Data Collection and Cleaning Report

Word2Vec from Gensim is used to calculate the semantic distances between pairs of keywords. Word2Vec requires training data which we retrieved from BBC news. Each keyword produces ~100 articles which are stored in a text file. These files are joined to produce one master file containing all ~1000 articles. We tokenize the contents of this file such that it is a valid input for the model. Our model is then trained on this tokenised data from which we can use it to calculate the semantic distances.

To ensure the relevance of the articles retrieved, we first work under the assumption that BBC's search feature returns results similar to the keywords searched for. We then implement a 'relevance filter' which checks to see if the contents of the article contains any of the keywords or their synonyms. As such, we accept all articles that successfully pass this filter as being relevant to our task.

We encounter two issues when evaluating the training data for our Word2Vec model. Firstly, keywords comprised of multiple words are not recognised by Word2Vec as one object. To mitigate this, we hyphenate keywords before training. The second issue is the lack of occurrences of specific keywords in our training data as this reduces the accuracy of our model. To mitigate this, we firstly retrieve more data by downloading the Wikipedia articles of each keyword and appending it to the training data. The improvements in keyword occurrences because of these Wikipedia articles is shown in figure 1. We secondly accept synonyms (the same synonyms that are used in the relevance filter) as being equivalent to their respective keywords. The number of additional keywords attained by this is shown in figure 2.

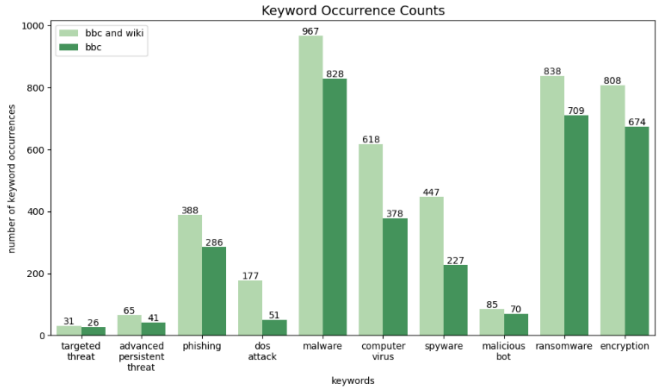

Figure 1 (Keyword Occurrence Counts.PNG)

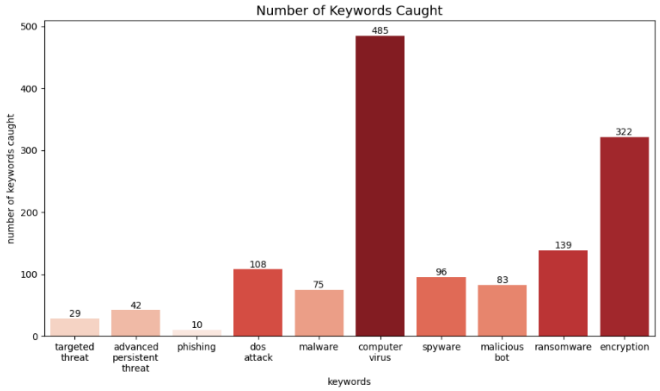

Figure 2 (Number of Keywords Caught.PNG)

As a second metric for semantic distances, we calculate the NGD (normalised google distance) for pairs of keywords. This is calculated by comparing the number of search results produced by google when searching each keyword on its own against searching both keywords together. Any Word2Vvec performance issues due to a lack of keyword occurrences are hopefully mitigated or reduced through the use of a second metric.

To compare our Word2Vec and NGD results, we normalise our two semantic distances dataframes to between 0 and 1. However, a smaller NGD distance means the two keywords are more closely related whereas for Word2Vec this is the opposite. To rectify this, we flip the values in the NGD dataframe. We do not calculate the distances for pairs of identical keywords. This allows our dataframe to show the most related pair of keywords with a value of 1 and least related with a value of 0. Between 0 and 1 the scale is linear. The semantic distances produced by Word2Vec and NGD are shown respectively in figure 3 and figure 4.

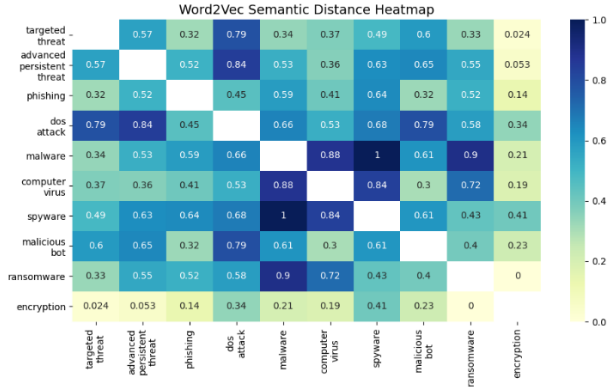

Figure 3 (Word2Vec Semantic Distance Heatmap.PNG)

