PSYC 5710: Introduction to Machine Learning and Data Mining

Assigment 1 - Text Mining

Hudson Golino, Department of Psychology, University of Virginia

Directions:

A. Work with the dataset containing 187 lyrics from the Beatles. This dataset contains four variables (see the table below).
B. Answer the questions that follows (below the table), and save the code you used in a R script.

Variables	Description
songs_title	Title of the Song
songs_Writers	Authors
songs_Song_Lyrics	Lyrics
Year	Year Released

Questions:

1. In our *Text Mining* class we saw a general process for transforming unstructured text data into structured, analyzable datasets. The general process can be summarized as follows:

> Importing text data; text reformating (e.g. transform the words to lowercase); preprocessing: stopword removal, removing punctuation, removing numbers, steeming; creating a document-term matrix.

Using the *Beatles* dataset, show which codes can be used to go from step 1 (importing the text data) to the last step (creating a document-term matrix). Use the *tm* package for processing the data, and explain each step.

2. Observe the document-term matrix below and answer the following questions:

- 2.1 What does *documents: 187* means?
- 2.2 What does *terms: 1719* means?
- 2.3 What does *Non-/sparse entries: 7000/314453* means?
- 2.4 What does *Sparsity: 98%* means?

```
dtm
```

```
## <<DocumentTermMatrix (documents: 187, terms: 1719)>>
## Non-/sparse entries: 7000/314453
## Sparsity           : 98%
## Maximal term length: 17
## Weighting          : term frequency (tf)
```

3. Without changing the maximum level of sparsity presented in the document-term matrix, the resulting dataset will have lots of *0's*. How can you reduce the sparsity of the document-term matrix via the *tm* package? (i.e. which function can you use?)

4. Change the maximum level of sparsity in the document-term matrix setting the *sparse* argument to: 0.99; 0.98; 0.97; and 0.96. What happens with the document term matrices as the level of sparsity decreases?

5. Create a new document term matrix named *dtm.beatles* with the maximum level of sparsity of .90. Convert this document-term matrix into a dataframe named *beatles.lyrics*. How many terms your *dtm.beatles* object contain?

6. Create a plot showing the distribution of the words frequency using the *ggplot2* package.

- 6.1 What are the three most frequent words?
- 6.2 What are the three least frequent words?

7. Create a dynamical heatmap of the words correlation matrix using the *plotly* package. Save the resulting plot into a *HTML* file named *beatles.cor.html*. For computing the correlation, use the `cor_auto` function from the *qgraph* package.

- 7.1 Based on the correlation heatmap, which pair of variables are more strongly positively correlated?

Challenge 1: Create a *github* account (https://github.com). Head over to GitHub and create a new repository named username.github.io, where username is your github username (or organization name). After that, go to your github project, click in *upload files* and drag and drop your *beatles.cor.html* file there.