Name: Arkapravo Ghosh NetID: ag2995

Project-1 Report

For vocab_size = 10,000

Document Type	Total #tokens	#Unique tokens	#Common tokens	#Exclusive Tokens
Legal	34238	3191	797	2394
Literature	73581	2615		1818

For vocab_size = 5,000

Document Type	Total #tokens	#Unique tokens	#Common tokens	#Exclusive Tokens
Legal	57710	1667	456	1211
Literature	104014	1261		805

This happens because, with a smaller vocabulary, the tokenizer has to break words down into smaller subword units to fit them within the vocabulary. As we reduce the vocab_size, the tokenizer is forced to split words into smaller units (subwords) more frequently because fewer whole words are part of the vocabulary. This increases the total number of tokens after tokenization.

For vocab_size = 20,000

The vocabulary size (obtained by printing leb(vocab)) = 14,917 (< 20,000). This means the maximum vocab_size that we can have is 14,917. If we set vocab_size >= 14,917, all the tokens in the tokenized texts will be exactly same as the original document (instead of having subwords with ## at the starting). As this is bad for model generalizability to new (unseen) words, we restrict ourselves to a smaller vocab_size = 10,000.

Final Results

Corpus contains 2495 sentences and 6104 unique words, when pre-tokenized with pre-trained **"**`bert-base-cased`**"** tokenizer which itself has a vocabulary size of 30,522.

Document Type	Total #tokens	#Unique tokens	#Common tokens	#Exclusive Tokens
Legal	34238	3191	797	2394
Literature	73581	2615		1818