**Solution**

**Part 1: Corpus Analysis**

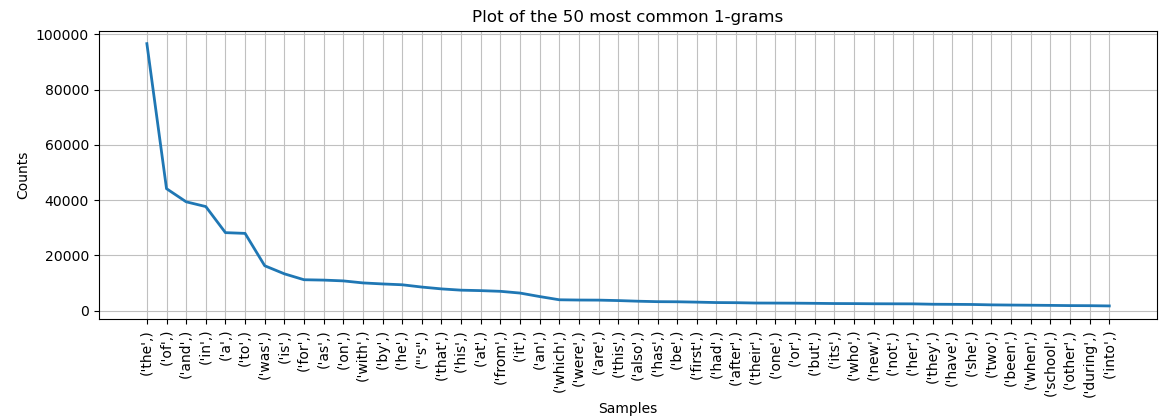
**Q1. Unigram analysis:**

**(a) Mention the total unique unigrams present in the corpus**

Total unique 1-grams : 75645

**(b) Plot the distribution of the unigram frequencies**

Top 50 most common words



**(c) How many (most frequent) uni-grams are required to cover the 90% of the complete corpus.**

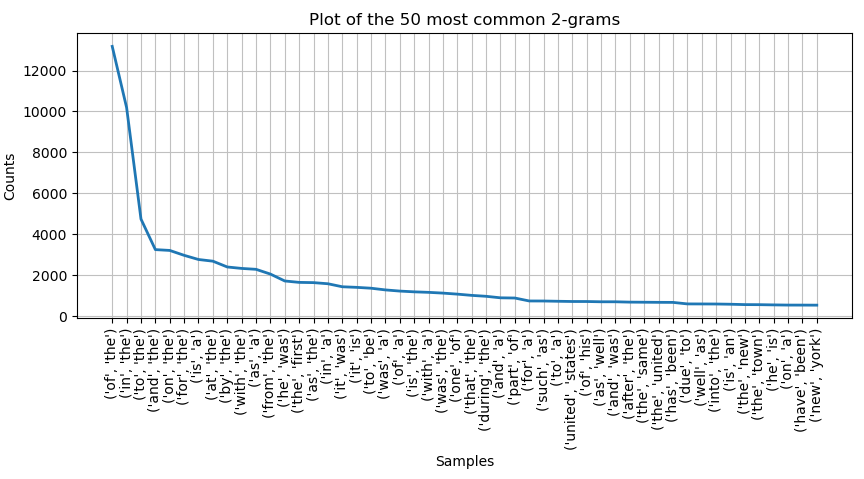
1-grams required to cover 90% of the complete corpus : 11714

