

# Tìm hiểu về Tokenization với OpenAI Tokenizer

Trong bài viết này, chúng ta sẽ cùng tìm hiểu cách **tokenization** hoạt động thông qua công cụ **OpenAI Tokenizer** – một công cụ miễn phí và dễ sử dụng mà bạn có thể tìm thấy bằng cách tìm kiếm từ khóa “OpenAI tokenizer” trên Google.

## Tokenization là gì?

Tokenization là quá trình mà các mô hình ngôn ngữ như ChatGPT sử dụng để **chia nhỏ văn bản thành các đơn vị nhỏ hơn gọi là tokens**. Mỗi token có thể là một từ, một phần của từ, dấu câu, hoặc thậm chí là khoảng trắng.

## Trải nghiệm thực tế với OpenAI Tokenizer

Bạn không cần tài khoản để dùng thử công cụ này. Hãy cùng thử một ví dụ đơn giản:

“Harry caught the golden snitch during the Quidditch match.”

Khi nhập vào, bạn sẽ thấy câu này được chia thành các token như:

- “Harry” là một token.
- “caught the golden snitch” là một cụm token.
- “during” là một token riêng.
- “Quidditch” được chia thành nhiều phần nhỏ như “Qu”, “id”, “ditch”.
- “match” và dấu “.” cũng là các token riêng biệt.

Mỗi token tương ứng với một **token ID** – một con số đại diện cho token đó. Ví dụ: “.” luôn có giá trị là **13**.

## Token có phân biệt chữ hoa, chữ thường và số nhiều không?

Có! Hãy thử câu:

|

"I ate an apple. Then I bought two apples with my Apple iPhone."

- "apple" (số ít, chữ thường) và "apples" (số nhiều) có token khác nhau.
- "Apple" (viết hoa, chỉ thương hiệu) cũng có token khác.
- "iPhone" được tách thành **hai token** vì nó là một từ phức.

Điều này cho thấy tokenizer **phân biệt rõ giữa các dạng viết khác nhau**, bao gồm cả chữ hoa, chữ thường, số ít và số nhiều.

## Token ID có ý nghĩa gì không?

Không. Các con số như 3366, 1810, hay 1641 chỉ là giá trị đại diện – **chúng không có ý nghĩa ngữ nghĩa nào**. Dù một số token ID có vẻ “gần nhau”, chúng không hề liên quan về mặt nghĩa.

## Các mô hình khác nhau - tokenizer có khác nhau không?

Có. Tùy vào mô hình GPT (ví dụ: GPT-3 hay GPT-4-turbo), **tokenizer có thể phân tách văn bản hơi khác nhau**. Nhưng bạn không cần lo lắng quá – vì:

- Mỗi mô hình sẽ tự sử dụng tokenizer phù hợp của nó.
- Việc phân tách token là bước đầu tiên mà mô hình xử lý trước khi hiểu và phản hồi bạn.

## Kết luận

Tokenization là một bước quan trọng trong cách các mô hình AI hiểu ngôn ngữ. Tuy nhiên, bạn **không cần phải thao tác thủ công**, vì các mô hình hiện đại sẽ làm điều này cho bạn. Dù vậy, hiểu rõ quá trình này giúp bạn sử dụng mô hình hiệu quả hơn, đặc biệt khi làm việc với giới hạn token trong API hoặc cần tối ưu hóa prompt.

Tác giả: **Đỗ Ngọc Tú**  
Công Ty Phần Mềm **VHTSoft**

Phiên bản #1

Được tạo 24 tháng 4 2025 13:50:09 bởi Đỗ Ngọc Tú

Được cập nhật 26 tháng 4 2025 02:30:05 bởi Đỗ Ngọc Tú