Overview Documentation API reference Log in Sign up

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)

```
([ role : user , content :process_text_to_imagen])

response = openai.ChatCompletion.create(
    model='gpt-3.5-turbo',
    messages=conversation,
    max_tokens=300)

img_prompt = response.choices[0]['message']['content'].strip()
print ("Los prompts para cada cliente son los siguientes:",
```

Clear Show example

Tokens
219

Characters
878

```
process_text_to_imagen = "en base a la respuesta anterior crea un
 text prompt para cada uno de los clientes, para que pueda ser usado
 como prompt en NightCafe y pueda generar una imagen de cada uno de
 ellos, asignando colores en funcion de su participacion transacc
ional en cada canal, ej si el cliente opero 50% en canal comercial
```

```
  presentalo de manera resumida categoria a la que pertenece el
  cliente y colores asignados "

conversation.append ({"role":"user","content":process_text_to_imagen
})

response = openai.ChatCompletion.create(
    model='gpt-3.5-turbo',
    messages=conversation,
    max_tokens=300)

                      sponse.choices[0]['message']['content'].strip()
print ("Los prompts para cada cliente son los siguientes:", img
```

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.