**Q2. Bigram analysis:**

1. **Mention the total unique bigrams present in the corpus.**

Total unique 2-grams : 581899

1. **Plot the distribution of the bigram frequencies.**



**(c) How many (most frequent) bi-grams are required to cover the 90% of the complete corpus.**

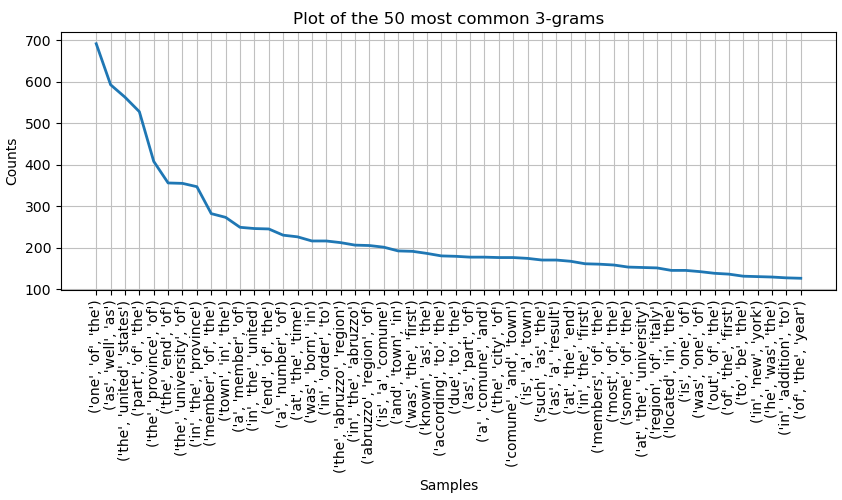
2-grams required to cover 90% of the complete corpus : 454207

**Q3. Trigram analysis:**

1. **Mention the total unique trigrams present in the corpus.**

Total unique 3-grams : 1044041

1. **Plot the distribution of the trigram frequencies.**



**(c) How many (most frequent) tri-grams are required to cover the 90% of the complete corpus.**

3-grams required to cover 90% of the complete corpus : 919608

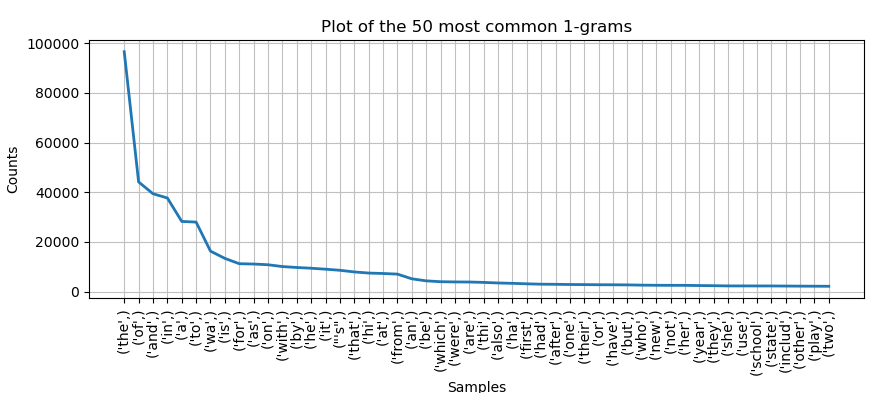
**Q4. Repeat Q1, Q2, and Q3 after performing the stemming process on the tokens.**

**Unigram analysis:**

1. **Mention the total unique unigrams present in the corpus.**

Total unique 1-grams : 60644

1. **Plot the distribution of the unigram frequencies.**



1. **How many (most frequent) uni-grams are required to cover the 90% of the complete corpus.**

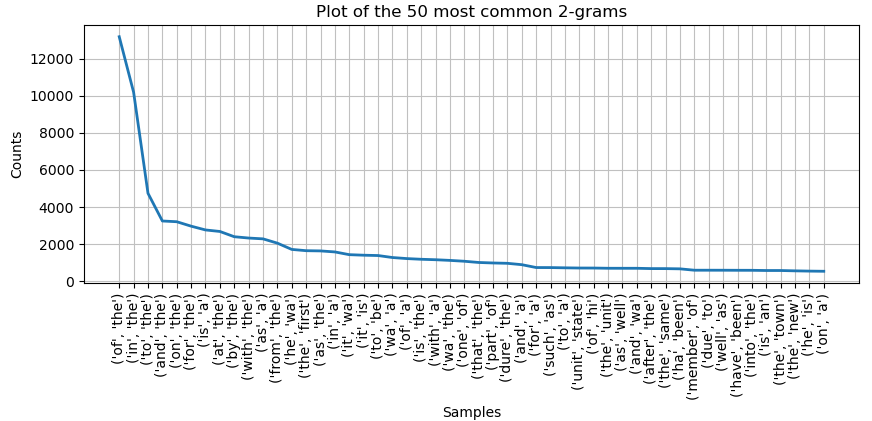
1-grams required to cover 90% of the complete corpus : 6605

