Token Counting with tiktoken

Debajyoti Maity

November 4, 2024

1 Introduction

This document describes the functionality of counting tokens in text using the `tiktoken` library. Tokenization is a critical step in natural language processing and machine learning, where text is converted into a format suitable for model consumption.

2 Tokenization

Tokenization is the process of splitting text into smaller units, called tokens. These tokens can be words, subwords, or characters, depending on the specific tokenizer used. In this implementation, we utilize the `cl100k_base` tokenizer from the `tiktoken` library, which is designed for use with OpenAI's models.

3 Functionality

The main functionality is encapsulated in the `count_tokens` function, which takes a string of text as input and returns the number of tokens it contains.

3.1 Function Definition

The function is defined as follows:

```
def count_tokens(text):
    """Count the number of tokens in a given text using tiktoken."""
    if not text:  # Handle empty string case
        return 0
    tokens = tokenizer.encode(text)  # Encode the text into tokens
    return len(tokens)  # Return the number of tokens
```

3.2 Examples

Below are some examples of using the `count_tokens` function to count tokens in various types of text:

- **Simple greeting:** `"Hello, world!"`
  Token count: `count_tokens("Hello, world!") = 4`

- **A question:** `"What is the capital of France?"`
  Token count: `count_tokens("What is the capital of France?") = 8`

- **Longer sentence:** `"This is a longer sentence to test the token counting functionality."`
  Token count: `count_tokens("This is a longer sentence to test the token counting functionality.") = 13`

- **Empty string:** `""`
  Token count: `count_tokens("") = 0`

- **Text with emojis:** `" Let’s test some emojis!  "`
  Token count: `count_tokens(" Let’s test some emojis!  ") = 7`

4 Conclusion

Counting tokens is essential for processing and understanding text in machine learning applications. This implementation provides a straightforward method to evaluate the number of tokens in any given text using the `tiktoken` library.