notebook

April 27, 2025

0.1 # Introduction

Fly away with Peter Pan!

Peter Pan has been the companion of many children, and went a long way, starting as a Christmas play and ending up as a Disney classic. Did you know that although the play was titled "Peter Pan, Or The Boy Who Wouldn't Grow Up", J. M. Barrie's novel was actually titled "Peter and Wendy"?

You're going to explore and analyze Peter Pan's text to answer the question in the instruction pane below. You are working with the text version available here at Project Gutenberg. Feel free to add as many cells as necessary. Finally, remember that you are only tested on your answer, not on the methods you use to arrive at the answer!

Note: If you haven't completed a DataCamp project before you should check out the Intro to Projects first to learn about the interface. Intermediate Importing Data in Python and Introduction to Natural Language Processing in Python teach the skills required to complete this project. Should you decide to use them, English stopwords have been downloaded from nltk and are available for you in your environment.

0.1.1 Step 1: Import libraries

[1]:
```
import requests
import nltk
from bs4 import BeautifulSoup
from collections import Counter
```

0.1.2 Step 2: Get HTML

[2]:
```
r = requests.get("https://www.gutenberg.org/files/16/16-h/16-h.htm")
r.encoding = 'utf-8'
html = r.text
```

0.1.3 Step 3: Get Text

[3]:
```
soup = BeautifulSoup(html)
text = soup.text
```

0.1.4 Step 4: Get Words

[4]:
```
tokenizer = nltk.tokenize.RegexpTokenizer("\w+")
tokens = tokenizer.tokenize(text)
words = [token.lower() for token in tokens]
```

0.1.5 Step 5: Stopwords

[5]:
```
nltk.download('stopwords')

# Make a list of stop words
stop_words = nltk.corpus.stopwords.words("english")

# Remove stopwords from tokens list
words_clean = [word for word in words if word not in stop_words]
```

```
[nltk_data] Downloading package stopwords to /home/repl/nltk_data…
```

```
---------------------------------------------------------------------------
PermissionError                           Traceback (most recent call last)
<ipython-input-5-db118785861d> in <module>
----> 1 nltk.download('stopwords')
      2
      3 # Make a list of stop words
      4 stop_words = nltk.corpus.stopwords.words("english")
      5

/usr/local/lib/python3.6/dist-packages/nltk/downloader.py in download(self,␣
 ↪info_or_id, download_dir, quiet, force, prefix, halt_on_error, raise_on_error
    668                                       subsequent_indent=prefix+prefix2+'␣
 ↪'*4))
    669
--> 670             for msg in self.incr_download(info_or_id, download_dir,␣
 ↪force):
    671                 # Error messages
    672                 if isinstance(msg, ErrorMessage):

/usr/local/lib/python3.6/dist-packages/nltk/downloader.py in incr_download(self ␣
 ↪info_or_id, download_dir, force)
    553         # Handle Packages (delegate to a helper function).
    554         else:
--> 555             for msg in self._download_package(info, download_dir, force ):
    556                 yield msg
    557

/usr/local/lib/python3.6/dist-packages/nltk/downloader.py in␣
 ↪_download_package(self, info, download_dir, force)
    604             if status == self.STALE:
```

```
    605                 yield StaleMessage(info)
--> 606             os.remove(filepath)
    607
    608         # Ensure the download_dir exists

PermissionError: [Errno 13] Permission denied: '/home/repl/nltk_data/corpora/
 ↪stopwords.zip'
```

0.1.6 Step 6: Count Words

[ ]:
```
count = Counter(words_clean)

# Get top 10 most common words
top_ten = count.most_common(10)
```

0.1.7 Step 7: Protaganists

[ ]:
```
protagonists = ["hook", "john", "peter", "wendy"]
```