Part one Preprocessing

As all the data provided are raw text files, it is necessary for us to conduct text preprocessing before calculating unsmoothed unigram and bigram for the text files. After going through most of the text files, we found that these emails begin with the words ‘Subject : ’,’ writes : ‘, ‘wrote :’, ‘said :’ and ‘Subject : Re’. In addition, we found that the signatures at the end of email are also not relevant to the email text. Hence, we used Python program to clean the heading and signatures to get rid of irrelevant text in the raw text files. In the preprocessed emails, we can still find many irrelevant symbols such as ‘> " | # : - ) ( \* [ ] ) { + = ^ \_\_ ~ / \\’. It is not sensible to include all these symbols in the email, as they do not have any meaning and may just indicate the format of the text files. Apart from that, the full email addresses in each email are also irrelevant to the content of the email. Including these irrelevant noise into our analysis can harm our bigram probability tables. As these symbols do not appear frequently, involving these rare case into language models can reduce the probability of generating clean text and this will also increase the probability of <UNK> which will be demonstrated further in the later session. After removing all the irrelevant noise in the text file, we also clean the format of raw text. We ensured that there is only one white space between words and beginning of each email is text instead symbols.

After obtaining cleaned text following previous procedures, we used NLTK to help us set up boundary ‘<s>’ between each sentences. In addition, we input ‘<s>’ at the beginning and the end of the email to help separate different emails.

Random Sentences Generation

In this part, we used all the files in each topic folders to generate sentences. We concatenate all files in one folders into a string as training dataset. Then we used both unigram and bigram to generate the random sentences. The first step for generating unigram and bigram language model is to use training data set and get the probability and conditional probability for both unigram and bigram.