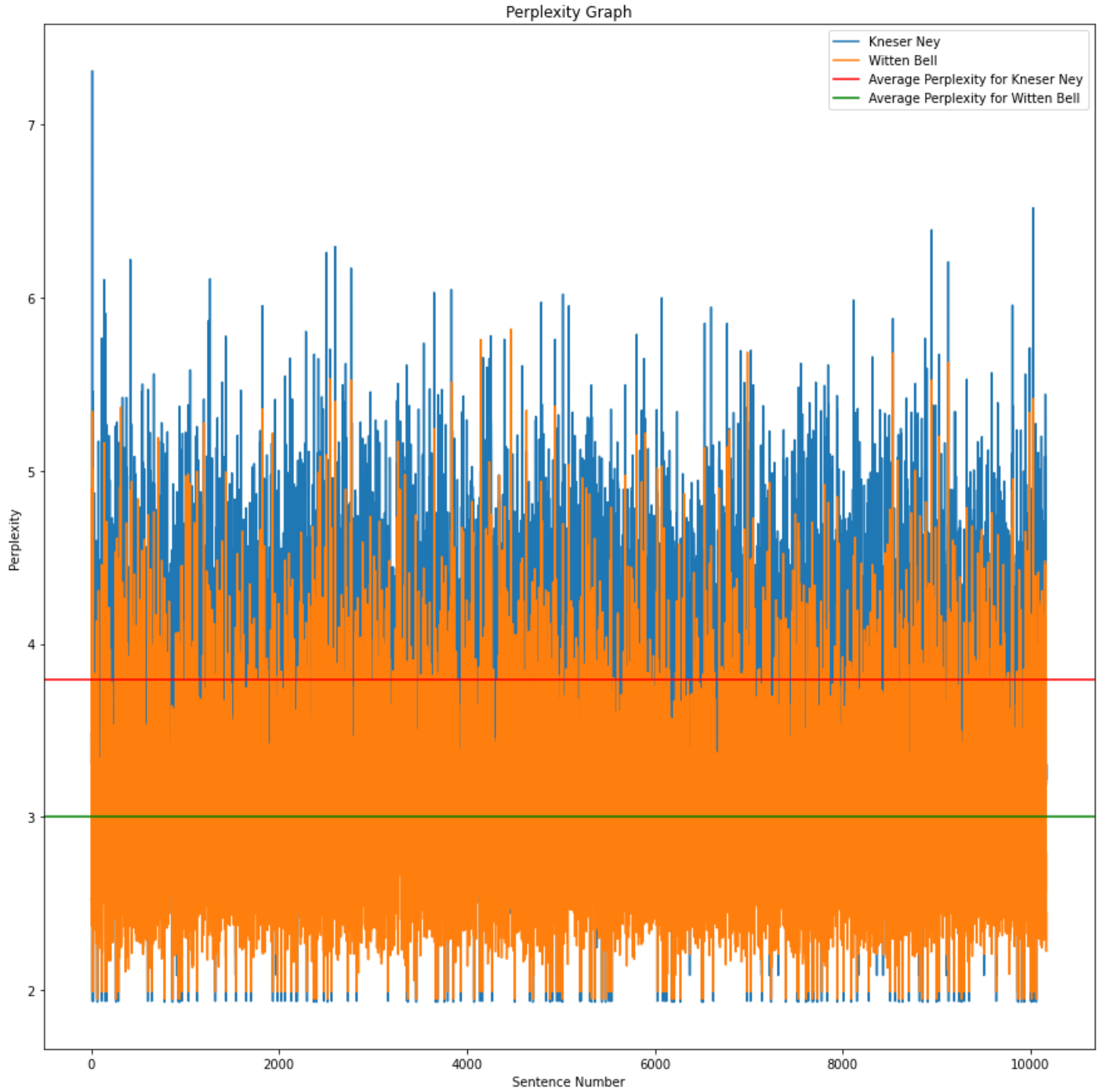
Report

Abhishek Sharma (2020101050)

