**Research Summary**

The first step of this project was to conduct research into some of the data sources we could consider for the NLP task. From the assessment 3 brief, ‘It is understood that public domain cyber-security data in this space is sparse and that you may have to work with other sources.’ Some of the sources of data we considered for review were:

1. Enron Spam Emails - <http://www2.aueb.gr/users/ion/data/enron-spam/>
   1. The emails are split into spam and ‘ham’ (normal) emails and constitute a large amount of language data from a cybersecurity point of view, considering the impact of spam and malware through email.
2. 2016 News Articles - <https://github.com/derekgreene/topic-model-tutorial/tree/master/data>
   1. This is a data set of 4551 news articles from 2016 covering topics ranging from Twitter banning white nationalists to Innovation funds in Africa. The data set is not focused on cybersecurity (although there are cybersecurity reports) but provides a large amount of data to be processed.
3. NCSC Reports - <https://www.ncsc.gov.uk/section/keep-up-to-date/threat-reports?q=&defaultTypes=report&sort=date%2Bdesc&writtenFor=Large+organisations>
   1. These are the weekly threat reports from the NCSC. This would allow us to cover a broad range of topics within cybersecurity and possibly look at the changing landscape of cybersecurity over the last 4 years but poses difficulty with collating the information initially.
4. ABC Million News Headlines Dataset - <https://raw.githubusercontent.com/ravishchawla/topic_modeling/master/data/abcnews-date-text.csv>
   1. Like point 2, this covers a broad range of news headlines (although excludes any of the actual reporting compared to point 2) and thus provides a vast amount of data to analyse but is not focused on cybersecurity.
5. Krebs on Security Articles - <https://github.com/87bdharr/Cyber-Topic-Modelling>
   1. This repo uses data pulled from Krebs on Security articles. This has similar issues to option 3 as it would involve scraping data from articles on the web instead of using a predetermined data set but has the advantage that it is very cyber focused.

Our team decided to go with option 1, the ‘Enron spam e-mails’ given a couple of considerations, such as:

- *The integrity, replicability and continuity of the data*: The e-mails were formatted in an initial gzip (zipped) state, which was unzipped and resulted in 6 separate folders – determined by the author of the e-mails in each respective folder. These were then placed in a single folder. All e-mails were represented in an individual and uniquely numbered plain text file. A ‘summary’ text file also featured in every of the 6 folders, explaining the correspondence and the separation of spam from ham.

- *Ease of ulterior performance analysis:* Given the summary files, we possess prior knowledge of what to expect the ratio spam:normal would look like. Moreover, each of the 6 folders had its e-mails separated in 2 subfolders: ‘Spam’ for all spam e-mails and ‘Ham’ for all normal correspondences.

- *Source validity:* The Enron case is well known, and all e-mails were sure to be good sources of insight. Moreover, those already categorized as spam were known to have been deemed so on the basis of something better than an industrial spam filter, e.g: yahoo, gmail etc.

The only hardship encountered was uploading the files to GitHub – as even in a compressed state they were slightly (28mb) over the upload limit supported by the browser version (25mb). Our team decided to eliminate one of the 6 folders at random (enron 4, specifically, was arbitrarily chosen) as enough data was available without it as well.