NLTK Text Splitter

Rather than just splitting on “\n\n”, we can use NLTK to split based on tokenizers.

1. How the text is split: by NLTK
2. How the chunk size is measured: by length function passed in (defaults to number of characters)

```
# 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 NLTKTextSplitter
text_splitter = NLTKTextSplitter(chunk_size=1000)
```

```
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.

And with an unwavering resolve that freedom will always triumph over tyranny.

Six days ago, Russia’s Vladimir Putin sought to shake the foundations of the free
world thinking he could make it bend to his menacing ways.
```

Skip to main content

```
But he badly miscalculated.

He thought he could roll into Ukraine and the world would roll over.

Instead he met a wall of strength he never imagined.

He met the Ukrainian people.

From President Zelenskyy to every Ukrainian, their fearlessness, their courage,
their determination, inspires the world.

Groups of citizens blocking tanks with their bodies.
```