**PREPROCESSING OVERVIEW:**

The goal of the preprocessing stage was meant to remove noise in the form of stop words and to convert the various news articles from the reuters dataset into a form that is easier to process. The stages included:

* Tokenization:
  + This involved breaking down the text in the various news articles into tokens (individual words)
* Stop Word Removal:
  + Stop words are then removed from among the various tokens.
* Lemmatization:
  + This is done to remove the effects of grammatical rules such as tenses so that similar words are treated the same.

**RESULTS AND CHALLENGES:**

Results included:

* A graph of a distribution of document length

  AI-generated content may be incorrect.Distribution of the word count among the various documents with the most frequent occurring between 0 to 100 words.

A graph with blue bars

AI-generated content may be incorrect.Counts of the top 10 words. Most of the words that were found within the top 10 had no grammatical meaning. Which implied that the preprocessing has some room for improvement.

* The distributions of the various parts of speech was also obtained which showed that nouns where the most common followed by adjectives, verbs and so on.