Personal Note 04: Text Embeddings

1. Text Embedding

- **Định nghĩa:**
  - Là kỹ thuật được sử dụng để biểu diễn văn bản dưới dạng các vector số học trong không gian nhiều chiều.
  - Thay vì biểu diễn văn bản dưới dạng chuỗi các từ hoặc ký tự, Text Embedding chuyển đổi các đơn vị văn bản như từ, câu, đoạn văn, hoặc thậm chí toàn bộ tài liệu thành các vector.
  - Các vector này chứa thông tin về ngữ nghĩa, ngữ cảnh, và cấu trúc của văn bản.
- **Mục tiêu:**
  - Tạo ra một biểu diễn số học của văn bản sao cho các văn bản có ý nghĩa tương tự sẽ có các vector gần nhau trong không gian vector.
  - Giúp các mô hình học máy có thể dễ dàng phân tích và xử lý văn bản trong nhiều bài toán như phân loại văn bản, tìm kiếm văn bản, dịch máy, và tóm tắt văn bản.
- **Đặc điểm:**
  - Text Embedding giúp nắm bắt được **ngữ nghĩa** của các từ, cụm từ, hoặc câu dựa trên ngữ cảnh mà chúng xuất hiện.
  - Mỗi văn bản được biểu diễn bằng một vector trong không gian nhiều chiều (thường từ hàng chục đến hàng trăm chiều).
- **Các mô hình tạo nên Text Embeddings:**
  - Word2Vec
  - Average Word2Vec
  - Doc2Vec

2. Word2Vec

- **Mục tiêu:** tạo ra các **word embedding**
    - Biểu diễn các từ trong không gian nhiều chiều sao cho các từ có ý nghĩa tương tự hoặc xuất hiện trong cùng ngữ cảnh sẽ nằm gần nhau trong không gian này.
    - Giúp các mô hình học máy hiểu và xử lý ngữ nghĩa của từ dễ dàng hơn.
- Hai mô hình chính trong Word2Vec:
    - **Continuous Bag of Words (CBOW)**:
        - Dự đoán **từ trung tâm (target word)** dựa trên **ngữ cảnh xung quanh (context)**.
        - Ví dụ: Cho câu "Tôi thích uống cà phê buổi sáng", CBOW sẽ sử dụng các từ "Tôi", "thích", "cà phê", "buổi", "sáng" để **dự đoán từ "uống"**.
    - **Skip-gram**:
        - Dự đoán các từ xung quanh dựa trên từ trung tâm.
        - Ví dụ: Với từ trung tâm "uống" trong câu "Tôi thích uống cà phê buổi sáng", Skip-gram sẽ dự đoán các từ "Tôi", "thích", "cà phê", "buổi", "sáng".
- Các lưu ý trong fine-tune model Word2Vec:
    - vector_size: thường nằm trong khoảng 100 hoặc 300.
    - vector_size lớn hơn cho ra kết quả tốt hơn vì đa dạng và tăng cường đặc trưng của từ
    - window thường được set là 5 hoặc 7.
- Avg Word2Vec giúp tính trung bình các vector của từ trong một văn bản. Sử dụng Avg Word2Vec giúp cho văn bản sử dụng duy nhất một vector đặc trưng (đại diện).

3. Mở rộng: Doc2Vec

- **Mục tiêu:** tạo ra các vector biểu diễn cho toàn bộ **đoạn văn** hoặc **tài liệu**, thay vì chỉ biểu diễn các từ đơn lẻ. Nó là một kỹ thuật học phân phối tương tự như Word2Vec, nhưng nó bổ sung thêm một vector đại diện cho cả đoạn văn (document) bên cạnh vector của từ.
- **Doc2Vec** có hai phương pháp phổ biến để huấn luyện:

- **Distributed Memory (DM)**: Phương pháp này giống như mô hình **CBOW** trong Word2Vec. Vector của đoạn văn và vector của các từ trong đoạn văn được sử dụng cùng nhau để dự đoán từ tiếp theo trong một ngữ cảnh.
- **Distributed Bag of Words (DBOW)**: Phương pháp này giống như mô hình **Skip-gram** trong Word2Vec. Ở đây, mô hình chỉ sử dụng vector của đoạn văn để dự đoán các từ trong đoạn văn.