| **title** | **author** | **date** | **output** |
| --- | --- | --- | --- |
| Assignment 3 - NLP | Charles Lang | 2/23/2017 | html\_document |

**Libraries**

#Make sure you install and load the following libraries

library(tm)

library(SnowballC)

library(wordcloud)

library(ggplot2)

library(dplyr)

library(tidyr)

library(topicmodels)

#IF USING A MAC PLEASE RUN THIS CODE

Sys.setlocale("LC\_ALL", "C")

**Import all document files and the list of weeks file**

#Create a list of all the files

file.list <- list.files(path="~/YOUR FILE PATH", pattern=".csv")

#Loop over file list importing them and binding them together

D1 <- do.call("rbind", lapply(file.list, read.csv, header = TRUE, stringsAsFactors = FALSE))

D2 <- read.csv("~/YOUR FILE PATH/week-list.csv", header = TRUE)

**Clean the htlm tags from your text**

**Merge with week list so you have a variable representing weeks for each entry**

**Process text using the tm package - Code has been altered to account for changes in the tm package**

#Convert the data frame to the corpus format that the tm package uses

corpus <- Corpus(VectorSource(D1$Notes2))

#Remove spaces

corpus <- tm\_map(corpus, stripWhitespace)

#Convert to lower case

corpus <- tm\_map(corpus, content\_transformer(tolower))

#Remove pre-defined stop words ('the', 'a', etc)

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

#Convert words to stems ("education" = "edu") for analysis, for more info see http://tartarus.org/~martin/PorterStemmer/

corpus <- tm\_map(corpus, stemDocument, lazy=TRUE)

#Remove numbers

corpus <- tm\_map(corpus, removeNumbers, lazy=TRUE)

#remove punctuation

corpus <- tm\_map(corpus, removePunctuation, lazy=TRUE)

**Create a Term Document Matrix**

#Convert corpus to a term document matrix - so each word can be analyzed individuallly

tdm.corpus <- TermDocumentMatrix(corpus)

**Sentiment Analysis**

**Match words in corpus to lexicons of positive & negative words**

#Upload positive and negative word lexicons

positive <- readLines("positive-words.txt")

negative <- readLines("negative-words.txt")

#Search for matches between each word and the two lexicons

D1$positive <- tm\_term\_score(tdm.corpus, positive)

D1$negative <- tm\_term\_score(tdm.corpus, negative)

#Generate an overall pos-neg score for each line

D1$score <- D1$positive - D1$negative

**Generate a graph of the sum of the sentiment score over weeks**

**LDA Topic Modelling**

#Term Frequency Inverse Document Frequency

dtm.tfi <- DocumentTermMatrix(corpus, control = list(weighting = weightTf))

#Remove very uncommon terms (term freq inverse document freq < 0.1)

dtm.tfi <- dtm.tfi[,dtm.tfi$v >= 0.1]

#Remove non-zero entries

rowTotals <- apply(dtm.tfi , 1, sum) #Find the sum of words in each Document

dtm.tfi <- dtm.tfi[rowTotals> 0, ] #Divide by sum across rows

lda.model = LDA(dtm.tfi, k = 3, seed = 150)

#Which terms are most common in each topic

terms(lda.model)

#Which documents belong to which topic

topics(lda.model)

**Main Task**

Your task is to generate a *single* visualization showing:

* Sentiment for each week and
* One important topic for that week