Text Assignment

February 5, 2025

Question: Which of the three books (A Doll's House, Wuthering Heights & Jane Eyre: An Autobiography) consists of the highest propotion of positive words?

[29]:
```
# import packages (you may need this)
import numpy as np
import pandas as pd
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import requests
import re
from urllib.parse import urlparse
import urllib.robotparser
from bs4 import BeautifulSoup

# This code checks the robots.txt file
def canFetch(url):

    parsed_uri = urlparse(url)
    domain = '{uri.scheme}://{uri.netloc}/'.format(uri=parsed_uri)

    rp = urllib.robotparser.RobotFileParser()
    rp.set_url(domain + "/robots.txt")
    try:
        rp.read()
        canFetchBool = rp.can_fetch("*", url)
    except:
        canFetchBool = None

    return canFetchBool
```

The output True for each of the URLs signify that the link can be fetched by any crawler.

[43]:
```
print(canFetch("https://www.gutenberg.org/cache/epub/1260/pg1260.txt")) # Jane␣
 ↪Eyre: An Autobiography
print(canFetch("https://www.gutenberg.org/cache/epub/2542/pg2542.txt")) # A␣
 ↪Doll's House
print(canFetch("https://www.gutenberg.org/cache/epub/768/pg768.txt")) #␣
 ↪Wurthering Heights
```

```
True
True
True
```

Then, I will fetch the url of each book and display the first 2000 characters them.

[40]:
```
url = "https://www.gutenberg.org/cache/epub/1260/pg1260.txt"

r = requests.get(url)

urlText = r.text # Extract the text content from the response
urlText[:2000] # Print the first 2000 characters
```

[40]:
```
'\ufeffThe Project Gutenberg eBook of Jane Eyre: An Autobiography\r\n
\r\nThis ebook is for the use of anyone anywhere in the United States
and\r\nmost other parts of the world at no cost and with almost no
restrictions\r\nwhatsoever. You may copy it, give it away or re-use it under the
terms\r\nof the Project Gutenberg License included with this ebook or
online\r\nat www.gutenberg.org. If you are not located in the United
States,\r\nyou will have to check the laws of the country where you are
located\r\nbefore using this eBook.\r\n\r\nTitle: Jane Eyre: An
Autobiography\r\n\r\nAuthor: Charlotte Brontë\r\n\r\nIllustrator: F. H.
Townsend\r\n\r\nRelease date: March 1, 1998 [eBook #1260]\r\n
Most recently updated: October 29, 2024\r\n\r\nLanguage: English\r\n\r\nCredits:
David Price\r\n\r\n\r\n*** START OF THE PROJECT GUTENBERG EBOOK JANE EYRE: AN
AUTOBIOGRAPHY ***\r\n\r\nJANE EYRE\r\nAN AUTOBIOGRAPHY\r\n\r\nby Charlotte
Brontë\r\n\r\n_ILLUSTRATED BY F. H. TOWNSEND_\r\n\r\nLondon\r\nSERVICE &
PATON\r\n5 HENRIETTA STREET\r\n1897\r\n\r\n_The Illustrations_\r\n_in this
Volume are the copyright of_\r\nSERVICE & PATON, _London_\r\n\r\nTO\r\nW. M.
THACKERAY, ESQ.,\r\n\r\nThis Work\r\nIS RESPECTFULLY INSCRIBED\r\n\r\nBY\r\nTHE
AUTHOR\r\n\r\n\r\n\r\n\r\n\r\nPREFACE\r\n\r\n\r\nA preface to the first edition of
"Jane Eyre" being unnecessary, I gave\r\nnone: this second edition demands a few
words both of acknowledgment\r\nand miscellaneous remark.\r\n\r\nMy thanks are
due in three quarters.\r\n\r\nTo the Public, for the indulgent ear it has
inclined to a plain tale\r\nwith few pretensions.\r\n\r\nTo the Press, for the
fair field its honest suffrage has opened to an\r\nobscure aspirant.\r\n\r\nTo
my Publishers, for the aid their tact, their energy, their practical\r\nsense
and frank liberality have afforded an unknown and
unrecommended\r\nAuthor.\r\n\r\nThe Press and the Public are but vague
personifications for me, and I\r\nmust thank them in vague terms; but my
Publishers are definite: so are\r\ncertain generous critics who have encouraged
me as only large-hearted\r\nand high-minded men know how to encourage a
strugglin'
```

[41]:
```
url2 = "https://www.gutenberg.org/cache/epub/2542/pg2542.txt"

r = requests.get(url2)
```

```
urlText2 = r.text # Extract the text content from the response
urlText2[:2000] # Print the first 2000 characters
```

