Team BLodgic
September 21, 2014
Assignment #7.2 Capstone Data Mining Execution

Summary of the output:

Our final output of our capstone data mining execution provides a shiny app word cloud from text mining. The word cloud itself is based off of the frequency count of words from Cyber Security Presentations given at various conferences. The corpus is 4518 line csv file called "Presentations_Mildred_North_America_Eventswhich.csv" which lists presentations and various other metadata from cyber security related conferences.

The R Code:

The code leverages the shiny web app framework for R. The code files are broken out into server.R, a server script, and ui.R, the user-interface script. An additional directory name of "files" (under the parent file directory Test1-app) holds the corpus csv file.

The following libraries are called to run the app.
library('tm') – Text Mining Package <http://cran.r-project.org/web/packages/tm/index.html>

library('plyr') – Tools for splitting, applying and combining data <http://cran.r-project.org/web/packages/plyr/index.html>

library('class') – functions for Classification <http://cran.r-project.org/web/packages/class/index.html>

library('SnowballC')- Snowball stemmers based on the C libstemmer UTF-8 library <http://cran.r-project.org/web/packages/SnowballC/index.html>

library('memoise') – Memoise functions to cache the results <http://cran.r-project.org/web/packages/memoise/index.html>

library('wordcloud) – Word Clouds <http://cran.r-project.org/web/packages/wordcloud/index.html>

Code Breakdown

The code is broken out into 4 main sections, data input, preprocess, transpose, and display.

Data Input

The corpus file directory "files" is called from its directory source.

Preprocess

The corpus is then preprocessed by performing the following:
-strip white space
-lower case the words
-remove punctuation
-remove stopwords
-stem words
-turn the csv file into a plaintext document

Transpose
-Identify the frequency of terms within the document

-translate the document into a term matrix
-build a data frame of the matrix-

Display
-Here the wordcloud library is called.
-The word cloud calls “Hfdf1” or the data frame matrix. The data frame includes the top frequent words (minus the stopwords)
-The plot is sent to shiny and a word cloud is generated

Meaninful insight

Our team learned a great deal on using the text mining packages within R to generate an expressive word cloud summary. This exercise has built a foundation for our future plans with text mining, analytics and web app displays. Our goal is to glean a further understanding of the pulse of the cyber security landscape. And by leveraging the text mining packages within R and the shiny web app framework, this exercise has certainty set the stage for our future development in this area.