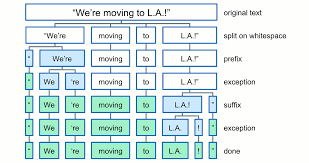
## **Tại sao chúng ta cần Tokenizer?**



Nhu cầu về tokenizer xuất phát từ câu hỏi "Làm thế nào chúng ta có thể làm cho máy đọc được?"

Con người học một ngôn ngữ bằng cách kết nối âm thanh với ý nghĩa và sau đó chúng ta học đọc và viết bằng ngôn ngữ đó. Máy móc không thể làm được điều đó nên chúng cần được cung cấp những đơn vị văn bản cơ bản nhất để bắt đầu xử lý văn bản.Đó là lúc tokenization phát huy tác dụng. Nó chia văn bản thành các đơn vị nhỏ hơn gọi là "token”

Để mô hình có thể học được gồm 2 bước

1. tokenize – quyết định thuật toán chúng tôi sẽ sử dụng để tạo mã thông báo.
2. Mã hóa token thành vectơ

**Word-Based tokenization:**

Chúng ta cần quyết định cách chuyển đổi văn bản thành các token nhỏ. Một phương pháp đơn giản và dễ hiểu mà hầu hết chúng ta đều đề xuất là sử dụng token dựa trên từ, chia văn bản bằng dấu cách.

Vấn đề đặt ra là nó có nguy cơ bị thiếu từ để huấn luyện. Mô hình không thể nhận ra các biến thể của từ không thuộc dữ liệu mà mô hình đã được đào tạo.Ngay cả khi bạn thêm một **công cụ bổ trợ** để giải quyết vấn đề này, bạn cũng đang thêm một bước bổ sung vào quy trình xử lý của mình.

**Việc xử lý tiếng lóng và chữ viết tắt cũng khó khăn.** Ngày nay, chúng ta sử dụng rất nhiều tiếng lóng và chữ viết tắt trong văn bản, chẳng hạn như "FOMO", "LOL", "tl;dr", v.v.

BPE ban đầu là một thuật toán nén dữ liệu mà bạn sử dụng để tìm ra cách tốt nhất để biểu diễn dữ liệu bằng cách xác định các cặp byte chung. Bây giờ chúng tôi sử dụng nó trong NLP để tìm cách trình bày văn bản tốt nhất bằng cách sử dụng số lượng mã thông báo nhỏ nhất.

Đây là cách nó hoạt động:

1. Thêm ký hiệu nhận dạng ( </w>) vào cuối mỗi từ để xác định phần cuối của từ rồi tính tần suất xuất hiện của từ trong văn bản.
2. Chia từ thành các ký tự rồi tính tần số ký tự.
3. Từ các mã thông báo ký tự, đối với số lần lặp được xác định trước, hãy đếm tần số của các cặp byte liên tiếp và hợp nhất cặp byte xuất hiện thường xuyên nhất.
4. Tiếp tục lặp lại cho đến khi bạn đạt đến giới hạn lặp lại (do bạn đặt) hoặc cho đến khi bạn đạt đến giới hạn mã thông báo.

## **Bước 1: Thêm định danh từ và tính tần suất từ**

## **Bước 2: Tách từ thành các ký tự rồi tính tần số ký tự**

## **Bước 3: Hợp nhất các cặp byte liên tiếp xuất hiện thường xuyên nhất**

## **Bước 4 - Lặp lại n lần để tìm các cặp (về tần số) tốt nhất để mã hóa và sau đó ghép chúng để tìm các từ con**

Vì vậy, khi lặp lại với từng cặp tốt nhất, chúng tôi hợp nhất (nối) cặp đó. Bạn có thể thấy rằng khi chúng tôi tính toán lại tần số, tần số mã thông báo ký tự ban đầu sẽ giảm xuống và tần số mã thông báo được ghép nối mới sẽ xuất hiện trong từ điển mã thông báo.

Nếu bạn nhìn vào số lượng mã thông báo được tạo, trước tiên nó sẽ tăng lên do chúng tôi tạo các cặp mới – nhưng số lượng bắt đầu giảm sau một số lần lặp.

