Word Co-occurrence - Fluke or Real?

Medha Prodduturi - SE20UCSE096

Word Co-occurrence analysis is widely used in various forms of research and aims to find similarities between word pairs and patterns. In this lab, we use nltk - natural language tool kit, pandas, and NumPy libraries to analyze two .txt files and understand the co-occurrence of words.

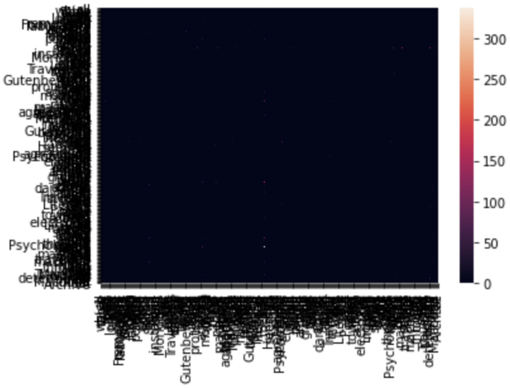
By using natural language processing to analyze these texts, we dwell into bigrams. Biagrams are two words that come together in a corpus. We first split the .txt file into tokens. Then, we split each token into 2, creating bigrams.

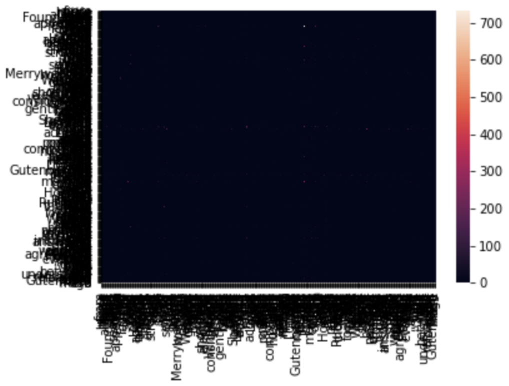
```
tokens = text.split()
bigrams = ngrams(tokens, 2)
```

Once we split the words into bigrams, we analyze the frequency of each word pair. This way, we understand the number of times a pair appears, ultimately proving word co-occurrence is real.

```
frequence = nltk.FreqDist(bigrams)
```

This gives us a result of a key, value pair: [(('of', 'the'), 337), (('in', 'the'), 176), (('I', 'had'), 127), (('I', 'was'), 108), (('and', 'the'), 102), (('to', 'the'), 99), (('of', 'a'), 76)...]





Looking at the heatmaps we have created through the data we have analyzed, we can see the entire graph is dark. This shows the abundance of overlapping word pairs or their frequency. This determines the existence and intention of word co-occurrence in any text file.