Average Perplexities of Language Models for Training and Testing Sets

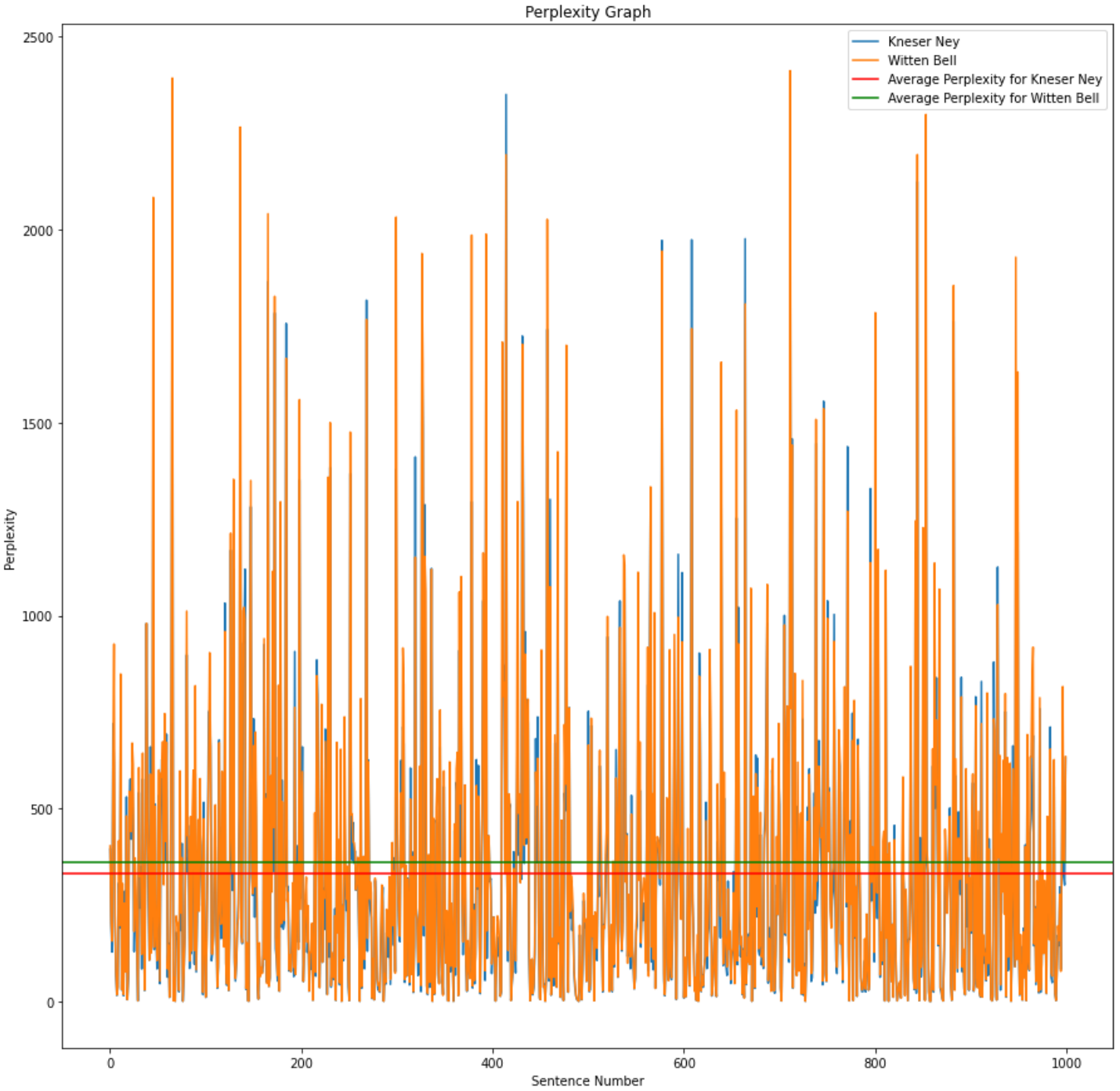
Model	Training Set	Testing Set
LM1	3.7962182265131403	333.5803939196564
LM2	4.132242178957028	639.1833368899953
LM3	3.0005709024527962	363.26914413343894
LM4	3.382472858056434	705.1494741178086
LM5	99.9233965899293	675.2893569838129
LM6	229.85889115195687	645.5694829013433

