

Assignment [3] :

- Take texts of 5 different authors (gutenberg.org) Or, 5 articles each from 5 different topics.
 - gutenberg.org has some the worlds most popular books.
 - Choose authors from different time/places/writing style.
- Do LM & Smoothing for each of these topics/authors separately
- Plot the log-log curve (frequency vs rank) of all these different LMs on to one graph
- Analyse the graph. What part of the zipf is same/different ?
- Also look at the zipf curve of each topic/author (this time Id vs Frequency). Looking at the differences, given a new distribution (a new zipf curve) can we classify the text as to which author it belongs to ?.