**Goal**

To create a simple program, in python, using Natural Language Tool Kit (NLTK), that will read 20 articles and return words that represent the article.

**Method**

Using the NLTK the program can access articles from the NLTK corpus Reuter. From these articles the program will access the raw data from the articles using the raw method from the built-in corpus reader. With this raw data the program can remove all punctuations and all stop words. From there the program will find the top used words, all the hypernyms of all words in article to see if there are any words that have the same hypernyms and synonyms of the top used words. The program will then loop that process again by grabbing the hypernyms of the previous hypernyms found, the synonyms of those hypernyms, and vice versa. The results will then be exported to a csv file which then can be read opened as an excel file and observed.

**Process**

Originally, to test simple results of the program, an early version of the program was created that only read one hardcoded article. In this case, the data used was Herman Melville’s *Moby Dick* from NLTK’s Brown Corpus. This program only found the top 15 used words, the synonyms of those words, and the synonyms of the synonyms, and so on. From this it was found that the results resembled many words that described the given article. Example being whale being one of the top synonyms. It is also noted that if this process is repeated the results will eventually only display one word.

From this I then created a program that will take the first 15 articles in the corpus Reuter using their file ids. This did not give the same results as the earlier program. From this it was found that the articles would have no repeated words or synonyms. To combat this we updated the program to find hypernyms which we were hoping will find more accurate results.

Attached are the results of the most updated version of the program.

**Results**