Corpus 1 : Pride and Prejudice - Jane Austen.txt



Average Perplexity for Kneser Ney: 3.7962182265131403

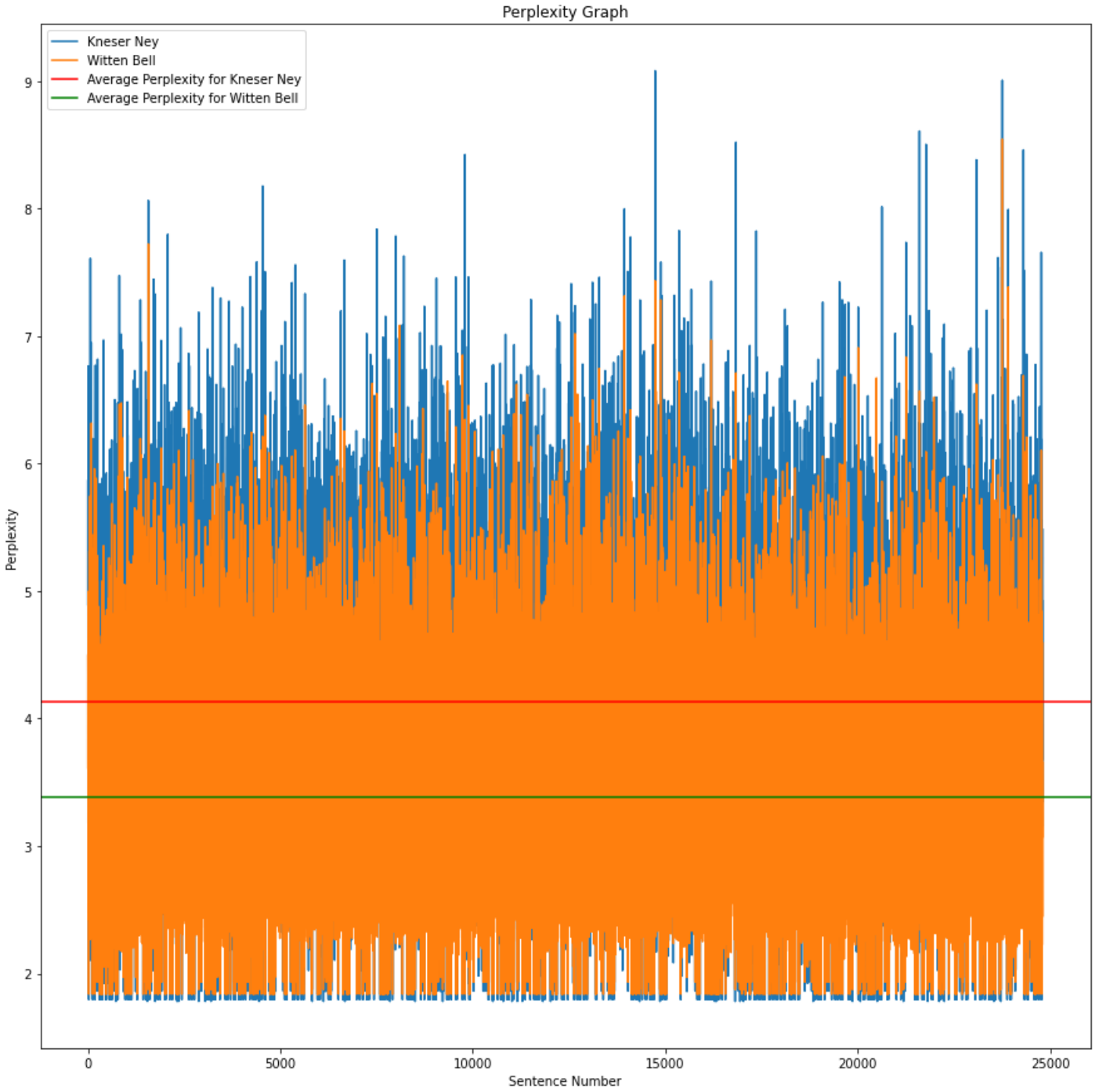
Average Perplexity for Witten Bell: 3.0005709024527962



