M8L1 Homework Assignment (due 12th November)

This homework assignment focuses on Text Mining theory. You will provide a written analysis based on the following information:

Consider the text from Dr. Seuss below:

"*You have brains in your head. You have feet in your shoes. You can steer yourself in any direction you choose. You're on your own, and you know what you know. And you are the guy who'll decide where to go.*"

Perform the following tasks (either by hand or in R):

1. Create a term by document matrix for the Dr. Seuss quote. Assume each sentence is a new document.
2. Calculate the td-idf for three terms in the text. Assume each sentence is a new document.
3. Write a regular expression to segment the Dr. Seuss quote in to seperate sentences.
4. Write a regular expression to tokenize the Dr. Seuss quote.
5. Create a frequency signature for the Dr. Seuss quote. Assume each sentence is a new document.

As a reminder, please provide a written analysis/report as an .Rmd file. Note: This is a graded assignment due by Sunday at 11:59 pm

M8L2 Homework Assignment (due 12th November)

This homework assignment focuses on Regular Expressions. You will provide a written analysis based on the following information:

Create regular expressions for the patterns below:

- Match any of the following punctuation characters in the ASCII table: !“#$%&’()+,
- Create one regular expression to match all common misspellings of calendar (see https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/C)
- Create one regular expression to match any character except line breaks.
- You need to validate a ZIP code (U.S. postal code), allowing both the five-digit and nine-digit (called ZIP+4) formats. The regex should match 02115 and 02115-5515, but not 2115, 2115-5515, 21155515,021155515, etc.
- You need to validate a legit any password for your website. Passwords have the following complexity requirements: Length between 8 and 32 characters, ASCII visible and space characters only, One or more uppercase letters, One or more lowercase letters, One or more special characters (ASCII punctuation)
- Load the file ML.Tweets.csv (it is online at ‘http://nikbearbrown.com/YouTube/MachineLearning/Twitter/’)
- Complete the following:
  - Extract a list of the top 9 users (e.g. @NikBearBrown)
  - Extract a list of the top 9 hashtags (e.g. #Bear)
  - Find the top 5 most positve tweets
  - Find the top 5 most negative tweets
  - Create a world cloud of 100 related tweets
  - Which tweets could be classified as game development?

As a reminder, please provide a written analysis/report as an .Rmd file. Note: This is a graded assignment due by Sunday at 11:59 pm.

M8L3 Homework Assignment (due 12th November)

This homework assignment focuses on Sentiment Analysis in R. You will provide a written analysis based on the following information:

Load the file ML.Tweets.csv (it is online at
'http://nikbearbrown.com/YouTube/MachineLearning/Twitter/'

- Complete following with ML.Tweets.csv:
    - Extract and rank a list of the important hashtags (using td-idf or word entropy).
    - Cluster the tweets using these hashtags.
    - Optional - Give the the clusters names based on their dominant hashtags.
    - Use the qdap polarity function to score the polarity of the tweets in ML.Tweets.csv.
    - Would creating a custom polarity.frame - A dataframe or environment containing a dataframe of positive/negative words and weights - based on the tags and words in these tweets improve the polarity score? Try it.

As a reminder, please provide a written analysis/report as an .Rmd file. Note: This is a graded assignment due by Sunday at 11:59 pm.