**Bigram analysis**

1. **Mention the total unique bigrams present in the corpus.**

Total unique 2-grams : 534154

1. **Plot the distribution of the bigram frequencies.**



1. **How many (most frequent) bi-grams are required to cover the 90% of the complete corpus.**

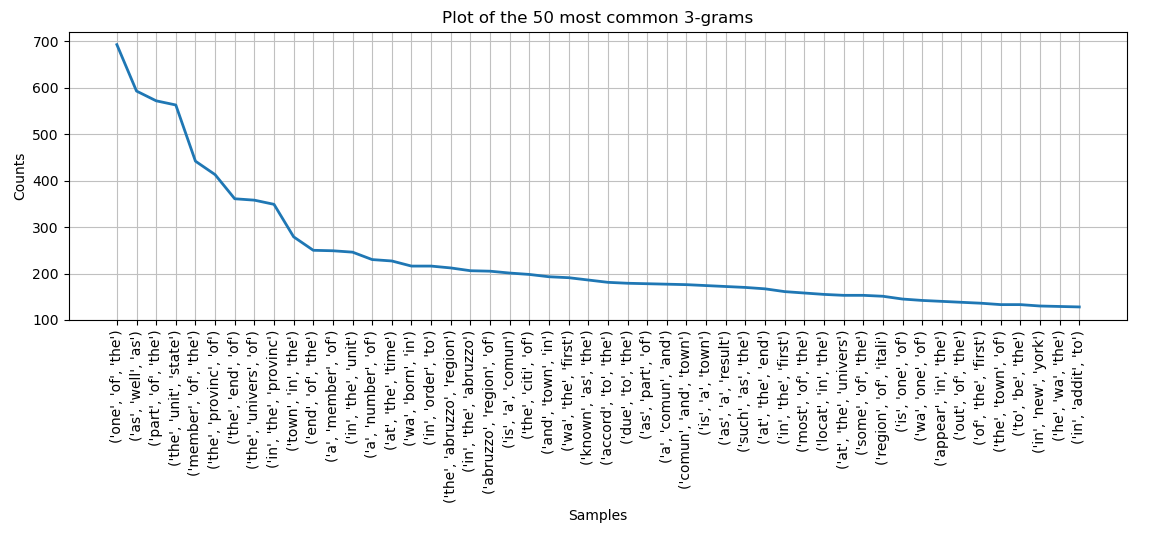
2-grams required to cover 90% of the complete corpus : 406462

**Trigram analysis:**

1. **Mention the total unique trigrams present in the corpus.**

Total unique 3-grams : 1025690

1. **Plot the distribution of the trigram frequencies.**



**(c) How many (most frequent) tri-grams are required to cover the 90% of the complete corpus.**

3-grams required to cover 90% of the complete corpus : 901257

**Q5 Briefly summarize and discuss the frequency distributions obtained in Q1 to Q4. Do these distributions approximately follow Zipf's law?**

Yes. It does follow zipfs law approximately. We can see that in charts plotted for every n-gram case atleast for top 5 ngrams before it becomes constant

**Q6. What library you used for tokenization and stemming? What were the underlying algorithms used by the library for these tasks?**

I used nltk tokenization and nltk stemming libraries. Nltk stemmer uses Porter Stemmer Algorithm. NLTK Tokenizer uses TreebankWordTokenizer along with PunktSentenceTokenizer for the specified language

**Q7. Report three examples based on your observation, where the tool used for tokenization did not tokenize the character sequence properly.**

**Part2: Vector-space based IR System**

1. Query = “”
2. Query =
3. Query =
4. Query =
5. Query =
6. Query =
7. Query =
8. Query =