ACL - Seminar 3

Ichim Ștefan

Two texts from Project Gutenberg were analyzed by tokenizing them (lowercase, alphabetic characters only) and computing frequency statistics. The implementation used regular expressions (`re.findall`) for tokenization, `collections.Counter` for word frequency counting, and `matplotlib` for visualizing the frequency distributions.

1 Hamlet

Corpus Source	**Tokens**	**Types**	**Top 7 Words**	**Coverage & Hapax**
Hamlet by Shakespeare `https://www.gutenberg.org/files/1524/1524-0.txt`	32,789	4,547	the (1112), and (986), to (735), of (680), i (635), a (561), you (559)	Top 7 coverage: 16.07% Words appearing once: 2632, 8.03% of tokens, 57.88% of types

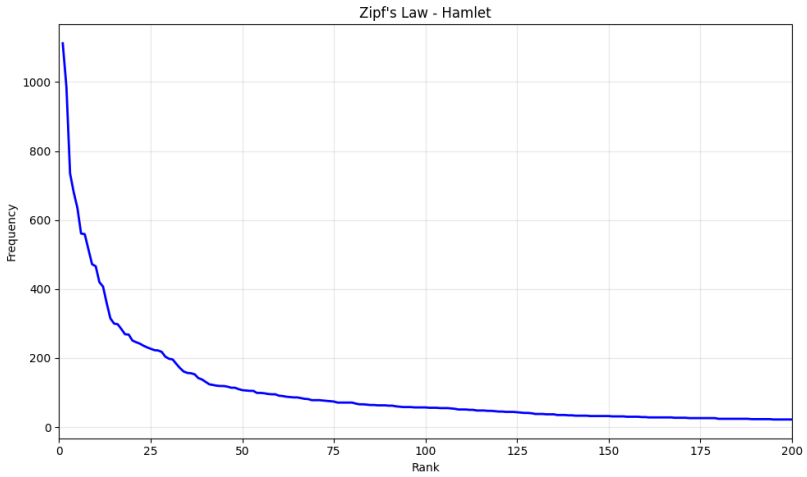

Figure 1: Frequency distribution for Hamlet

2 Don Quixote

Corpus Source	Tokens	Types	Top 7 Words	Coverage & Hapax
Don Quixote by Cervantes `https://www.gutenberg.org/files/996/996-0.txt`	433,477	15,586	the (22479), and (17719), to (14008), of (13493), that (7994), in (7338), a (7197)	Top 7 coverage: 20.81% Words appearing once: 5663, 1.31% of tokens, 36.33% of types

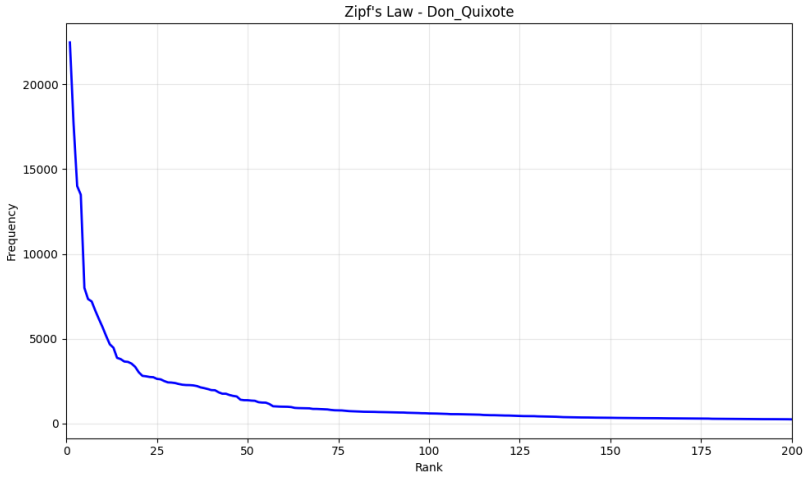

Figure 2: Frequency distribution for Don Quixote