Name: _____

Exercise 11, makeup 11/18/21 – UMass CS 490A
Turn in via Gradescope by next Monday

Let's explore the subword-based tokenization strategies of large, pre-trained language models.

Use the following Google CoLab notebook:
https://colab.research.google.com/drive/1v3iustM3huxMVSItknowzWGwdrQpZoco

Part I. Experiment with the tokenizers of BERT, GPT-2, and RoBERTa.

A. What are some **small** words that are segmented into **multiple** tokens by the tokenizers? Identify 2-3 words that are segmented into multiple tokens by **all** three tokenizers, as well as 2-3 words that are segment into multiple tokens by **1 or 2 (but not all)** of the tokenizers. Include both the word and their tokenizations in your answer.

B. What are **long** words that are **not split** into multiple tokens (i.e., remains one token) by the tokenizers? Identify 2-3 words that are not split by **any** of the three tokenizers, as well as 2-3 words that are not split by **1 or 2 (but not all)** of the tokenizers. Include both the word and their tokenizations in your answer.

Part II.

C. Which of these models use the same tokenizer?

D. Which tokenization strategy seems better? Why?