Tokenization – "Băm nhỏ" ngôn ngữ để AI hiểu, qua thế giới phù thủy Harry Potter

Một khái niệm cực kỳ quan trọng trong lĩnh vực Trí tuệ nhân tạo và Xử lý ngôn ngữ tự nhiên (NLP): **Tokenization** – hay còn gọi là quá trình tách văn bản thành các đơn vị nhỏ hơn gọi là "tokens".

Và để cho việc học trở nên thú vị hơn, hãy cùng quay về với thế giới phù thủy **Harry Potter** nhé!

Tokenization là gì?

Hãy tưởng tượng bạn đang luyện đọc câu thần chú:
"Wingardium Leviosa", và Hermione nhắc bạn: *"Không phải Leviosá, mà là Leviosa!"*

Để đọc đúng câu thần chú, bạn phải phát âm đúng từng âm tiết.
=> Tương tự, AI cũng cần "nghe hiểu" ngôn ngữ bằng cách **chia nhỏ câu chữ thành những phần dễ xử lý hơn** – đó chính là **tokenization**.

Token có thể là:

- Một từ: *"Harry", "phù thủy", "đũa phép"*
- Một phần của từ: *"phù-" và "-thủy"*
- Hoặc một ký tự: *"H", "a", "r", "r", "y"*

Ví dụ cụ thể:

Giả sử bạn có câu sau:

> **"Harry bắt được trái banh vàng trong trận Quidditch."**

- Nếu dùng **token hóa theo từ (word tokenization)**:
  - Token sẽ là: `Harry`, `bắt`, `được`, `trái`, `banh`, `vàng`, `trong`, `trận`, `Quidditch`.

- Nhưng nếu gặp từ lạ như **“Quidditch”**, mô hình AI có thể không hiểu ngay.
  => Lúc này, ta dùng **token hóa theo phần từ (subword tokenization)**:
  Ví dụ: `Quid`, `ditch` → giúp AI nhận diện từ dễ hơn bằng cách chia nhỏ.
- Với ngôn ngữ phức tạp hoặc tên riêng, có thể dùng **token hóa theo ký tự (character tokenization)**:
  - Ví dụ: *“Alohomora”* → tách thành: `A`, `l`, `o`, `h`, `o`, `m`, `o`, `r`, `a`

Tại sao tokenization lại quan trọng?

Nó giúp AI **hiểu chính xác ngữ cảnh và ý nghĩa** của câu.
Ví dụ:

> **“Giáo sư Snape đưa thuốc của Harry cho Neville.”**

Nếu token hóa và xử lý không chính xác, AI có thể hiểu nhầm rằng *thuốc là của Neville* chứ không phải *của Harry*.

Việc tách câu đúng, kết hợp với bối cảnh, giúp mô hình hiểu rằng:

- *Người sở hữu ban đầu là Harry.*
- *Người nhận là Neville.*
- *Người thực hiện hành động là Snape.*

Thách thức trong tokenization

- Từ vựng hiếm, đặc biệt như:
  “Expecto Patronum”, “Horcrux”, “Voldemort” – nếu không được token hóa đúng, AI sẽ không hiểu được ý nghĩa.
- Tên riêng hoặc từ ghép:
  “Trường Hogwarts”, “Ký ức Pensieve”, “Bảo bối tử thần” – cần giữ nguyên như một token duy nhất để không làm sai lệch ý nghĩa.
- Ngôn ngữ không dùng bảng chữ cái Latinh (như tiếng Trung, tiếng Nhật) – nơi mỗi ký tự có thể là một từ hoặc một âm tiết – sẽ cần xử lý đặc biệt hơn.

Tokenization trong mô hình của OpenAI

Mỗi mô hình ngôn ngữ có cách token hóa riêng, và tokenizer của OpenAI đã được tối ưu cho các mô hình GPT như GPT-3 hay GPT-4.
=> Ví dụ: thay vì coi *“asphodel”* là một token duy nhất, nó có thể tách thành: `asp`, `ho`, `del` nếu cần thiết.

Tóm lại

- Tokenization là **bước đầu tiên và quan trọng nhất** để AI hiểu ngôn ngữ.
- Có 3 phương pháp chính:
  ➤ **Token hóa từ** – đơn giản, dễ hiểu.
  ➤ **Token hóa phần từ** – phù hợp cho từ vựng hiếm.
  ➤ **Token hóa ký tự** – dành cho ngôn ngữ đặc biệt.
- Tokenization **không chỉ là tách từ**, mà là cách để **tối ưu hóa khả năng hiểu ngôn ngữ của mô hình AI.**

Tokenization nhìn thì đơn giản, nhưng lại là chiếc chìa khóa đầu tiên mở ra cánh cửa hiểu ngôn ngữ của máy móc.
Và trong phần tiếp theo, chúng ta sẽ cùng nhau thử nghiệm với tokenizer thực tế của OpenAI để thấy rõ mọi thứ vận hành ra sao nhé!

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #1
Được tạo 23 tháng 4 2025 17:48:38 bởi Đỗ Ngọc Tú
Được cập nhật 26 tháng 4 2025 02:30:05 bởi Đỗ Ngọc Tú