Problem 3

a.

Based on Heaps' Law: $v = k.n^{\beta}$, where v is vocabulary size(number of unique words), n is the number of words in corpus.

Here, the β=0.5, and we don't need to know the value k.

Thus, $v=k.n^{0.5}$ (equation 1).

Assuming p be the proportion of a collection of a text before 90% of the vocabulary has been encountered, then according to the Heaps' Law:

$90\%*v=k.(p*n^{0.5})$ (equation 2)

Based on equation1 and equation2, we can get: p=0.81

Thus, the proportion of a collection of text must be 0.81 before 90% of its vocabulary has been encountered.

b.

I have written the code to verify Heap's Law on the Alice in Wonderland text.

Based on my programing, I got the following information:

β=0.094 k=427.582 n=24931031 v=2632

According to the results,

$V=k.n^{\beta}$ = 2120 and the actual unique number is 2632

Because 2120 ≈ 2632, the Heap's Law has been verified.