**Answer2. (b)**

Now while building retrieval index, it's decided to omit all words that occur fewer than five times (i.e., one to four times). According to Zipf's law

Proportion of the Vocabulary words from Alice in Wonderland after omit =

Highest Rank of words (frequency 1) – Highest Rank of words (frequency 5) / Total number of words

Basically after using the above formulae we are getting the number of words whose frequencies are 1, 2, 3 and 4. After getting the words with frequency fewer than five times, we have to divide the quantity by Total number of vocabulary words to get the proportion need to be omitted.

The proportion of total words omitted from the collection
Highest Rank of Word having Frequency 1 = 2632
Highest Rank of Word having Frequency 5 = 697

Total Number of Vocabulary words = Highest Rank = 2632

(2632 – 697) / 2632 = 0.735

**Actual Proportion of vocabulary words omitted = 0.73 or 73%**

**Predicted Proportion can be calculated through**

Predicted Proportion for Number of occurrence (n=1) = (1 /n *(n + 1))
= 0.500 --------(1)
Predicted Proportion for Number of occurrence (n=2) = (1 /n* (n + 1))
= 0.167 ---------(2)

Predicted Proportion for Number of occurrence (n=3) = (1 /n *(n + 1))
= 0.083 ---------(3)

Predicted Proportion for Number of occurrence (n=4) = (1 /n *(n + 1))
= 0.050 ---------(4)

**By adding (1), (2), (3) and (4)**

**Predicted Proportion of vocabulary words omitted = 0.80 or 80%**

The proportion of words that actually be omitted from Alice in wonderland

Number of Words with the frequency 1 = (Highest Rank of word having frequency1) – (Highest Rank of word having frequency2) = 2632 – 1488 = 1144

Number of Words with the frequency 2 = ((Highest Rank of word having frequency2) – (Highest Rank of word having frequency3)) * 2 = (1488 – 1079) * 2 = 818

Number of Words with the frequency 3 = ((Highest Rank of word having frequency3) – (Highest Rank of word having frequency4)) * 3= (1079 – 846) * 3 = 699

Number of Words with the frequency 4 = ((Highest Rank of word having frequency4) – (Highest Rank of word having frequency5)) * 4 = (846 - 697) * 4 = 596

Total Number of words in the text = 26693

**Proportion of words that is omitted from total words=**
**(1144 + 818 + 699 + 596) / 26693 = 0.12 or 12%**