CSSE 315 – Natural Language Processing
Rose-Hulman Institute of Technology

# Quiz 02 SOLUTIONS

Name (Print):______________________ Date:__________

1. Which of the following is NOT a common tokenization technique in NLP?
   - a.) Word-level tokenization
   - b.) **Sentence-level tokenization**
   - c.) Character-level tokenization
   - d.) Subword-level tokenization

2. What is a major advantage of subword-level tokenization over word-level tokenization?
   - a.) It requires less memory
   - b.) **It handles out-of-vocabulary words better**
   - c.) It is faster to compute
   - d.) It is easier to implement

3. The word ”studies” is transformed into ”studi”. Is this an example of stemming or lemmatization? Explain your answer: **stemming, it is not a dictionary form**

4. In the traditional NLP pipeline, we include stemming/lemmatization. Why does the newer NLP pipeline omit this step? Explain your answer: **subword techniques or character tokenization do not include stemming/lemmatization**

5. What is the primary purpose of feature engineering in NLP?
   - a.) To reduce the size of the text data
   - b.) To improve the readability of the text data
   - c.) To correct grammatical errors in the text data
   - d.) **To represent text data in a numerical format suitable for machine learning algorithms**

6. Which of the following statements about Count Frequency (CF) is TRUE?
   - a.) CF considers the length of the document (total of all tokens)

- b.) **CF is useful for identifying words that appear frequently in a specific document**
- c.) CF value is always smaller than TF (Term Frequency)
- d.) CF is a technique to measure the rarity of a term

7. Provide an example of unigram, bi-gram, and tri-gram in the following sequence:

   The quick brown fox jumps over the lazy dog

   - a.) Unigram: the
   - b.) Bi-gram: brown fox
   - c.) Tri-gram: quick brown fox

8. Which tokenization method is currently used by GPT models?

   - a.) Word-Level
   - b.) Character-Level
   - c.) **Byte-Pair Encoding**
   - d.) Syntax-Level

9. Identify which vector space(s) consider sparse and which one(s) are dense

   - a.) Image vector (pixels): dense
   - b.) Audio spectrogram (sound waves): dense
   - c.) Word vectors: sparse

10. Which of the following statements about stop words is FALSE?

    - a.) Stop words are usually the most frequent words in a document
    - b.) **Stop words carry significant semantic meaning in most NLP tasks**
    - c.) Stop words are often function words like “the”, “a”, and “is”
    - d.) Removing stop words improves the performance of NLP models

11. Why is TF-IDF often preferred over just using Term Frequency (TF) in NLP tasks?

    - a.) TF-IDF is simpler to calculate than TF
    - b.) TF-IDF gives higher weight to very common words
    - c.) **TF-IDF helps distinguish important words from common (frequent) words**
    - d.) TF-IDF gives smaller weight to rare words

12. Which tokenization technique generally requires the largest vocabulary size?

    - a.) Character-level tokenization
    - b.) **Word-level tokenization**
    - c.) Subword-level tokenization