Question 6:

Number of Ground truth tokens: 203

1) **mBERT Tokenizer :**

   Total 850 mBERT2000 calculated tokens

   Precision: 0.0798611111111111

   Recall: 0.116751269035533

   F1_score 0.09484536082474226

   I used different max_length parameter (1000 and 2000) in the mBERT tokenizer, as given in the question, but number of tokens generated for both is same i.e. 850 tokens. So, length of input tokens might not be a critical factor for this tokenizer when dealing with Devanagari script. And by looking at such huge number of tokens, we can say that it generated tokens (1 to 3) characters as one tokens

2) **BPE Tokenizer:**
   Total 296 calculated tokens (using vocab_size:1000)
   Precision: 0.45318352059925093
   Recall: 0.6142131979695431
   F1_score 0.5215517241379309

   Total 290 calculated tokens (using vocab_size:1000)
   Precision: 0.47509578544061304
   Recall: 0.6294416243654822
   F1_score 0.5414847161572052

   It looks like BPE considered one word or oneword+some characters as one token. Also, high values of precision, recall, F_1 score indicates that it actually predicted many tokens as word or word groups. It could identify almost 63% of tokens as correct word-groups out of total word groups

3) **Unigram Tokenizer Performance Variation:**
   Total 852 tokens calculated using Unigram Tokenizer
   Precision: 0.06060606060606061
   Recall: 0.08121827411167512
   F1_score 0.06941431670281994

   As precision, recall, and F1 scores are very low by using Unigram tokenizer, we can say that it isn't a good tokenizer for Devnagari script. Probably it couldn't capture characteristics of script. And it generated such huge number of tokens.

4) **WhiteSpace Tokenizer:**

Total 288 tokens calculated using whiteSpace Tokenizer
Precision: 0.4864864864864865
Recall: 0.6395939086294417
F1_score 0.5526315789473685

The Whitespace tokenizer stands out with higher precision, recall, and F1 scores compared to other tokenizers. As we know it just tokenizes based on white spaces, it could identify words. However, because of this working technique , it totally misses word groups.

In summary, we can say that for Devnagari script, tokenizers which tokenizes using white space, or tokenizes a word as a token performs good as tokenizer. However, even though whitespace tokenizer worked good in comparison with other tokenizers, it couldn't identify word groups. So, sometimes tokens generated by whitespace tokenizer may not make any sense alone. But that token may make sense when combined with some of the previous words. But, in comparison with mBERT, BPE, Unigram going with the simpler approach of whitespace tokenizer makes sense.