Tokenizer

Learn about language model tokenization

OpenAI's large language models (sometimes referred to as GPT's) process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

It's important to note that the exact tokenization process varies between models. Newer models like GPT-3.5 and GPT-4 use a different tokenizer than previous models, and will produce different tokens for the same input text.

GPT-3.5 & GPT-4 **GPT-3 (Legacy)**

```
response = openai.ChatCompletion.create(
    model='gpt-3.5-turbo',
    messages=conversation,
    max_tokens= 200
)

message=response.choices[0]['message']
```

Clear **Show example**

Tokens
350

Characters
1354

```
context =  "Eres un analista transaccional, determinas la clas
ificacion de cada cliente segun las reglas definidas presentadolo
 de manera resumida"
prompt = " Los canales en los que puede interactuar cada cliente o
 customer son Comercial, Backoffice y Caja, mientras mayor partip
 perteneciente al canal en el que tenga una participación porcentual
 superior al 50%. Si la participación en dos o más canales es
 superior al 30%, se clasifica al cliente como Multi-canal, el
 Cliente1 tiene una proporcion transaccional Backoffice 42%  Caja 33
%  Comercial 25% , el Cliente2 tiene una proporcion transaccional
 Backoffice 40% Caja 40% Comercial 20%, el Cliente3 tiene una
 proporcion transaccional Backoffice 20% Caja 60% Comercial 20%, el
 Cliente4 tiene una proporcion transaccional Backoffice 20% Caja 10%
 Comercial 70%, el Cliente5 tiene una proporcion transaccional Back
office 27%  Caja 27%  Comercial 45% "

                [
    {"role":"system","content":context},
```

Text Token IDs

A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).

If you need a programmatic interface for tokenizing text, check out our tiktoken package for Python. For JavaScript, the community-supported @dbdq/tiktoken package works with mo GPT models.