Assignment for Session 3: Text Processing and Representation with Word2Vec

---

Objective

Create a personalized Jupyter Notebook implementing **text preprocessing** and **Word2Vec-based representation** for two sample texts—one in **Arabic** and one in **English**. The texts should contain typical text-related issues to showcase the use of preprocessing methods to clean and prepare the data.

Assignment Details

1. Choose Your Focus

For this assignment, you are required to:

- **Preprocess** two texts—one in **Arabic** and one in **English**—containing text issues like misspellings, mixed languages, and punctuation problems.
- **Represent** the cleaned text using **Word2Vec** embeddings.

---

2. Tasks to Implement

1. **Preprocessing (40 Marks)**

    - Apply suitable techniques to clean the texts.
    - Techniques to use include:
        - **Normalization**: Convert text to a consistent format.
        - **Tokenization**: Split text into individual words or subword tokens.
        - **Stopword Removal**: Remove unnecessary words that don’t contribute to meaning.
        - **Spelling Correction**: Fix misspellings and abbreviations.
        - **Handling Mixed Languages**: Handle Arabic and English words properly.

    Output: Provide cleaned and processed versions of both texts (Arabic and English).

2. **Word2Vec Representation (40 Marks)**

    - Train a **Word2Vec model** on the cleaned texts.
    - Extract embeddings for individual words from the Word2Vec model.

    Output: Show word embeddings for at least five words from both the Arabic and English texts.

3. **Optional (Creativity Bonus - 20 Marks)**

    - Visualize the embeddings using **t-SNE** or **PCA** for a 2D scatter plot of the word vectors.

---

3. Explain Your Code

- **Markdown Explanation:** Add detailed explanations in markdown cells to clarify each step of your approach.
- **Justification:** Discuss why certain preprocessing steps were necessary and how they align with the session's concepts.
- **Reflection:** Reflect on the challenges you faced and how you addressed them.