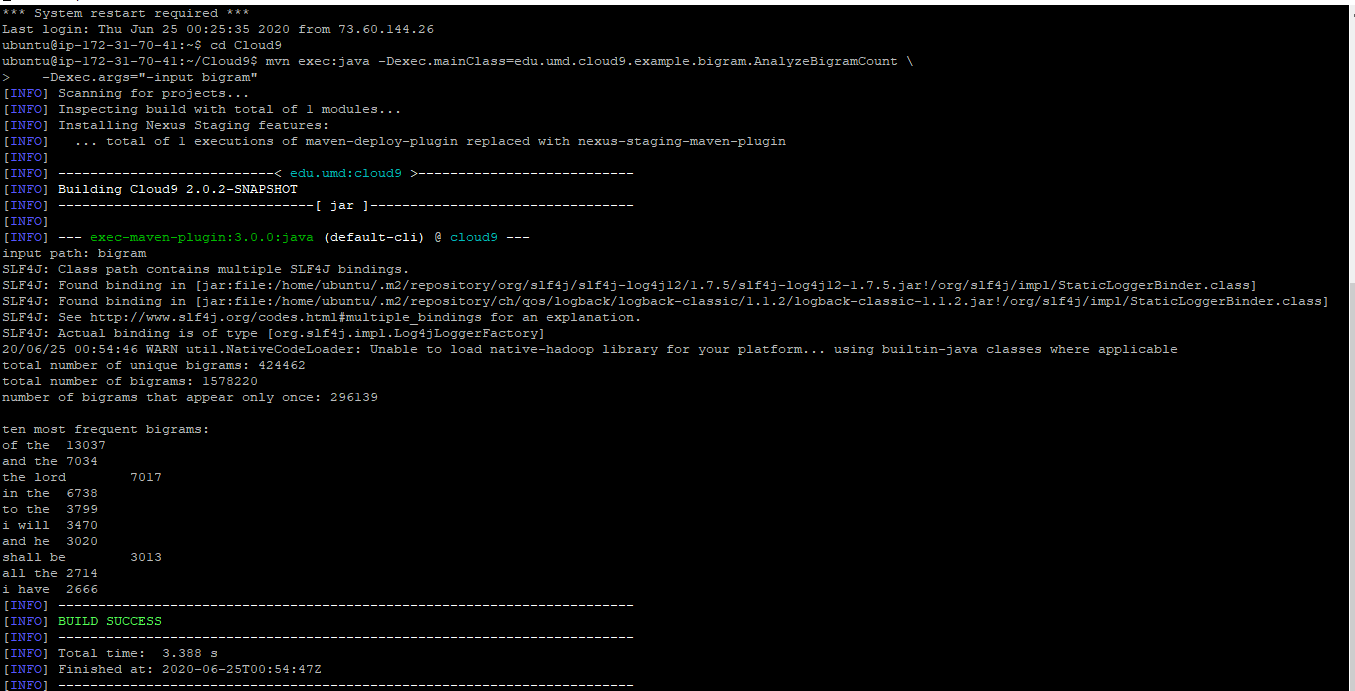
**Project 1**

I did Hadoop setup as mentioned in week1 documentation and I did cluster setup as well.

**Part-1**

I ran BigramCount jar as per the example given in Cloud9 to know the bigram count, This is the result from instance.



1. How many unique bigrams are there?

Ans : 424462

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

Ans : Ten most frequent bigrams:

of the 13037

and the 7034

the lord 7017

in the 6738

to the 3799

i will 3470

and he 3020

shall be 3013

all the 2714

i have 2666

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?

