Chào bạn, "Làm sạch" và "Phân đoạn" là hai bước tiền xử lý cực kỳ quan trọng và phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Chúng giúp chuẩn bị dữ liệu văn bản thô để máy tính có thể "hiểu" và xử lý hiệu quả hơn.

**1. Làm sạch văn bản (Text Cleaning)**

**Định nghĩa:** **Làm sạch văn bản** là quá trình loại bỏ hoặc sửa đổi những phần "nhiễu" (noise) không mong muốn khỏi văn bản. Mục tiêu là chuẩn hóa văn bản, giảm sự phức tạp và chỉ giữ lại những thông tin cốt lõi, giúp cải thiện độ chính xác của các mô hình máy học.

**Các kỹ thuật làm sạch phổ biến và ví dụ:**

Giả sử chúng ta có câu văn bản thô sau:

" Chào Mừng bạn đến với Việt Nam!!! Đây là một quốc gia TUYỆT vời, phải không? #VN123 "

* **Chuyển thành chữ thường (Lowercasing):** Đưa tất cả các ký tự về dạng chữ thường để máy tính không coi "Việt Nam" và "việt nam" là hai từ khác nhau.
  + **Kết quả:** " chào mừng bạn đến với việt nam!!! đây là một quốc gia tuyệt vời, phải không? #vn123 "
* **Xóa bỏ dấu câu và ký tự đặc biệt (Removing Punctuation & Special Characters):** Loại bỏ các dấu như !, ?, # thường không mang nhiều ý nghĩa ngữ nghĩa trong nhiều bài toán.
  + **Kết quả:** " chào mừng bạn đến với việt nam đây là một quốc gia tuyệt vời phải không vn123 "
* **Xóa bỏ số (Removing Numbers):** Loại bỏ số nếu chúng không quan trọng cho tác vụ đang thực hiện.
  + **Kết quả:** " chào mừng bạn đến với việt nam đây là một quốc gia tuyệt vời phải không vn "
* **Xóa bỏ khoảng trắng thừa (Stripping Extra Whitespace):** Loại bỏ các khoảng trắng thừa ở đầu, cuối và giữa các từ.
  + **Kết quả:** "chào mừng bạn đến với việt nam đây là một quốc gia tuyệt vời phải không vn"
* **Xóa bỏ từ dừng (Removing Stop Words):** "Từ dừng" là những từ xuất hiện rất phổ biến nhưng thường không mang nhiều ý nghĩa (ví dụ: là, và, của, một, cái, thì, ở...). Việc loại bỏ chúng giúp mô hình tập trung vào các từ khóa quan trọng hơn.
  + *Danh sách từ dừng tiếng Việt (ví dụ):* là, một, phải không
  + **Kết quả cuối cùng sau khi làm sạch:** "chào mừng bạn đến với việt nam đây quốc gia tuyệt vời vn"

**2. Phân đoạn văn bản (Text Chunking/Segmentation)**

**Định nghĩa:** **Phân đoạn văn bản** là quá trình chia một văn bản dài thành các đoạn (chunks) nhỏ hơn, có kích thước cố định hoặc có ý nghĩa ngữ nghĩa. Việc này rất quan trọng khi làm việc với các mô hình ngôn ngữ lớn (LLMs) vì chúng có giới hạn về số lượng từ/token có thể xử lý trong một lần (gọi là "context window").

**Tại sao phải phân đoạn?**

* **Vượt qua giới hạn của mô hình:** Hầu hết các mô hình không thể "đọc" hết một cuốn sách hay một tài liệu dài vài trăm trang trong một lần.
* **Tăng tính liên quan:** Khi tìm kiếm thông tin, việc trả về một đoạn văn bản ngắn, tập trung vào đúng câu hỏi sẽ hữu ích hơn là trả về cả một chương sách dài.
* **Hiệu quả chi phí và tốc độ:** Xử lý các đoạn nhỏ sẽ nhanh và tốn ít tài nguyên tính toán hơn.

**Các kỹ thuật phân đoạn phổ biến và ví dụ:**

Giả sử chúng ta có một đoạn văn bản dài:

"Trí tuệ nhân tạo (AI) đang thay đổi thế giới. Nó được ứng dụng trong y tế để chẩn đoán bệnh sớm. Trong ngành công nghiệp ô tô, AI giúp phát triển xe tự lái. Giáo dục cũng không ngoại lệ, với các hệ thống học tập cá nhân hóa. Tuy nhiên, việc phát triển AI cũng đặt ra nhiều thách thức về đạo đức và an ninh."

* **Phân đoạn theo kích thước cố định (Fixed-size Chunking):** Chia văn bản thành các đoạn có số lượng ký tự hoặc từ bằng nhau. Đây là cách đơn giản nhất nhưng có thể cắt ngang giữa câu hoặc từ, làm mất ngữ nghĩa.
  + **Ví dụ (chia mỗi 100 ký tự):**
    - **Chunk 1:** "Trí tuệ nhân tạo (AI) đang thay đổi thế giới. Nó được ứng dụng trong y tế để chẩn đoán bệnh sớm. Tro"
    - **Chunk 2:** "ng ngành công nghiệp ô tô, AI giúp phát triển xe tự lái. Giáo dục cũng không ngoại lệ, với các hệ thố"
    - ... (và cứ thế tiếp tục)
* **Phân đoạn theo câu (Sentence-based Chunking):** Chia văn bản dựa trên các dấu kết thúc câu (., ?, !). Cách này tốt hơn vì nó giữ được trọn vẹn ý nghĩa của từng câu.
  + **Ví dụ:**
    - **Chunk 1:** "Trí tuệ nhân tạo (AI) đang thay đổi thế giới."
    - **Chunk 2:** "Nó được ứng dụng trong y tế để chẩn đoán bệnh sớm."
    - **Chunk 3:** "Trong ngành công nghiệp ô tô, AI giúp phát triển xe tự lái."
    - ... (và cứ thế tiếp tục)
* **Phân đoạn đệ quy (Recursive Chunking):** Đây là phương pháp thông minh và được sử dụng rộng rãi. Nó cố gắng chia văn bản theo các dấu phân cách có ý nghĩa nhất (ví dụ: xuống dòng 2 lần \n\n), nếu các đoạn vẫn quá dài, nó sẽ tiếp tục chia theo dấu phân cách nhỏ hơn (ví dụ: xuống dòng 1 lần \n), rồi đến câu, rồi đến từ... Cách này giúp giữ lại cấu trúc ngữ nghĩa tốt nhất có thể.
* **Phân đoạn có chồng lấn (Chunk Overlap):** Để tránh mất ngữ cảnh ở điểm nối giữa hai đoạn, người ta thường cho đoạn sau bắt đầu bằng một phần cuối của đoạn trước.
  + **Ví dụ (chồng lấn 1 câu):**
    - **Chunk 1:** "Câu A. Câu B. Câu C."
    - **Chunk 2:** "Câu C. Câu D. Câu E." (Câu C được lặp lại để giữ mạch văn)

Tóm lại, **Làm sạch** là "dọn dẹp" văn bản, còn **Phân đoạn** là "cắt nhỏ" văn bản một cách hợp lý. Cả hai đều là những bước nền tảng để xây dựng các ứng dụng NLP mạnh mẽ và chính xác.