Name: _____ ID: _____

Exercise 3 Tokenization

Tokenize the following sentences. Go through each one and perform manual tokenization before doing so with word_tokenize(). Report both answer types and discuss any shortcomings of each approach (e.g., words that should not have been split) and/or discrepancies you have observed between the two approaches.

1. What you don't want to be done for yourself, don't do to others.

   Manual:

   NLTK:

2. As San Francisco begins proof-of-vaccine mandate, here's how to show proof from your phone:

   Manual:

   NLTK:

3. Dial 202-123-4567 and ask to talk to Ela Israeli to claim your €100. This offer expires on March 21, 2022. Email abc@ggk.eu for more information. Email verification code: 456-666

   Manual:

   NLTK:

4. The maximal effect is observed at the IL-10 concentration of 20 U/ml. The transcripts were detected in all the CD4− CD8−, CD4+ CD8+, CD4+ CD8−, and CD4− CD8+ cell populations.

   Manual:

   NLTK:

5. If your home has "hard water" (i.e., a high mineral content), your sinks, showers, and tubs no doubt bear white or yellow buildup as a result.

   Manual:

   NLTK:

Discussion: shortcomings and/or discrepancies:

NOTE: I worked with [NAME] _____ to complete this exercise.