**Student Name:** Kirk Kelly

**ID Number:** 2800750

**Course:** MSDS692

**Assignment:** A Positivity Analysis

**Date of Submission:** October 14, 2017

**Professor:** Geena Kim

Data Science Project Development Proposal

**Summary**

The solution analyzes the positivity of tweets, use that positivity to predict the positivity of statements outside of twitter, which use the same words that are in the tweets.

**Project Tasks**

The program uses the doc2vec deep learning algorithm. doc2vec draws context from phrases and is popular for movie reviews and sentiment classification. Text2vec will be used for sentiment analysis. Text2vec was written by Dmitry Selivanov.

709,376 tweets were collected in the file named TrainingData.csv. Those tweets were used to train a model, which was named modelTrainingWithGlmnet.RDS. The sentiment grading is from 0 to 1, therefore, .36 to .65 could be called neutral. The graph will use .5 for neutrality.

The data was split by using createDataPartition. A vocabulary-based vectorization (VBV) was done via preprocessing and tokenization, by using tolower, word\_tokenizer, and itoken. Also, the terms were collected and marked with unique IDs.

Document Term Matrix(DTM) was used to train the model, to find the frequency of terms. This was done by using create\_vocabulary, vocab\_vectorizer, and create\_dtm. The vectorizer function maps terms to indices, removes stop words and unusual terms.

A TF-IDF model was defined using tf-idf. TF-IDF is used to weigh the term frequency (word occurrence frequency) and inverse document frequency (idf decreases the weight for common words and increase the weight for uncommon words) of the data. To fit the model fit\_transform was used. The glmnet package was used for 5-fold cross-validation, using a L1 penalty and a small number of loops to reduce processing time. Prediction was done by the predict method.

The database connection was done by using the odbcDriverConnect method. The data was mapped to a data frame with dmat\_at.

The preprocessing, tokenization, vocabulary creation, DTM creation, data transformation with tf-idf, classification model loading, and prediction was done as before but on the new dataset. The same vectorizer that was used for the tweets, was used for the new data set. The probability that a randomly chosen positive case should outnumber a randomly chosen negative case, is shown by the area under the curve (AUC). The newly calculated weights were added to the new dataset.

A graph was used to show the probability of possibility of individual statements in the data set.

**Challenges**

Collection of the YouTube data then transforming it to the data from the tweets. I had to download a software to scrape data from Youtube. Loading the data into the database was not a problem, however getting it into the same format of the tweet data set took more time.

**Testing**

The IDs of the data sets were used to match the tweet data set in the csv file, to the YouTube data in the data base. The IDs were assigned starting at 1, then incremented by 1, in the order of the days in the data, therefore, earlier days have lower IDs.

A few results were checked mathematically.

**Milestones**

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone | Tasks | Hrs | Date |
| 1 | Data filtering, solution analysis and design stage. Source the data and create system mockup | 23 | Sep 10, 2017 |
| 2 | Design the architecture | 43 | Sep 17, 2017 |
| 3 | Write out the work plan | 11 | Sep 24, 2017 |
| 4 | Develop the database | 43 | Sep 24, 2017 |
| 5 | ETL design | 43 | Oct 8, 2017 |
| 6 | Verify data | 9 | Oct 15, 2017 |
| 7 | Test solution | 11 | Oct 15, 2017 |
| 8 | Peer review of the data and Power BI results, will be done by a peer in my department at work | 7 | Oct 22, 2017 |
| 9 | Complete the documentation | 9 | Oct 29, 2017 |
| 10 | Verify the results of the program | 11 | Nov 5, 2017 |

Total estimation of man hours: 210

References:

Bryl, S. Twitter Sentiment Analysis. Analyzecore.com. Data Scientist at MacPaw Inc. 2017 – 2017.