NLP

Assignment 3

Submitted by : Ganesh Borle (201505587)

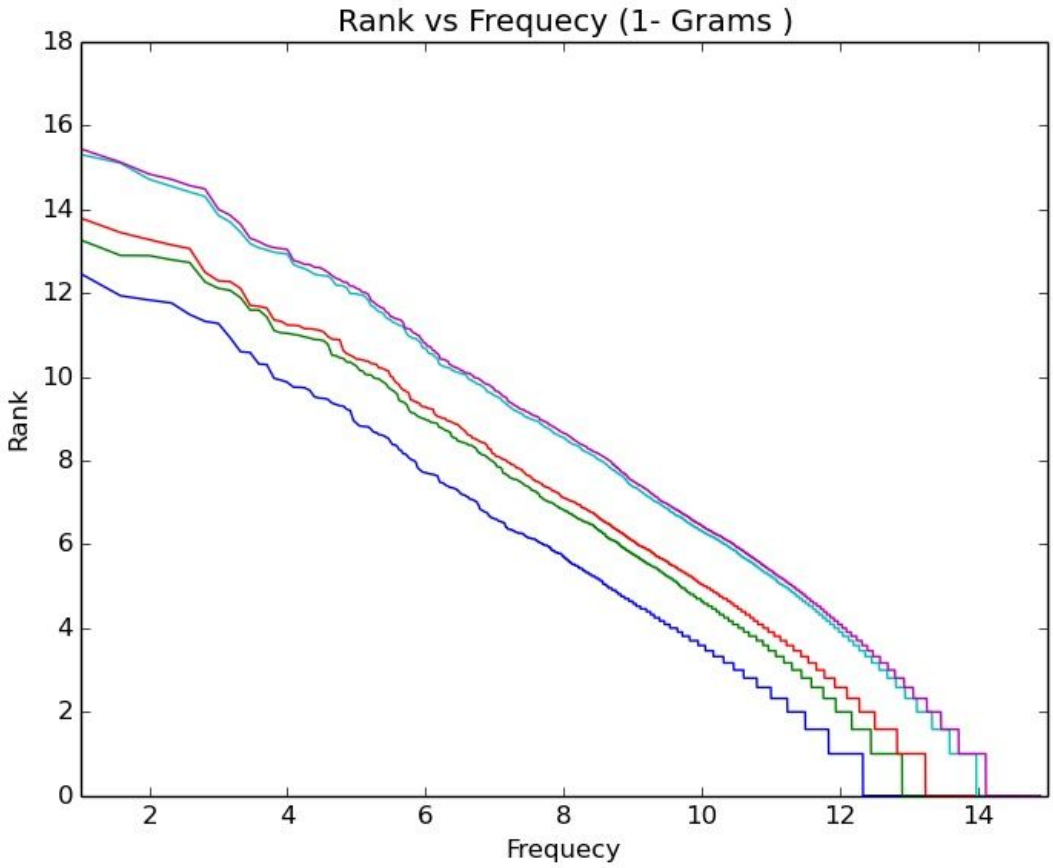
Corpus :

The corpus text is taken from the below authors:

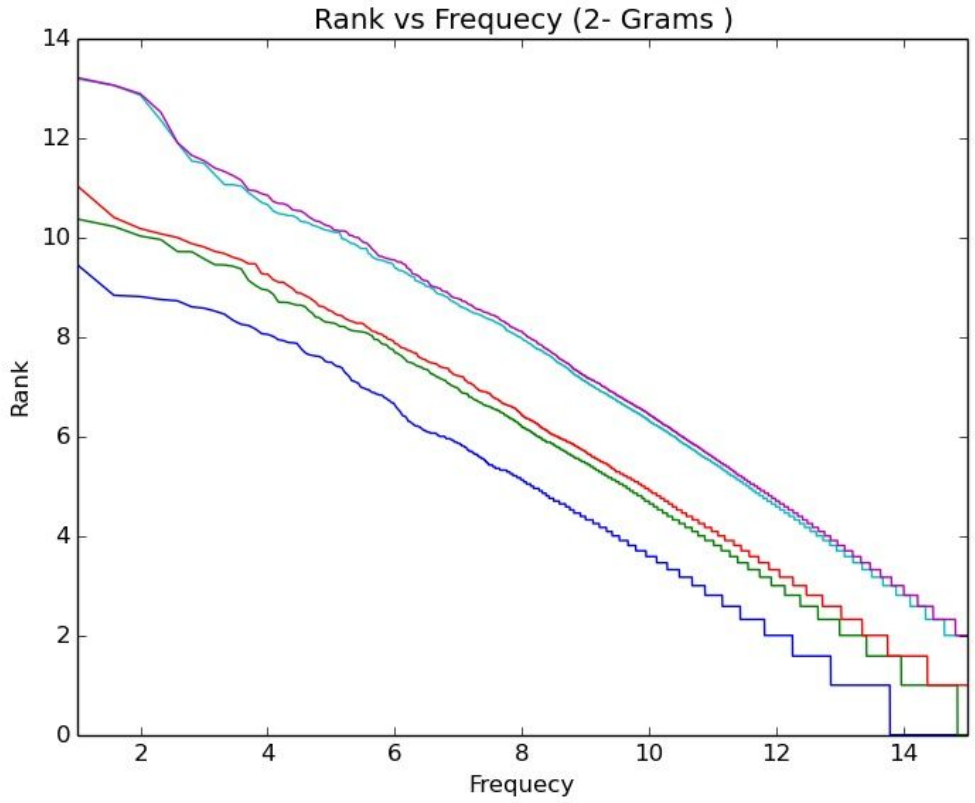
1. Alexandre Dumas
2. Benjamin Franklin
3. Hermann Ebbinghaus
4. Jane Austen
5. Jonathan Swift

Graphs:

