tiktoken (OpenAI) Length Function

You can also use tiktoken, a open source tokenizer package from OpenAI to estimate tokens used. Will probably be more accurate for their models.

1. How the text is split: by character passed in
2. How the chunk size is measured: by `tiktoken` tokenizer

```
# This is a long document we can split up.
with open('../../../state_of_the_union.txt') as f:
    state_of_the_union = f.read()
from langchain.text_splitter import CharacterTextSplitter
```

```
text_splitter = CharacterTextSplitter.from_tiktoken_encoder(chunk_size=100,
chunk_overlap=0)
texts = text_splitter.split_text(state_of_the_union)
```

```
print(texts[0])
```

```
Madam Speaker, Madam Vice President, our First Lady and Second Gentleman. Members
of Congress and the Cabinet. Justices of the Supreme Court. My fellow Americans.

Last year COVID-19 kept us apart. This year we are finally together again.

Tonight, we meet as Democrats Republicans and Independents. But most importantly
as Americans.

With a duty to one another to the American people to the Constitution.
```