Ôn tập kiến thức trước Lab 1 & Lab 2

Giảng viên: Phạm Ngọc Hải

September 16, 2025

Giới thiệu

- Tuần này chúng ta sẽ thực hành Lab 1 (Text Tokenization) và Lab 2 (Count Vectorization).
- Đây là những kiến thức nền tảng quan trọng trong Xử lý Ngôn ngữ Tự nhiên (NLP).
- Buổi ôn tập này sẽ giúp các bạn củng cố lại các khái niệm cơ bản trước khi bắt đầu thực hành.

1.1. Tokenization là gì?

- Quá trình chia một đoạn văn bản thành các đơn vị nhỏ hơn gọi là **token**.
- Token có thể là từ, cụm từ, ký tự, hoặc các đơn vị có ý nghĩa khác.
- **Ví dụ:** "Hello, world!" $\rightarrow$ ["Hello", ",", " ", "world", "!"] (tùy thuộc vào cách tách).

1.2. Tại sao Tokenization quan trọng?

- Là bước tiền xử lý cơ bản và thiết yếu trong hầu hết các tác vụ NLP.
- Giúp chuẩn hóa dữ liệu, giảm độ phức tạp của văn bản.
- Tạo ra các đơn vị đầu vào cho các mô hình NLP tiếp theo.

1.3. Các loại Tokenization phổ biến

- **Tách theo khoảng trắng (Whitespace Tokenization):** Đơn giản nhất, chia văn bản dựa trên các ký tự khoảng trắng.
  - *Nhược điểm:* Không xử lý được dấu câu dính liền từ, từ ghép.
- **Tách theo dấu câu (Punctuation Tokenization):** Tách riêng các dấu câu ra khỏi từ.
  - **Ví dụ:** "Hello, world!" $\rightarrow$ ["Hello", ",", "world", "!"]
- **Tách dựa trên biểu thức chính quy (Regex Tokenization):** Sử dụng các quy tắc regex để định nghĩa token.
  - Linh hoạt và mạnh mẽ hơn, có thể xử lý nhiều trường hợp phức tạp.
  - **Ví dụ regex:**
    `w+|[`
    `w`
    `s]` (tách từ hoặc các ký tự không phải từ/khoảng trắng).

2.1. Biểu diễn văn bản là gì?

- Chuyển đổi dữ liệu văn bản (phi cấu trúc) thành các định dạng số (có cấu trúc) mà các thuật toán học máy có thể hiểu và xử lý được.
- Văn bản không thể trực tiếp đưa vào mô hình học máy.

2.2. Giới thiệu mô hình Bag-of-Words (BoW)

- Là một mô hình đơn giản để biểu diễn văn bản.
- Mỗi văn bản được xem như một "túi" các từ, không quan tâm đến ngữ pháp hay thứ tự từ.
- Chỉ quan tâm đến việc từ nào xuất hiện và tần suất xuất hiện của chúng.

2.3. CountVectorizer hoạt động như thế nào?

- **Bước 1: Xây dựng từ vựng (Vocabulary Building):**
  - Duyệt qua toàn bộ tập hợp các văn bản (corpus).
  - Thu thập tất cả các từ (token) duy nhất xuất hiện trong corpus.
  - Gán một chỉ số (index) duy nhất cho mỗi từ trong từ vựng.
  - **Kết quả:** Một từ điển ánh xạ từ $\rightarrow$ chỉ số (ví dụ: `"hello": 0, "world": 1, ...`).
- **Bước 2: Chuyển đổi văn bản thành vector (Document-Term Matrix):**
  - Với mỗi văn bản trong corpus:
    - Tạo một vector có kích thước bằng số lượng từ trong từ vựng.
    - Mỗi vị trí trong vector tương ứng với một từ trong từ vựng.
    - Giá trị tại vị trí đó là số lần từ tương ứng xuất hiện trong văn bản.
  - **Kết quả:** Một ma trận mà mỗi hàng là một vector biểu diễn một văn bản, và mỗi cột là một từ trong từ vựng.

2.4. Ví dụ đơn giản

- **Corpus:**
  1. "I love NLP"
  2. "I love programming"
  3. "NLP is a subfield of AI"
- **Từ vựng (sau khi tokenization và xây dựng):**
  - `"i": 0, "love": 1, "nlp": 2, "programming": 3, "is": 4, "a": 5, "subfield": 6, "of": 7, "ai": 8`
- **Ma trận Document-Term:**
  1. `[1, 1, 1, 0, 0, 0, 0, 0, 0]` (I love NLP)
  2. `[1, 1, 0, 1, 0, 0, 0, 0, 0]` (I love programming)
  3. `[0, 0, 1, 0, 1, 1, 1, 1, 1]` (NLP is a subfield of AI)

Chúc các bạn ôn tập hiệu quả và thực hành tốt!