Ans: 52508/1578220 = 3.33%

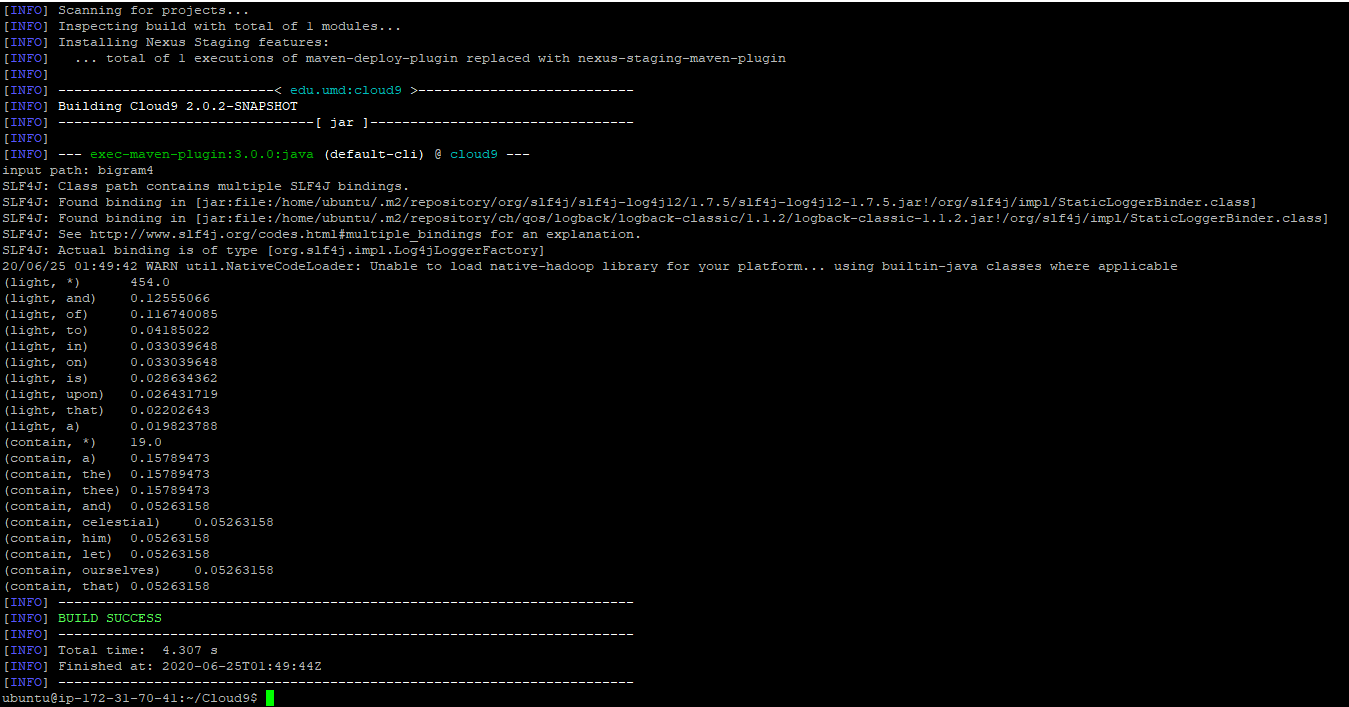
1. How many bigrams appear only once?

Ans: 296139

**Part -2**

Similar like part1, I ran Bigramfrequency jar file to know the frequency of specific words in bigram.

This is the result from my instance.



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

Ans:

(light, \*) 454.0

(light, and) 0.12555066

(light, of) 0.116740085

(light, to) 0.04185022

(light, in) 0.033039648

(light, on) 0.033039648

(light, is) 0.028634362

(light, upon) 0.026431719

(light, that) 0.02202643

(light, a) 0.019823788

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

Ans:

(contain, \*) 19.0

(contain, a) 0.15789473

(contain, the) 0.15789473

(contain, thee) 0.15789473

(contain, and) 0.05263158

(contain, celestial) 0.05263158

(contain, him) 0.05263158

(contain, let) 0.05263158

(contain, ourselves) 0.05263158

(contain, that) 0.05263158

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?).

Ans: 424462 / 417882 = 0.000243