SC**$**normalized\_text))

*# Pre-process data*

corpus <- **tm\_map**(corpus, tolower)

corpus <- **tm\_map**(corpus, removePunctuation)

corpus <- **tm\_map**(corpus, removeWords, **stopwords**("english"))

corpus <- **tm\_map**(corpus, removeWords, UniqueLowIDF[1**:**500])

corpus <- **tm\_map**(corpus, stemDocument)

dtm = **DocumentTermMatrix**(corpus)

*# Remove sparse terms*

dtm = **removeSparseTerms**(dtm, 0.997)

*# Create data frame*

labeledTerms = **as.data.frame**(**as.matrix**(dtm))

labeledTerms = labeledTerms[**rowSums**(**abs**(labeledTerms)) **!=** 0,]

##############################################################################

*#LDA Modelling Starts*

###############################################################################

*# set a seed so that the output of the model is predictable*

simpsons\_lda <- **LDA**(labeledTerms, k = 2, control = **list**(seed = 13))

*#The tidytext package provides this method for extracting the per-topic-per-word probabilities,*

*# called β (“beta”), from the model*

simpsons\_topics <- **tidy**(simpsons\_lda, matrix = "beta")

simpsons\_top\_terms <- simpsons\_topics **%>%**

**group\_by**(topic) **%>%**

**top\_n**(10, beta) **%>%**

**ungroup**() **%>%**

**arrange**(topic, **-**beta)

simpsons\_top\_terms **%>%**

**mutate**(term = **reorder**(term, beta)) **%>%**

**ggplot**(**aes**(term, beta, fill = **factor**(topic))) **+**

**geom\_col**(show.legend = FALSE) **+**

**facet\_wrap**(**~** topic, scales = "free") **+**

**coord\_flip**() **+** **theme\_bw**()

