# Mini-Project 1 - MTH 461

## Contents



**Instructions:**

- Please provide complete solutions for each problem. If it involves mathematical computations, explanations, or analysis, please provide your reasoning or detailed solutions.
- Note that some problems have multiple solutions or ways to solve it. Make sure that your solutions are clear enough to showcase your work and understanding of the material.
- Creativity and collaborations are encouraged. Use all of the resources you have and what you need to complete the module. Each student must take personal responsibility and submit their work individually. Please abide by the University of Portland Academic Honor Principle.
- There are two ways you can write your answers, a: by handwriting (either physically or digitally), or b: by typing on a template document with file type options, which can be downloaded from the course website.
- If you had handwritten your answers/solutions on a physical paper, make sure to label it properly and please scan your document using a scanner app for convenience. Suggestions: (1) "Tiny Scanner" for Android or (2) "Scanner App" for iOS.
- **Please save your work as one pdf file, don't put your name in any part of the document, and submit it to the Teams Assignments for this course. Your document upload will correspond to your name automatically in Teams.**
- If you have questions or concerns, please feel free to ask the instructor.

**R Packages:**

- Below are pre-loaded general packages required for this module assignment. You can load more packages here or throughout the module if necessary.
- Note that you need to install R packages before you can use them. You can use the `install.packages()` in the R console, or go to the "Tools" tab and click "Install Packages..." in R Studio.
- Be careful on loading R packages because sometimes any two packages can have conflicting functions when calling them.

```
# pre-load packages here
library(tidyverse)
library(textdata)
library(tidytext)
```

# Linguistic Patterns in Text Data Sets

## Sentiment Analysis of Sentences

## Data Sets and Lexicons

1. **Sentence Polarity Dataset v1.0**

This text data set is a collection of 5331 positive and 5331 negative labeled sentences. (Pang & Lee, 2005). This data set can be downloaded from Cornell - Sentence Polarity Movie Reviews.

```
# load raw data sets of positive and negative sentences
data_pos <- readLines("rt-polaritydata/rt-polarity.pos")
data_neg <- readLines("rt-polaritydata/rt-polarity.neg")

# convert the raw data to R data frame
df_sent_polar <- tibble(sentiment = rep(c("P","N"),each=length(data_pos)),
                        sentence = c(data_pos,data_neg))
```

2. **NRC Emotion Lexicon**

The `lexicon_nrc_eil()` lexicon from the `textdata` package, which is the sentiment and emotions lexicon from the National Research Council (NRC) of Canada. The NRC lexicon is a data set of categorized words according to there associated human emotions. The `lexicon_nrc_eil()` function loads a set of words with associated emotions in 4 categories with a numerical variable measuring its intensity (Mohammad, 2018). See the Sentiment and Emotion Lexicon for more information.

```
# load NRC emotion lexicons
emotions_4_with_intensity <- textdata::lexicon_nrc_eil()
```

3. **Stopwords Lexicon**

The `stop_words` list from the `tidytext` package, which is the stopword list retrieved from the SMART lexicon for text categorization benchmarking as explained by Lewis et al. (2004).

```
# load the stopwords lexicon
swd_list <- tidytext::stop_words %>%
  filter(lexicon == "SMART") %>%
  select(word)
swd_list <- swd_list$word
```

## Exercises

**Categorizing Emotions in News Topics**

**Load Data Sets**

**Exercises**

## References

Lewis, D. D., Yang, Y., Russell-Rose, T., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, *5*(Apr), 361–397.

Mohammad, S. M. (2018). Word affect intensities. *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the ACL*.