Task 1: Tokenizer Training

Fertility Score Matrix:

Sno	Name of Tokenizer	Fertility Score	Dataset size
1.	BertWordPieceTokenizer	1.2671739693118529	6 GB (15000 files)
2.	ByteLevelBPETokenizer	3.2839517885373644	6 GB (15000 files)
3.	GPT2TokenizerFast	7.328868252078862	6 GB (15000 files)
4.	SentencePieceBPETokenizer	1.2406445450652963	6 GB (15000 files)
5.	SentencePieceUnigramTokenizer	1.432706148830655	6 GB (15000 files)
6.	SpaCyTokenizer	1.5319781839145085	6 GB (15000 files)