

### Answer3 (b)

Now according to the Heaps' Law:

$$v = k \cdot n^{\beta} \text{-----}(1)$$

Where  $\beta = 0.5$ ,  $v$  is vocabulary size (number of unique words),  $n$  is the number of words in corpus,

To Verify Heap's law on the Alice in Wonderland text. We should know the Number of Words Processes and Number of Unique Word Seen.

After executing the file "**Problem3b-code.py**" given in the **Python** folder using **run.sh**, we get the following values as given below

Model parameters  $k$  and  $\beta$  are calculated using the least square method.

Vocabulary words Count ( $v$ ): 2632

Total Number of words present in the Text ( $n$ ): 26693

The Value of  $k = 5.28$

The Value of  $\beta = 0.614$

After putting the values of  $n$ ,  $k$  and  $\beta$  into the equation (1)

$$v = 5.28 * (26693)^{0.614}$$

After calculating the above equation we get

$$v \approx 2632 \text{-----}(2)$$

we can analyze that Calculated Vocabulary word count value is same as we got after running the code. Hence Heap's law verifies on Alice in Wonderland Text.