QUESTION - 6

	UNIGRAM-1000	UNIGRAM -1000	BPE-1000	BPE-2000	mBERT - 1000	mBERT - 2000	Indic-BERT -1000	Indic-BERT -2000	WHITE SPACE
PRECISION	0.017241379310344827	0.022988505747126436	0.011494252873563218	0.017241379310344827	0.005747126436781609	0.005747126436781609	0.011494252873563218	0.011494252873563218	0.040229885057471264
RECALL	0.0038412291933418692	0.005847953216374269	0.002472187886279357	0.004213483146067416	0.0012406947890818859	0.0012406947890818859	0.003189792663476874	0.003189792663476874	0.015486725663716814
F-SCORE	0.0062827225130890054	0.009324009324009324	0.004069175991861648	0.006772009029345373	0.0020408163265306124	0.0020408163265306124	0.004993757802746566	0.004993757802746566	0.022364217252396165

COMPARISON :

The PrecisIon, Recall and F-score value is highest for White space tokenizer because other tokenizers are giving tokens that are even dividing the single words. That is causing a large mismatch with the word tokens that I have marked in question 3 as per definition of word group. And white space tokenizer is words from spaces which increases its match from ground truth list. Hence, the true positive high for white space tokenizer.

But according to the description of tokenizers we should get the highest precision for Indic-BERT tokenizer as it is trained for Hindi language.