

The creation of the JSON file essential for the D3 Visualisations is created by the functions in the ‘parsingEnron’ python executable. The class starts off by creating an object of type ‘Email’ in order to store the parsed data from the given dataset. The object accepts the email sender, recipients and body as parameters.

Parsing

The root directory “maildirtest” which holds the dataset provided is iterated over with three iterators ‘directory’, ‘subdirectory’ and ‘filenames’, where ‘filenames’ is iterated over with iterator ‘filename’ in a nested for loop. The file is opened using prebuilt os functions, where the parsed data is stored in a variable ‘data’.

The sender of the emails is found with the use of the *email.parser* library. An instance of the library was created where, in this case, ‘from’ was passed as an argument to return the ‘from’ section of the email. A similar operation is carried out when reading the ‘to’ section of the email where the result of the *emailParser* instance is saved into the recipients variable. The same operation is carried out for the ‘bcc’ and the ‘cc’ fields. Since all of the ‘to’, ‘bcc’ and ‘cc’ fields indicate a recipient, the recipient variable is set to be a list for a set of recipients. Similarly, the body of the email is parsed.

An object of type Email is instantiated and appended to the list of all emails in the dataset. Similarly, a list of all the email addresses is created. Afterwards, a dictionary is created of all the combined documents passed between two distinct people by iterating through the emails and recipients lists.

Preprocessing

Preprocessing entails the carrying out of arranging the data to be valid for TF-IDF weighting and Cosine Similarity. Preprocessing consists of tokenisation, casefolding, stopword removal, stemming and symbol removal. Tokenisation is the process of turning sensitive data into nonsensitive data or tokens, this happens with the information being swapped into an algorithmically generated number with the same length and format of the initial data. Casefolding converts all the data passed into lowercased symbols. A pre set value of stop words, such as “and”, “then” and “is” are removed from the dataset through the execution of stop word removal, this selection selects data which is more important or relevant to the document. Stemming is the process of setting a particular word to its most basic form, for example: “Stemming” or “Stemmed” are outputted to be “Stem”. Symbols usually carry negligible importance in a document, therefore a variable symbol is iterated over and checks whether a variable in data contains a symbol, if a symbol is found, the data is appended and placed in a temporary array.

Calculating TF-IDF