MIS 506 – Module 2 – Assignment 1
Pre-Processing Text for Quantitative Text Analytics

In this assignment you will complete a variety of tasks related to cleaning and preprocessing text.

I recommend creating a new RStudio project for every assignment and for each lecture as you follow-along. Using a good directory structure will make it much easier for you to find your work later. For this assignment, you might create a directory for all Module 2 work. Within this directory, create a folder called “Cleaning Text” (or similar). Create an R Project (with an appropriate name) in this folder.

Submission Instruction:

You have two options:

1. Do all the work in an R Markdown document. Knit your document into a Word file and submit the Word file as the deliverable for this assignment.
2. Compress (or zip) your entire project folder (Module 2 - Assignment1) and upload the file in canvas.

Deliverables:

One: Read-in one of the following data sets as a data frame (data sets are available on Canvas). To do this follow the Module 2 lecture notes.

- ***Dataset 1:*** Amazon Book Reviews
- ***Dataset 2:*** Prince Lyrics
- ***Dataset 3:*** Artist Songs

Answer the following questions about the dataset:

a) Describe this dataset.
b) How many variables and observations are there?
c) Which column contains the text data that you are going to analyze?

Two: Tokenize the text and transform it to a tidy data structure and count the most popular words in the text.

Three: Prepare the text for analysis by removing the stop words, undesirable words, numbers, whitespaces, special characters, and any other necessary steps.

Four: Again count the most popular words in the text.

Five: Create a visualization of the most common words in the text and explain your results.