**Dataset Chosen:** Shakespeare Dataset

**Data Cleaning Steps:**

* In this Stage, I Removed some of the unnecessary special characters that are found in the text corpus , so that I can have better understanding of the data.

**Word Distribution:**

I Have taken the corpus in to a single string and started to take the frequencies of the each word.  
Thus helping me in calculating the probability of each word against the whole corpus words.

**Word Correlation:**

I approached this problem by taking the bi-grams-sequence of two adjacent elements. For each bigram combination i took their combined frequency

As an Example :-

(a,b):221

(a,c):321

(c,a):10

**Text Generation:**

I took the bigrams from the corpus.

W1-> [x,y,w2,.....]

W2-> [x,y,w3,.....]

W3->[a,b,c...]

I have chosen the start token randomly ,As I felt that this First word should be started in a randomized fashion. And now for the selected token ,I looked at bigrams history list and chosen the random word among the words that I have for my selected token in the biagram. I repeated this approach for all the 30 words. I have taken the 30 words as my sentence generation length.

**Sequence Prediction:**

I have taken the sequence of words as input.

Consider an example like below

w1 w2 w3 w4

For the last word i.e. w4 ,i took all the bigrams combinations(state transitions).

W4 -> bigram combination -> {a,b,c,d,..}

For the entire bigram list I calculated the probability from all the input word sequence.

P(a/w1), p(a/w2), p(a,w3), p(a,w4).

The process is repeated for all the input words and bigram combination of the last word. Finally, I have chosen the word which has more probability.