Hugging Face Length Function

Most LLMs are constrained by the number of tokens that you can pass in, which is not the same as the number of characters. In order to get a more accurate estimate, we can use Hugging Face tokenizers to count the text length.

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

```
from transformers import GPT2TokenizerFast

tokenizer = GPT2TokenizerFast.from_pretrained("gpt2")
```

```
# 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_huggingface_tokenizer(tokenizer,
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.
```