**PART I:**

1. **How many unique bigrams are there?**





1. **List the top ten most frequent bigrams and their counts.**



1. **What fraction of all bigrams occurrences does the top ten bigrams account for? That is, what is the cumulative frequency of the top ten bigrams?**

Adding all the topmost frequent bigrams: 52,508

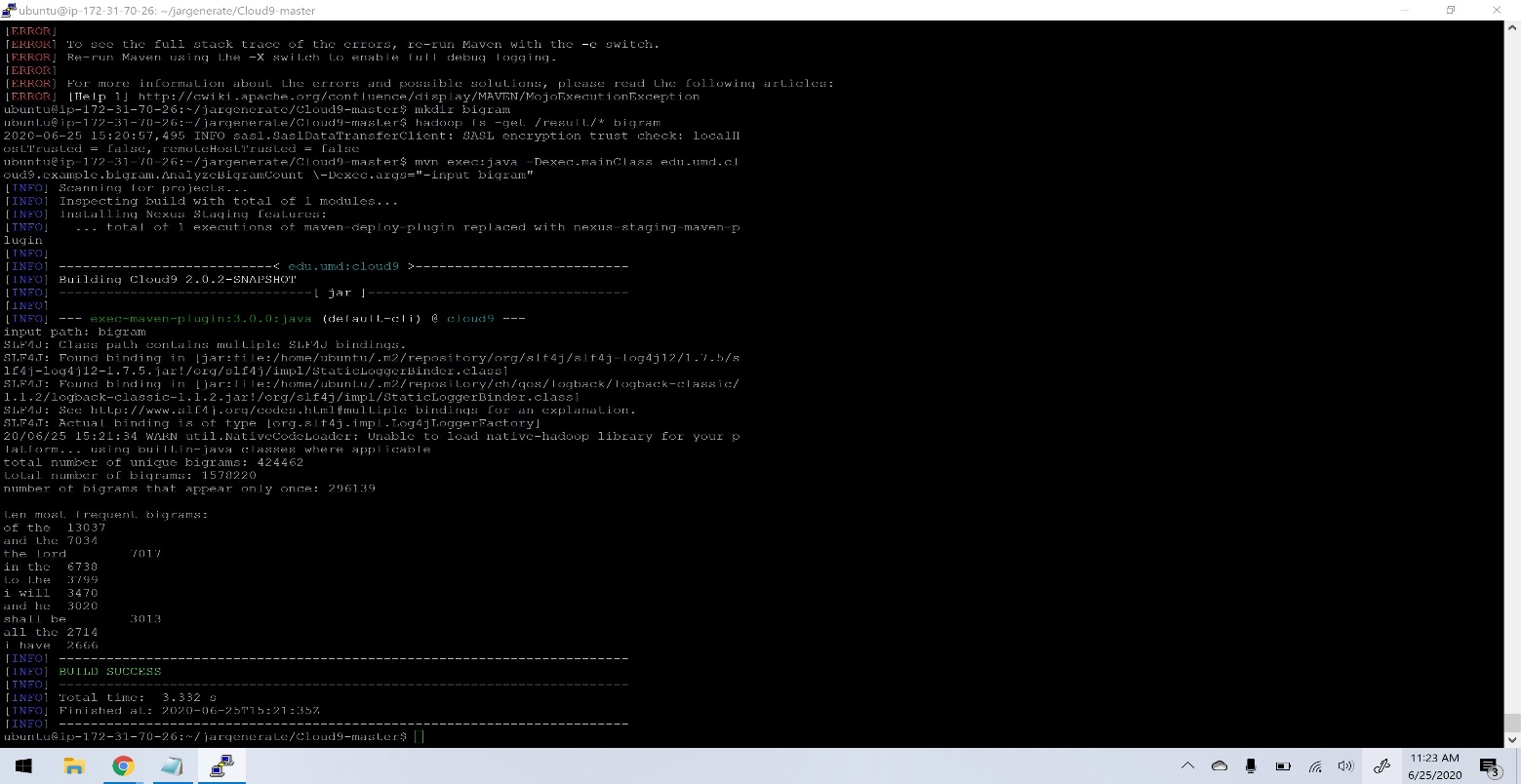
Total number of bigrams = 1578220

Cumulative frequency of the top ten bigrams = 52508/1578220 = 3.33%

1. **How many bigrams appear only once?**

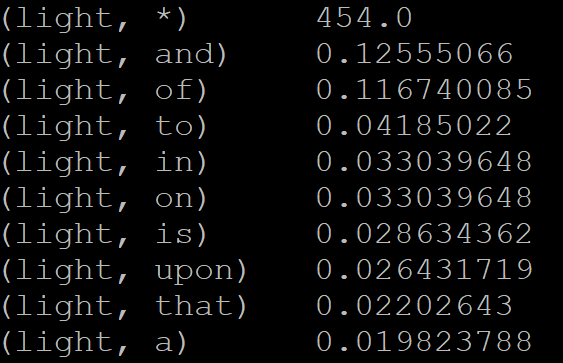


**Full Screen Short of the execution of Bigram count:**

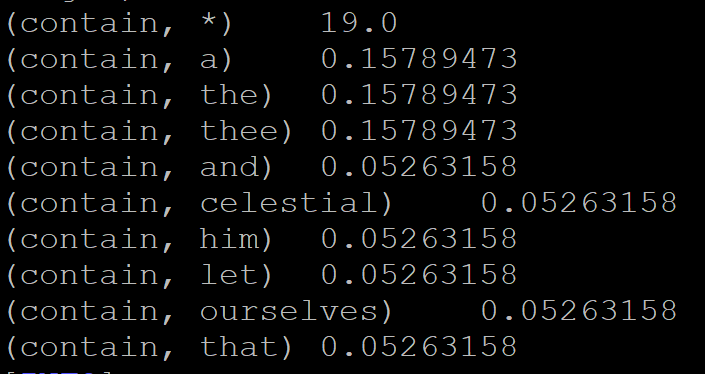


**Part II:**

1. **What are the five most frequent words following the word "light"? What is the frequency of observing each word?**



1. **Same question, except for the word "contain".**



1. **If there are a total of N words in your vocabulary, then there are a total of N2 possible values for F(Wn|Wn-1)—in theory, every word can follow every other word (including itself). What fraction of these values are non-zero? In other words, what proportion of all possible events are actually observed? To give a concrete example, let's say that following the word "happy", you only observe 100 different words in the text collection. This means that N-100 words are never seen after "happy" (perhaps the distribution of happiness is quite limited?).**

424462 / 417882 = 0.0002430724

**Full Screen Short of the execution of Bigram counts to Relative Frequencies:**

