WEEK 2: TEXT PREPROCESSING

a) Kỹ thuật text preprocessing là gì?

Text preprocessing là tổng hợp các bước xử lý dữ liệu thô trước khi đưa vào mô hình học máy hoặc LLM để làm sạch, chuẩn hóa và biến đổi văn bản thành dạng mà máy tính có thể hiểu và khai thác được.

b) Vai trò và tầm quan trọng của các kỹ thuật này. Tại sao cần tiền xử lý text ?

- **Vai trò:**
    - Làm sạch dữ liệu: loại bỏ ký tự thừa, emoji, link, stopwords,…
    - Chuẩn hóa văn bản: chuyển chữ viết hoa thành viết thường, stemming/lemmatization.
    - Biến đổi: chuyển text thành vector/embedding để máy tính hiểu.
    - Giúp dữ liệu gọn nhẹ, rõ nghĩa, thuận lợi cho việc trích xuất đặc trưng.
- **Tầm quan trọng:**
    - Tiết kiệm tài nguyên tính toán (bộ nhớ, thời gian huấn luyện).
    - Tránh mô hình học sai do dữ liệu nhiễu, không thống nhất.
    - Giúp mô hình tăng độ chính xác và khả năng tổng quát hóa.
- **Lý do cần xử lý text:**
    - Văn bản thực tế chứa nhiều nhiễu, ký tự đặc biệt, từ viết tắt, lỗi chính tả.
    - Ngôn ngữ tự nhiên không đồng nhất (một từ nhiều dạng khác nhau).

    - Máy tính không hiểu trực tiếp chữ, cần biến đổi sang dạng số.

c) Mô tả ngắn gọn những kỹ thuật text preprocessing truyền thống. Cho ví dụ và đoạn code kết quả dựa trên dataset cung cấp.

- Các kỹ thuật text preprocessing truyền thống:
    - Tokenization: Chia nhỏ chuỗi văn bản thành các đơn vị nhỏ hơn "token", đó có thể là từ, ký tự hay dấu câu.
    - Lowercasing: Đưa tất cả chữ về dạng thường.
    - Stopword removal: Loại bỏ từ ít giá trị (the, is, are, and).
    - Stemming: Cắt về gốc từ (studying -> study).
    - Lemmatization: Chuẩn hóa từ về dạng chuẩn trong từ điển (better -> good).
    - Punctuation removal: Bỏ dấu câu.
    - Number removal: Bỏ số nếu không cần.

d) Mô tả ngắn gọn những kỹ thuật text preprocessing sử dụng trong các ứng dụng LLM, hoặc để xây dựng mô hình RAG. Cho ví dụ.

- Lọc chất lượng (Quality Filtering): Sử dụng các bộ phân loại hoặc các quy tắc heuristic để tự động loại bỏ dữ liệu chất lượng thấp.
- Loại bỏ trùng lặp (Deduplication): Xóa các bản sao ở nhiều cấp độ (câu, tài liệu, tập dữ liệu) để tăng sự đa dạng và ổn định cho mô hình.

- Loại bỏ thông tin cá nhân (Privacy Redaction): Phát hiện và xóa các thông tin nhạy cảm như tên, địa chỉ, số điện thoại để bảo vệ quyền riêng tư người dùng.
- Tokenization nâng cao: Các LLM hiện đại không dùng tokenization dựa trên từ đơn giản mà sử dụng các thuật toán subword như Byte Pair Encoding (BPE).

e) Làm sao xây dựng một pipeline bao gồm nhiều bước xử lý text

- **Phương pháp truyền thống:**

1. **Làm sạch văn bản (Text Cleaning):**
    - Chuyển đổi về chữ thường (Lowercasing): Đảm bảo tính nhất quán (Apple -> apple).
    - Loại bỏ các yếu tố nhiễu: Xóa các thẻ HTML, URL, ký tự đặc biệt, dấu câu, hoặc các con số không cần thiết.
2. **Tách từ (Tokenization):**
    - Phân tách văn bản thành các đơn vị nhỏ hơn(“token”) hoặc câu (sentence tokenization).
3. **Loại bỏ từ dừng (Stopword Removal):**
    - Xóa các từ xuất hiện phổ biến nhưng ít mang ý nghĩa ngữ nghĩa trong hầu hết các ngữ cảnh (ví dụ: "và", "là", "của", "thì", "một"). Việc này giúp giảm chiều dữ liệu và tập trung vào các từ khóa quan trọng.

4. **Chuẩn hóa từ (Stemming & Lemmatization):**
    - Stemming: Một phương pháp dựa trên quy tắc để đưa từ về dạng gốc bằng cách cắt bỏ phần hậu tố (running, ran, runs đều trở thành run). Phương pháp này nhanh nhưng đôi khi tạo ra các từ không có trong từ điển.
    - Lemmatization: Sử dụng từ điển và phân tích ngữ pháp để đưa từ về dạng nguyên thể (lemma) của nó (better -> good, was -> be). Phương pháp này chính xác hơn nhưng yêu cầu tài nguyên tính toán cao hơn.
5. **Vector hóa (Vectorization):** Biến đổi các từ đã được làm sạch thành các vector số để mô hình máy học có thể xử lý.
    - Các phương pháp phổ biến:
        - Bag-of-Words (BoW): Biểu diễn văn bản bằng một túi chứa các từ và tần suất xuất hiện của chúng, bỏ qua thứ tự từ.
        - TF-IDF (Term Frequency-Inverse Document Frequency): Đánh giá tầm quan trọng của một từ trong một văn bản cụ thể so với toàn bộ kho văn bản, giúp làm nổi bật các từ đặc trưng.
        - Word Embeddings (pre-trained): Sử dụng các vector từ đã được huấn luyện sẵn trên kho dữ liệu lớn (như Word2Vec, GloVe, FastText) để biểu diễn mỗi từ bằng một vector dày đặc, nắm bắt được mối quan hệ ngữ nghĩa giữa các từ.

- **Phương pháp LLM hiện đại:**

1. Thu thập dữ liệu (Data Collection)
2. Lọc Chất lượng (Quality Filtering): Tự động loại bỏ nội dung chất lượng thấp, spam, văn bản lặp lại vô nghĩa, hoặc văn bản được tạo tự động.
3. Khử Trùng lặp (Deduplication): Tăng sự đa dạng của dữ liệu, ngăn mô hình học thuộc lòng các mẫu lặp lại và cải thiện sự ổn định trong quá trình huấn luyện.
4. Ẩn/Xóa Thông tin Cá nhân: Bảo vệ quyền riêng tư người dùng
5. Tokenization Nâng cao (Advanced Tokenization): Phá vỡ văn bản thành các tokens mà mô hình có thể hiểu được, đồng thời xử lý hiệu quả các từ hiếm và từ ngoài bộ từ vựng (out-of-vocabulary).

<u>Câu hỏi:</u>

1. Để xây dựng một bộ lọc chất lượng, ngoài việc dùng Wikipedia làm “mẫu dương tính”, trong thực tế người ta còn dùng những phương pháp hoặc nguồn dữ liệu nào khác để định nghĩa 'chất lượng cao' và 'chất lượng thấp' ạ?"
2. Em thấy có nhiều thuật toán subword như BPE, WordPiece, SentencePiece. Sự khác biệt về hiệu năng thực tế của chúng khi áp dụng cho các ngôn ngữ khác nhau (đặc biệt là tiếng Việt) là gì ạ?