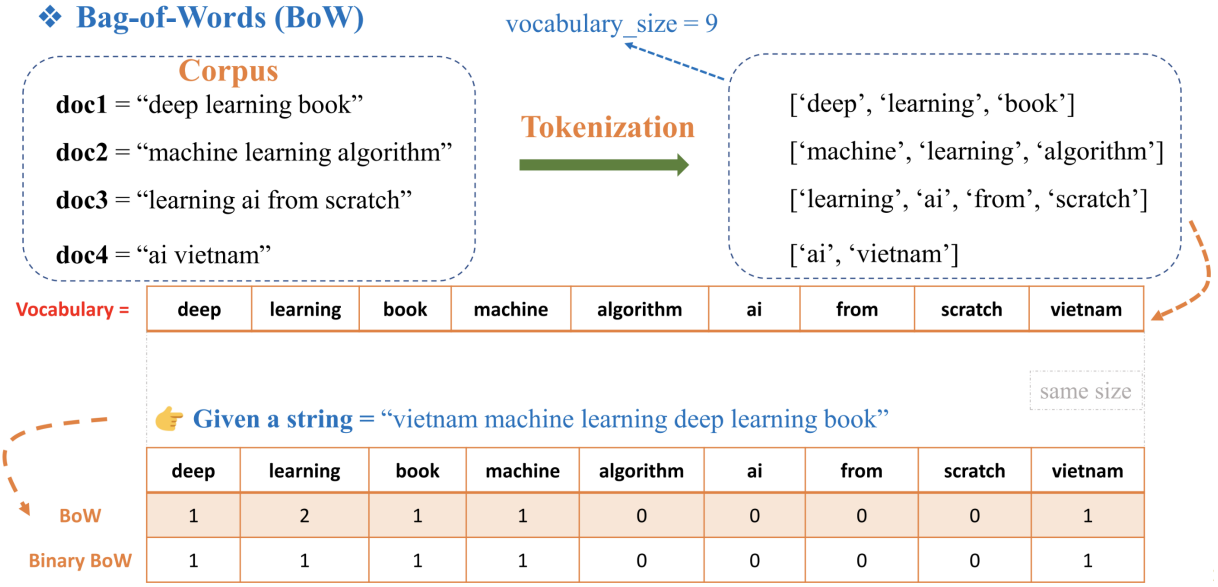
Section 7 - Bag of Words (NLP)

Hoàng-Nguyên Vũ

1. **Mô tả:**

   - **Bag of Words** là một thuật toán hỗ trợ xử lý ngôn ngữ tự nhiên và mục đích của BoW là phân loại text hay văn bản. Ý tưởng của BoW là phân tích và phân nhóm dựa theo "Bag of Words" (corpus). Với test data mới, tiến hành tìm ra số lần từng từ của test data xuất hiện trong "Bag". Cách thức thực hiện như sau:
     - **Bước 1:** Chia nhỏ văn bản thành các từ riêng lẻ.
     - **Bước 2:** Tạo một tập hợp các từ xuất hiện trong văn bản. Tập hợp này không có phần tử trùng nhau.
     - **Bước 3:** Biểu diễn văn bản input ở dạng vector: Mỗi câu (mỗi input) được biểu diễn bằng một vector, với mỗi phần tử trong vector thể hiện số lần xuất hiện của từ đó trong input.



2. **Bài tập:** Tạo Bag-Of-Word cho tập dataset sau: *corpus* = ["Tôi thích môn Toán", "Tôi thích AI", "Tôi thích âm nhạc"]. Sau đó tạo list có tên *vector* để lưu vector sau khi thực hiện bước Tokenization đoạn văn bản sau: **Tôi thích AI thích Toán**, biết Bag-Of-Word được sắp theo thứ tự tăng dần

```
corpus = [''' Your Code Here ''']
# Your code here
```

Output:
- **Tôi thích AI thích Toán:** [1, 1, 1, 0, 0, 2, 0]
- **Bag-of-Words:** [AI, Toán, Tôi, môn, nhạc, thích, âm]