Data Science - Data Collection & Data Cleaning Report

zpqf41

April 24, 2021

In this report, I discuss the method I have employed to calculate semantic distances between phrases. For the text corpus relating to each keyword, I utilise data downloaded from BBC News [1]. More specifically, I download the webpage content of the top 100 articles relevant to a given keyword.

The Algorithm

The complexity of understanding human language makes the field of natural language processing particularly difficult. Words can have different meanings in different situations, and they can also be arranged in any number of ways. In addition, context is typically also required to interpret a sentence correctly. Semantic analysis aims to understand the meaning of words and how they are interpreted. In my implementation, I assign a value to two given keywords to represent how similar they are, based on a given text corpus.

For my implementation, I use Google's word2vec model [2][3][4] using the Gensim Python module. The word2vec model is trained by moving through the text corpus with a sliding window to provide a prediction of the current word using its neighbouring words [5]. It is not a deep neural network, it works by turning text into a numeric form that a deep neural can process as an input [5].

My implementation takes two phrases as its input. The top 100 articles relevant to these phrases are then used as the text corpus to train the word2vec model. Once the model has been trained, it can produce the similarity of two words. The solution needs to work with a phrase, not only words. My implementation finds the similarity of all words in the first phrase, and all words in the second phase, and takes the average of these. This ensures that the meanings of word2vec word vectors are preserved.

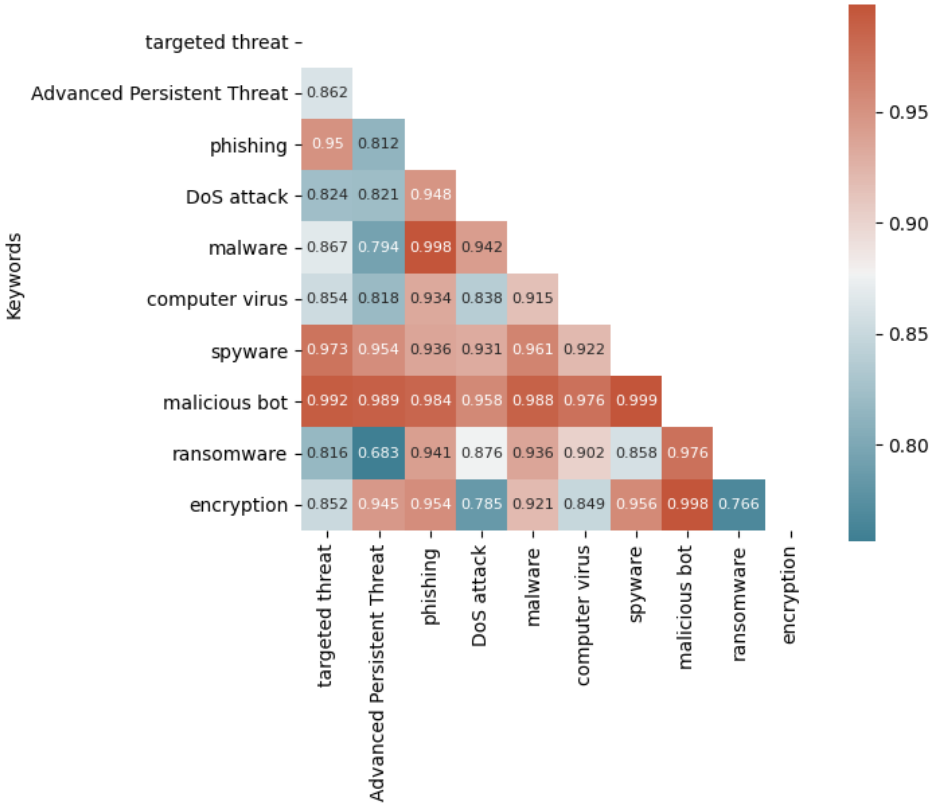

Figure 1: A heatmap for the similarity of keywords.

The Results

Figure 1 is a heatmap describing the similarities of ten given keywords; a greater value and red colour represent a closer similarity. It is expected that these words all have a high similarity. The weakest semantic distance is "ransomware" and "Advanced Persistent Threat" with a score of 0.683. The strongest semantic distance is "malicious bot" and "spyware" with a score of 0.999. These are expected results.

The Receiver Operating Characteristic (ROC) curve in Figure 2 is for a logistic classifier on training data containing tweets with positive and negative meanings [6]. It takes the word2vec model and analyses just how accurate it is. It shows that the accuracy of the model is around 80%.

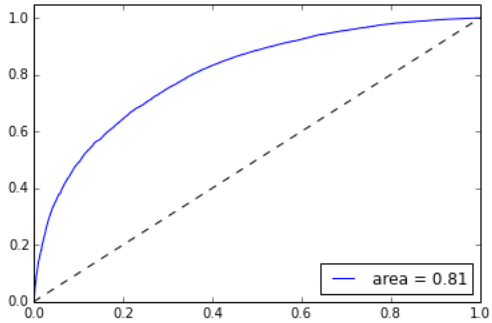

Figure 2: An example ROC curve for a logistic classifier on training data of tweets. [6]

References

[1] BBC News. https://www.bbc.co.uk/news

[2] T. Mikolov, K. Chen, G. S. Corrado, J. A. Dean, "Computing numeric representations of words in a high-dimensional space", U.S. Patent 9037464B1. May. 05, 2015.

[3] T. Mikolov, K. Chen, G. S. Corrado, J. A. Dean, "Efficient Estimation of Word Representations in Vector Space". Sep. 07, 2013.

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality". Oct. 16, 2013.

[5] S. Li, "Understanding Word2vec Embedding in Practice". Dec. 04, 2019. [Online]. Available: https://towardsdatascience.com/understanding-word2vec-embedding-in-practice-3e9b8985953

[6] M. Czerny, "Modern Methods for Sentiment Analysis". Dec. 21, 2017. [Online]. Available: https://medium.com/district-data-labs/modern-methods-for-sentiment-analysis-694eaf725244