### ****Text Similarity****

In this exercise, we compared different methods of computing similarity between product texts, using three distinct approaches: **TF-IDF + cosine similarity**, **Jaccard**, and **BERT embeddings**. The analyzed texts are Nike product descriptions, containing marketing or technical terms related to shoes, apparel, and sports accessories.

#### ****1. TF-IDF + Cosine Similarity****

The TF-IDF method followed by a cosine similarity measure helps detect lexical overlaps between texts.

* Some products show moderately high scores, especially those whose descriptions share frequent commercial terms like “shoe”, “Nike SB”, “durability”, “comfort” or “performance”.
* This suggests the method is effective in spotting texts with common words, even when those words appear in different contexts or positions.
* However, when the vocabulary changes even for similar products similarity drops significantly. This approach is therefore very sensitive to wording variations, limiting its ability to detect deeper semantic similarities.

#### ****2. Jaccard Similarity****

Jaccard similarity, applied to word sets (or n-grams), gives overall much lower scores.

* This method only considers the proportion of **exact word intersection** between two texts, which penalizes alternative phrasings.
* Even for two very similar products, like two pairs of skate shoes, the score may be low if the exact words differ.
* This confirms that the Jaccard method is **too strict** in a linguistically rich context like marketing, where lexical variety is high.

#### ****3. BERT Embeddings****

The BERT-based method offers a radically different perspective:

* It encodes the **overall meaning** of sentences rather than just word co-occurrences.
* Thanks to semantic embeddings, BERT can group texts that talk about the **same type of product**, even if the wording differs significantly.
* For example, two descriptions evoking similar ideas of comfort or sports performance will be considered close, even if one talks about “optimal support” and the other about “secure fit”.
* This capacity to recognize **semantic proximity rather than surface form** makes BERT far more robust for tasks where **meaning matters more than wording**.

### ****Conclusion****

* **TF-IDF** effectively captures lexical similarity but is limited when vocabulary changes.
* **Jaccard** proves too rigid in a context where texts are short and phrasing is varied.
* **BERT**, on the other hand, offers a much finer analysis of meaning, making it the most relevant tool to identify similarities between product descriptions in real scenarios.