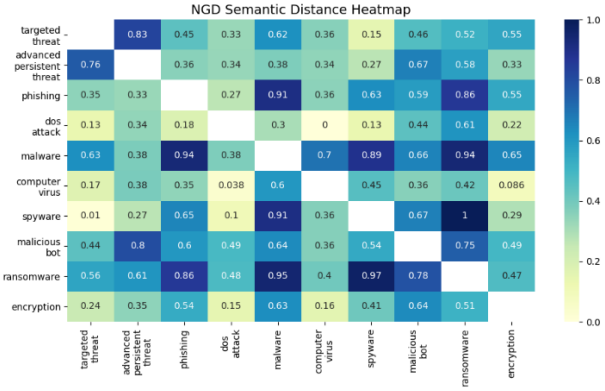

Figure 4 (NGD Semantic Distance Heatmap.PNG)

We can see that both metrics agree for some of the more similar cases such as ‘ransomware’ and ‘malware’ or ‘spyware’ and ‘malware’. This is encouraging as we should expect similar results, however, neither metric is perfect so there are some striking discrepancies such as the similarity between ‘dos attack and ‘advanced persistent threat’. These inconsistencies between metrics due to performance issues of either algorithm will be reduced by combining both outputs.

To amalgamate our results, we take the average of both semantic distance heatmaps. This is valid since both dataframes are normalised so have the same scale. The final output of our algorithm for problem 3 is shown in figure 5. We expect our heatmap to be symmetrical, however, NGD returns slightly different values when searching for pairs of keywords through google in different orders. An intuitive example is ‘heat wave’ vs ‘wave heat’, obviously the former would return more results. From this heatmap, we can see that ‘spyware and ‘malware are the most closely related keywords whilst ‘encryption’ and ‘targeted threat’ are the least related in terms of semantic distance.

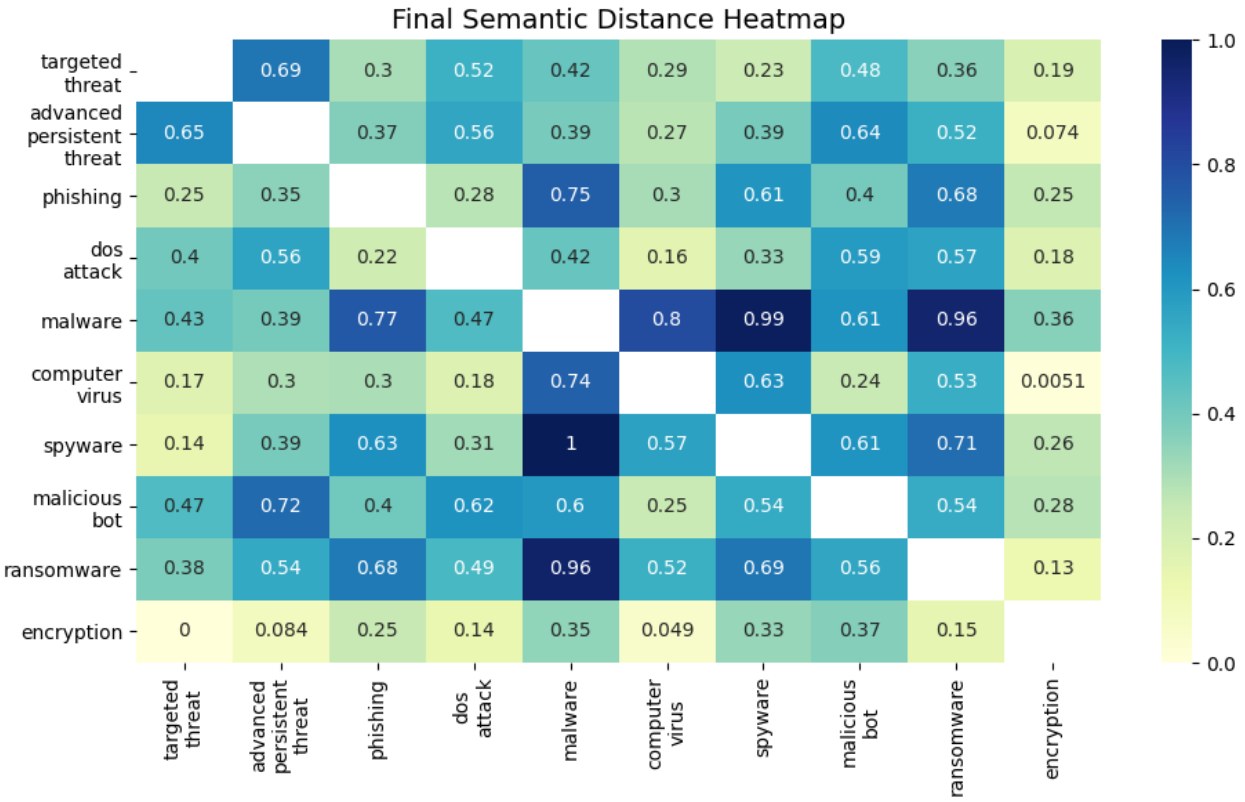

Figure 5 (Final Semantic Distance Heatmap.PNG)