Average Perplexity for Kneser Ney: 333.5803939196564

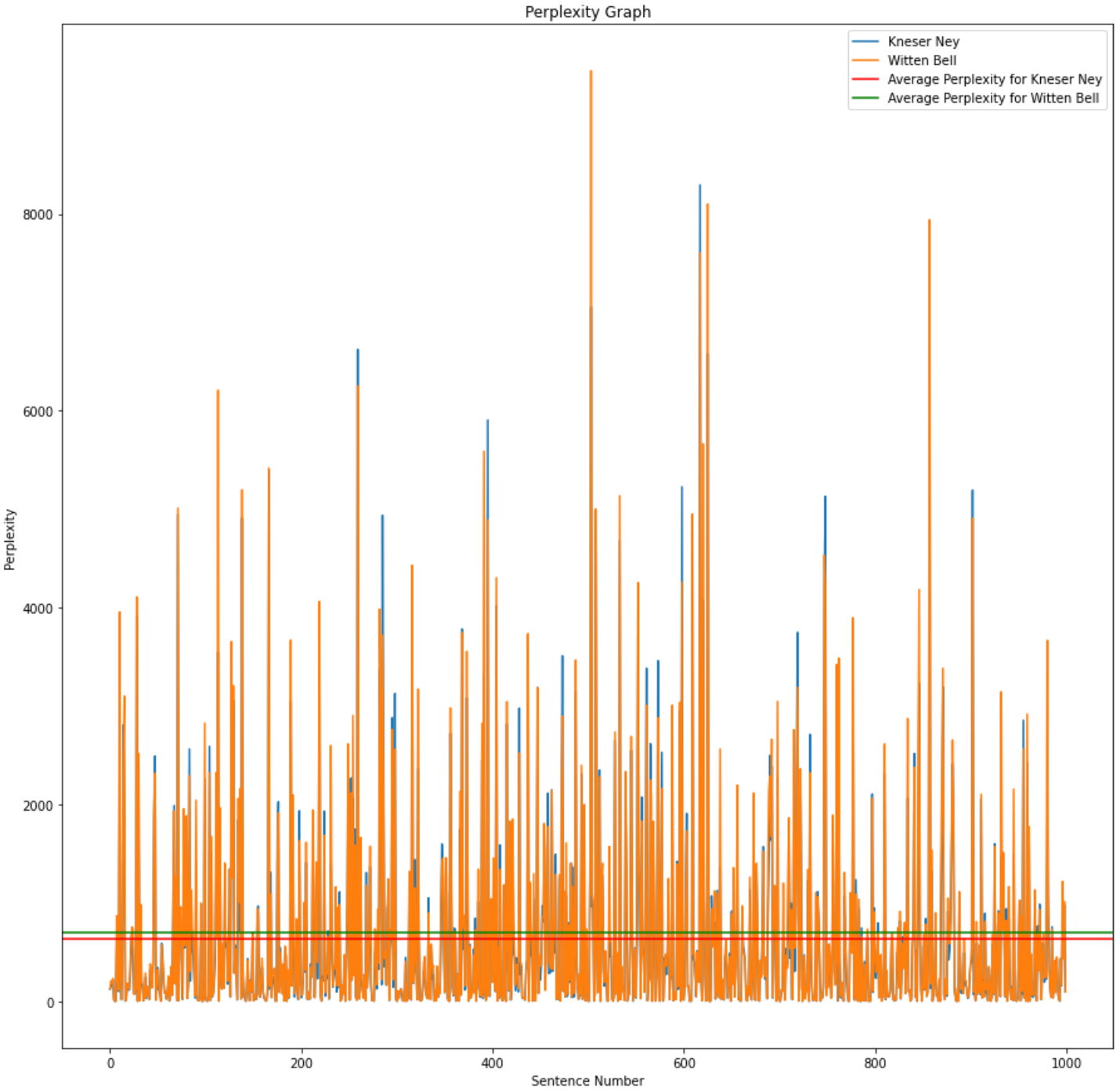
Average Perplexity for Witten Bell: 363.26914413343894

Corpus 2 : Ulysses- James Joyce.txt



Average Perplexity for Kneser Ney: 4.132242178957028

Average Perplexity for Witten Bell: 3.382472858056434



Average Perplexity for Kneser Ney: 639.1833368899953

Average Perplexity for Witten Bell: 705.1494741178086

Observation

1. For the first corpus, the Kneser Ney smoothing method has a lower perplexity value on the testing set, while the Witten Bell method has a lower perplexity value on the training set. This suggests that the Kneser Ney method is better suited for generalization to new data, while the Witten Bell method may be more effective in modeling the characteristics of the training data.
2. For the second corpus, the Kneser Ney smoothing method again has a lower perplexity value on the testing set, while the Witten Bell method has a lower perplexity value on the training set. This indicates that the observations from the first corpus hold true for the second corpus as well.
3. The higher perplexity values for the second corpus on both Kneser Ney and Witten Bell smoothing methods suggest that this corpus may be more difficult to model accurately than the first corpus. This may be due to a number of factors such as a larger vocabulary size, more complex sentence structures, or greater variability in the topics covered by the text.