Name: ___________________ **Roll No:** _____________

Natural Language Processing (NLP DS-A) **Dr. Zohair Ahmed**

Activity Sheet 1 (Solution): Text Representation Techniques in NLP (BoW, TF-IDF, N-grams)

Part 1: Bag of Words (BoW)

Answer 1: BoW Representation

- **Vocabulary**:
  ["I", "love", "programming", "Python"]
- **BoW Vectors**:
  - Text 1 ("I love programming"): [1, 1, 1, 0]
  - Text 2 ("I love Python programming"): [1, 1, 1, 1]

Explanation:

- The vector represents the presence of words from the vocabulary.
- "I", "love", and "programming" appear in both texts, but "Python" only appears in Text 2.

Part 2: TF-IDF (Term Frequency-Inverse Document Frequency)

$$\text{TF-IDF}(w, d) = TF(w, d) \times IDF(w)$$

$$IDF(w) = \text{lo g}\left(\frac{N}{df(w)}\right)$$

Answer 2: TF-IDF Calculation

- **Step 1**: **TF** for "fox" in **Document 1**:
  - The word "fox" appears once in Document 1, so $TF(\text{fox}) = \frac{1}{4} = \mathbf{0.25}$
- **Step 2**: **df** for "fox":
  - "fox" appears in **Documents 1 and 2**, so $\boldsymbol{df(\text{"}fox\text{"}) = 2}$.
- **Step 3**: **TF-IDF** for **"fox"** in Document 1 and **N = 3** (total documents)
  - Using the formula:

$$\text{TF-IDF}(\text{fox}) = 0.25 \times \log\left(\frac{3}{2}\right) = 0.25 \times \log(1.5)$$

$$0.25 \times 0.1761 \approx \mathbf{0.044}$$

Part 3: N-grams

Answer 3: N-grams Generation

- **Unigrams (1-grams)**:
  - ["Data", "science", "is", "the", "future", "of", "technology"]
- **Bigrams (2-grams)**:
  - ["Data science", "science is", "is the", "the future", "future of", "of technology"]
- **Trigrams (3-grams)**:
  - ["Data science is", "science is the", "is the future", "the future of", "future of technology"]

Part 4: Implementing BoW and TF-IDF in Python

Answer 4: Code Implementation Outputs

1. **BoW Representation (CountVectorizer)**:
   - **Vocabulary**: ["I", "am", "learning", "NLP"]
   - **BoW Matrix**:
   - [1, 1, 1, 1]
   - [0, 1, 0, 1]

2. **TF-IDF Representation (TfidfVectorizer)**:
    - **TF-IDF Matrix**:
    - [0.577, 0.577, 0.577, 0.577]
    - [0.577, 0.577, 0.577, 0.577]

Explanation:

- For the BoW vector, each row represents the frequency of words in the sentence. The first sentence contains "I", "am", "learning", and "NLP", so their values are 1 (they appear once).
- The TF-IDF matrix represents the importance of each word within the document, with higher values indicating greater importance (words that are not common in the whole corpus).

Part 5: Reflection

Answer 5: Short Answer

1. **Difference between TF-IDF and BoW?**
    - **Answer**: **BoW** counts word frequencies without considering how common or rare the words are in the entire corpus, while **TF-IDF** adjusts the frequency by considering the inverse document frequency, giving less importance to common words and more to rare words. TF-IDF helps in distinguishing important words that are unique to each document.
2. **How do N-grams help in capturing context? Give an example.**
    - **Answer**: N-grams capture relationships between consecutive words, allowing the model to understand the context better. For example, in the bigram "data science", the context of both words together gives more meaning than treating "data" and "science" separately as unigrams.