**Text Similarity Techniques**

**Objective**

The goal of this task was to investigate and compare different approaches for measuring the similarity between two texts. These approaches include both syntactic methods (focused on word form) and semantic methods (focused on meaning).

We evaluated the following techniques:

* **Cosine Similarity** (syntactic)
* **Jaccard Similarity** (syntactic)
* **BERT-based Semantic Similarity** using SentenceTransformer

**Input Texts**

* Text 1: *"The astronaut prepared for the mission with intense training."*
* Text 2: *"Before flying to space, the cosmonaut trained rigorously."*

These sentences convey similar ideas but use very different wording. This makes them ideal for evaluating how each method handles both surface-level and deeper linguistic similarity.

**Methods Used**

1. **Cosine Similarity (TF-IDF Based)**
   * Used TfidfVectorizer() to convert words into weighted vectors based on term frequency and inverse document frequency.
   * Applied cosine\_similarity() from Scikit-learn to compare the vectors.
2. **Jaccard Similarity (Set-Based)**
   * Lowercased and tokenized both texts into sets of unique words.
   * Computed similarity using set intersection and union:
     + A & B for common elements
     + A | B for total unique elements
   * Limitation: Only measures shared word presence, ignoring meaning.
3. **BERT Semantic Similarity**
   * Used the lightweight **all-MiniLM-L6-v2** model from the sentence-transformers library.
   * Encoded each sentence into a 384-dimensional vector.
   * Calculated similarity with util.cos\_sim(), which is optimized for deep learning tensors.
   * Advantage: Captures meaning even when the vocabulary is entirely different.

**Results Overview**

| **Method** | **Score** | **Type** |
| --- | --- | --- |
| Cosine Similarity (TF-IDF) | 0.1230 | Syntactic |
| Jaccard Similarity | 0.6667 | Syntactic |
| BERT Semantic Similarity | 0.6187 | Semantic |

**Analysis**

* **Cosine Similarity (0.1230)**
  + Very low score due to minimal word overlap.
  + Focuses only on shared terms, missing semantic relationships.
* **Jaccard Similarity (0.6667)**
  + Surprisingly high score due to common function words like "the" and "for".
  + Ignores meaning, structure, or context.
* **BERT Similarity (0.6187)**
  + Accurately recognizes semantic equivalence between phrases like "astronaut" and "cosmonaut", or "training" and "trained".
  + Best suited for applications like paraphrase detection or textual entailment.

**Conclusion**

* Syntactic techniques such as Cosine and Jaccard are limited, especially when different words are used to express the same idea.
* BERT-based models, though more complex, offer far more meaningful insights and align better with human judgment.
* For practical NLP applications requiring true understanding, semantic models like BERT are the preferred choice.