Report for Assignment-02

1. to find a given word/token is a sentence ending word or not we'll use the bi-gram plotting in the 'bigram_result.txt' to find P(**X**|</s>). which can be easily found using the script as those lists in which list[1] is '</s>' and list[0] is our word/token '**X**'.

1a. In my perspective the one with Bigram(n=2) seems more reasonable as we are getting lower bridging conditions in case of finding probability of a n-gram as an ending pair.
So, It doesn't always matter the lower or higher number of 'n' it all depends upon the corpus we are using.
So no n as a general can be said to work always for all datasets.

1b. Figure '1b.png' will be the zipf's law for the unigram token/words.

1c. Figure '1c.png' will be showing the plot of P(**X**|</s>), mean plot of Rank Vs frequency of all the bigrams which are ending the sentences, I.e containing '</s>' as the second word.

1d. Using Bi gram only I tried to generate certain no. of sentences like

* I take “The” as starting word:
“ The Russian Version of Kim Kardashian https://t.co/zTG5sHp9vm https://t.co/riMihTyoGR”

* I take “best” as starting word:
“best D.”

* I take “RT” as starting word:
“RT @VP: Happy 55th, Barack! A brother to me, a best friend forever. https://t.co/uNsxouTKOO ”

and many more can be generated.