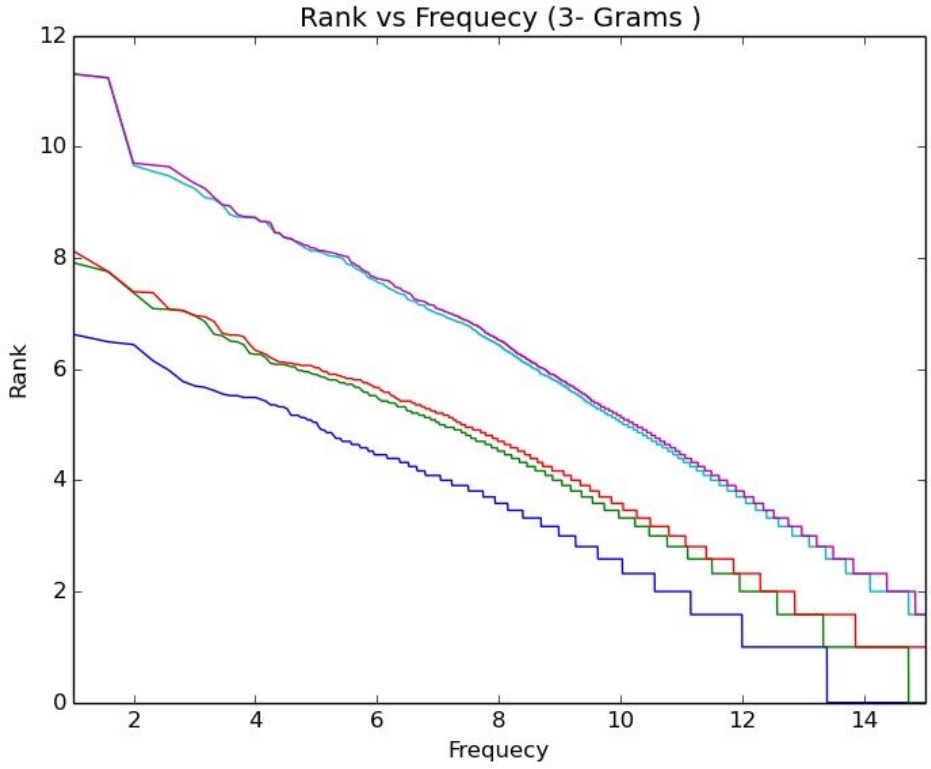
- **Unigram**



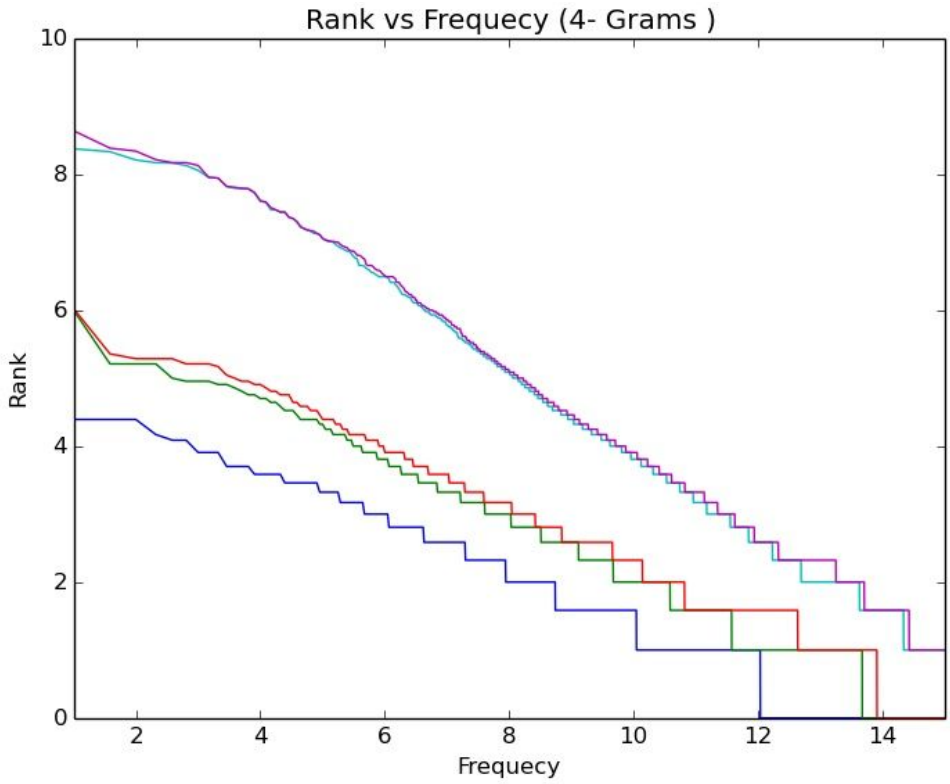
- **Bigram**



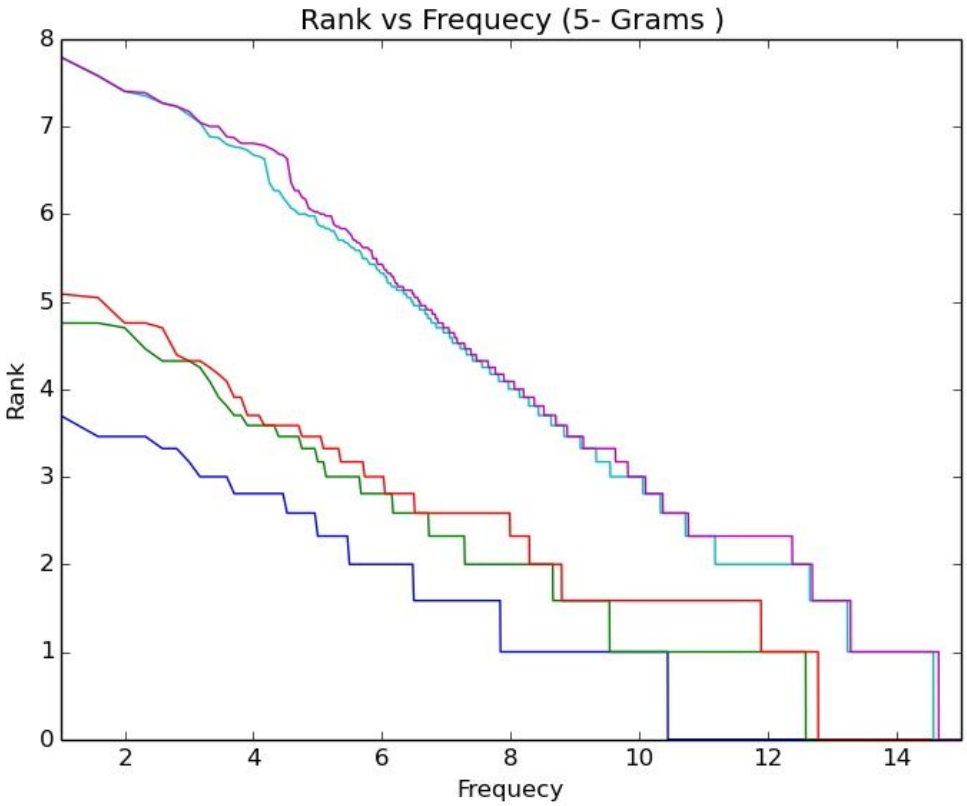
- **Trigram**



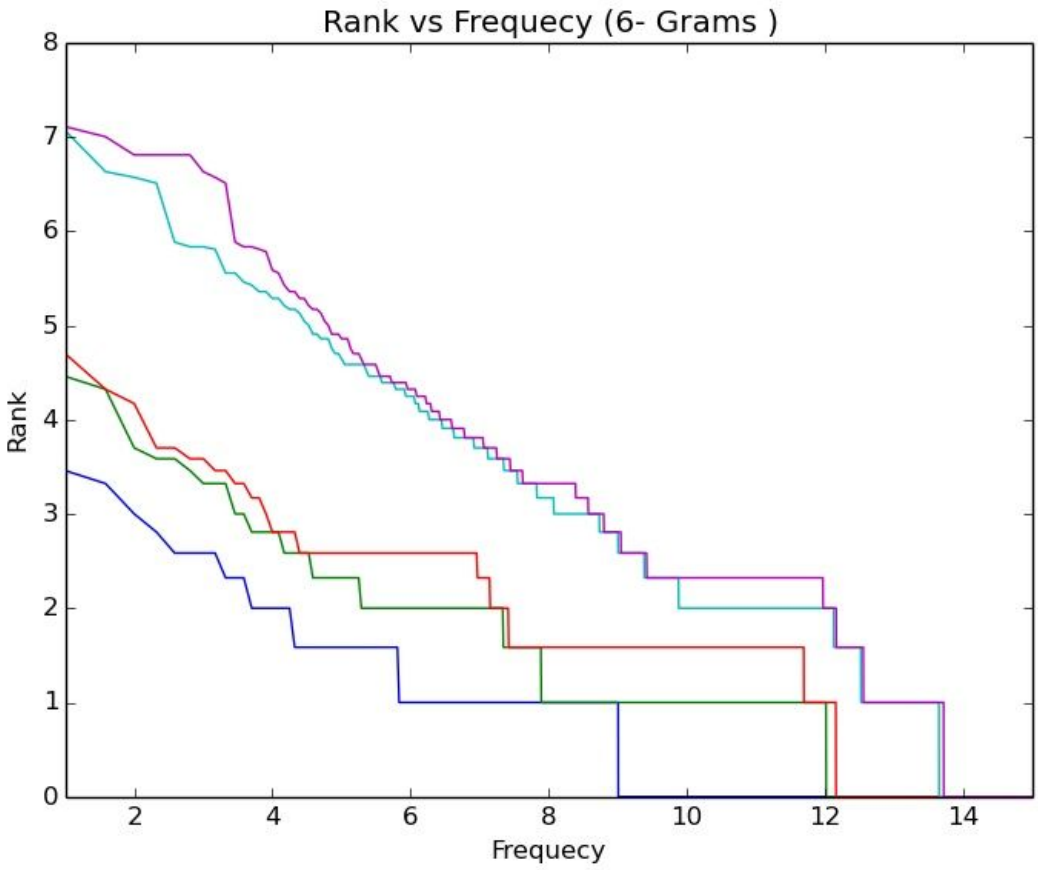
- **4 - gram**


Rank vs Frequecy (4- Grams)
Rank
Frequecy
0
2
4
6
8
10
12
14

- **5-Gram**


Rank vs Frequecy (5- Grams)
Rank
Frequecy
0
1
2
3
4
5
6
7
8
10
12
14

- **6-Gram**



Can we identify the author through the Zipf plot of the book ?

Yes. Different authors have different vocabulary in terms of the frequency of the words and there will also be a lot of non-common words with moderately-high frequency that could be used to classify the text or a book into their authors. The Zipf plot of the new book is likely to be similar to the Zipf curve of other books of the same author. Here, we will make the LM of the input text and compare it with the built models, higher the curve fitting, more chances of input text belonging to that author.