**Assessment 3 Introduction**

Our group is faced with the task of applying topic-modelling to non-trivial data in a cyber security context. Our first challenge was to identify the data set we will perform topic modelling on. This formed most of our research time and involved looking for cybersecurity related data sets, meta data, byte-code and reports that can be modelled using topic modelling. In the documentation and code below, we will outline: the research we undertook and the reason for deciding on using the Enron spam data set, the topic modelling we perform and an analysis of the results obtained. We phrase our question surrounding this data set as: ‘Analyse the Enron spam vs normal data set to create and visualise a topic model and understand the features of spam emails.’

We first consider LDA modelling using the genism package and then move on to a td-idf implementation of LDA on the data set. Our third model uses NMF to create topics with the data and provides a different perspective from LDA. We will then analyse the models using the performance metrics below:

* Classification of the data set into spam and normal emails – the emails are classified into spam and normal emails already and thus we can create predictive models for the documents
* Perplexity and coherence scores – we calculate both in the first two cases using genism packages and in the final case we calculate coherence due to the lack of off-the-shelf implementation for the NMF model

Overall, the project aims to apply topic modelling to the Enron data set to create meaningful visualisations, accurate classification results and to develop our own skills in topic modelling.