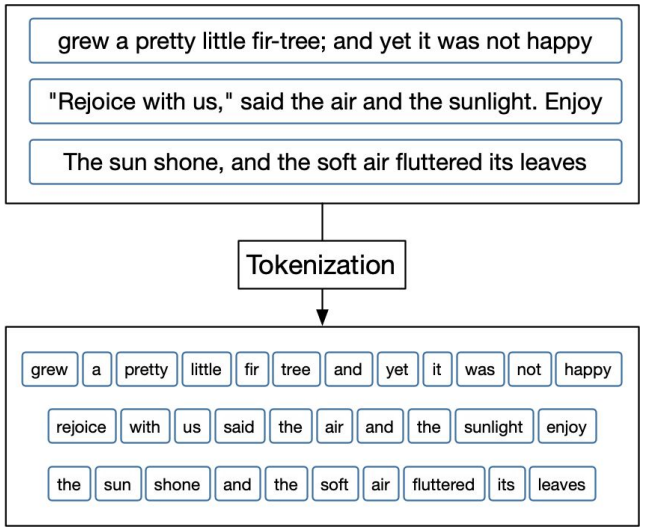
What are Tokens?

Examine token limits what are and how to gracefully handle limits.

What Is Tokenization?

Sentences, words and characters are transformed into lots of ***individual tokens.***



Token Limits

- ***Restricted by the models*** that we choose to use.
- Each model has an input token limit and and output token limit.

Exercise: View OpenAI Pricing

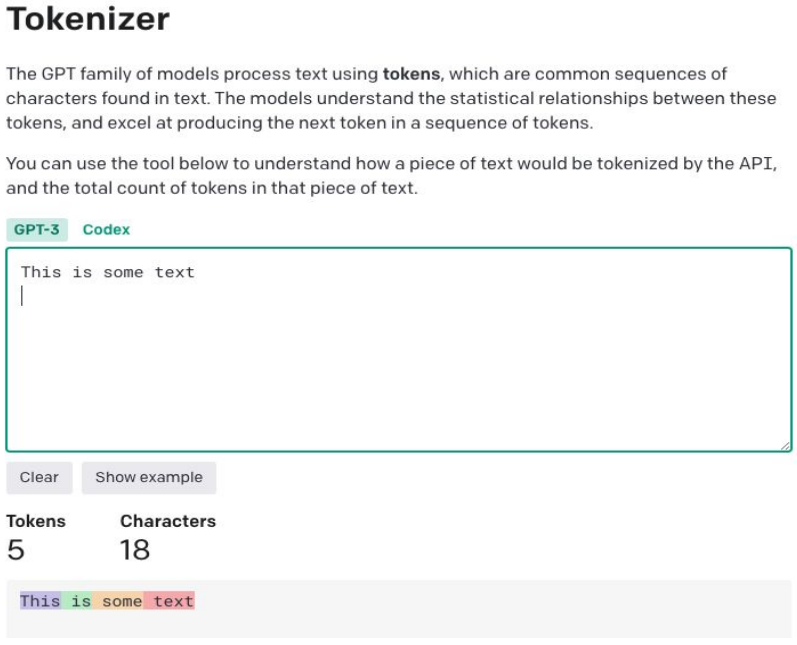
Navigate to OpenAI's pricing by Google searching ***"OpenAI pricing"***

Explore the different models and their pricing for:

- Input tokens
- Cached input tokens
- Output tokens

How to get the token limit for ChatGPT and GPT-X

https://platform.openai.com/tokenizer



How to get the token limit - Automatically

```python
import tiktoken
enc = tiktoken.get_encoding("o200k_base")
assert enc.decode(enc.encode("hello world")) == "hello world"

# To get the tokeniser corresponding to a specific model in the OpenAI API:
enc = tiktoken.encoding_for_model("gpt-4o")
```

Approaches To Avoid Hitting Token Limits

- Use different models
- Shortening your prompt/context
- Chunking
- Windowed chunks
- Summarisation