[41]: "\ufeffThe Project Gutenberg eBook of A Doll's House : a play\r\n   \r\nThis ebook is for the use of anyone anywhere in the United States and\r\nmost other parts of the world at no cost and with almost no restrictions\r\nwhatsoever. You may copy it, give it away or re-use it under the terms\r\nof the Project Gutenberg License included with this ebook or online\r\nat www.gutenberg.org. If you are not located in the United States,\r\nyou will have to check the laws of the country where you are located\r\nbefore using this eBook.\r\n\r\nTitle: A Doll's House : a play\r\n\r\nAuthor: Henrik Ibsen\r\n\r\nRelease date: March 1, 2001 [eBook #2542]\r\n   Most recently updated: March 27, 2024\r\n\r\nLanguage: English\r\n\r\nCredits: Martin Adamson and David Widger\r\n\r\n\r\n*** START OF THE PROJECT GUTENBERG EBOOK A DOLL'S HOUSE : A PLAY ***\r\n\r\n\r\n\r\n\r\n\r\nA Doll’s House\r\n\r\nby Henrik Ibsen\r\n\r\n\r\nContents\r\n\r\n ACT I.\r\n ACT II.\r\n ACT III.\r\n\r\n\r\n\r\n DRAMATIS PERSONAE\r\n\r\nTorvald Helmer.\r\nNora, his wife.\r\nDoctor Rank.\r\nMrs Linde.\r\nNils Krogstad.\r\nHelmer’s three young children.\r\nAnne, their nurse.\r\nA Housemaid.\r\nA Porter.\r\n\r\n_[The action takes place in Helmer’s house.]_\r\n\r\n\r\n\r\n\r\nA DOLL’S HOUSE\r\n\r\n\r\n\r\n\r\nACT I\r\n\r\n\r\n_[SCENE.-A room furnished comfortably and tastefully, but not\r\nextravagantly. At the back, a door to the right leads to the\r\nentrance-hall, another to the left leads to Helmer’s study. Between the\r\ndoors stands a piano. In the middle of the left-hand wall is a door,\r\nand beyond it a window. Near the window are a round table, arm-chairs\r\nand a small sofa. In the right-hand wall, at the farther end, another\r\ndoor; and on the same side, nearer the footlights, a stove, two easy\r\nchairs and a rocking-chair; between the stove and the door, a small\r\ntable. Engravings on the walls; a cabinet with china and other small\r\nobjects; a small book-case with well-bound books. The floors are\r\ncarpeted, and a fire burns in the stove. It is winter._\r\n\r\n_A bell rings in the hall; shortly afterwards the door is heard to\r\nopen. Enter NORA, "

[42]:
```
url3 = "https://www.gutenberg.org/cache/epub/768/pg768.txt"

r = requests.get(url3)

urlText3 = r.text # Extract the text content from the response
urlText3[:2000] # Print the first 2000 characters
```

[42]: '\ufeffThe Project Gutenberg eBook of Wuthering Heights\r\n   \r\nThis ebook is for the use of anyone anywhere in the United States and\r\nmost other parts of the world at no cost and with almost no restrictions\r\nwhatsoever. You may copy it, give it away or re-use it under the terms\r\nof the Project Gutenberg License included with this ebook or online\r\nat www.gutenberg.org. If you are not located in the United States,\r\nyou will have to check the laws of the country where you are located\r\nbefore using this eBook.\r\n\r\nTitle:

```
Wuthering Heights\r\n\r\nAuthor: Emily Brontë\r\n\r\nRelease date: December 1,
1996 [eBook #768]\r\n                Most recently updated: January 18,
2022\r\n\r\nLanguage: English\r\n\r\nCredits: David Price\r\n\r\n\r\n*** START
OF THE PROJECT GUTENBERG EBOOK WUTHERING HEIGHTS
***\r\n\r\n\r\n\r\n\r\n\r\nWuthering Heights\r\n\r\nby Emily
Brontë\r\n\r\n\r\n\r\n\r\n\r\nCHAPTER I\r\n\r\n\r\n1801-I have just returned from a
visit to my landlord-the solitary\r\nneighbour that I shall be troubled with.
This is certainly a beautiful\r\ncountry! In all England, I do not believe that
I could have fixed on a\r\nsituation so completely removed from the stir of
society. A perfect\r\nmisanthropist's Heaven-and Mr. Heathcliff and I are such a
suitable\r\npair to divide the desolation between us. A capital fellow! He
little\r\nimagined how my heart warmed towards him when I beheld his black
eyes\r\nwithdraw so suspiciously under their brows, as I rode up, and when
his\r\nfingers sheltered themselves, with a jealous resolution, still
further\r\nin his waistcoat, as I announced my name.\r\n\r\n"Mr. Heathcliff?" I
said.\r\n\r\nA nod was the answer.\r\n\r\n"Mr. Lockwood, your new tenant, sir. I
do myself the honour of calling\r\nas soon as possible after my arrival, to
express the hope that I have\r\nnot inconvenienced you by my perseverance in
soliciting the occupation\r\nof Thrushcross Grange: I heard yesterday you had
had some thoughts-"\r\n\r\n"Thrushcross Grange is my own, sir," he interrupted,
wincing. "I should\r\nnot allow any one to inconvenience me, if I could hinder
it-'
```

The code below appends each URL to the List in order to convert it into a DataFrame in the next code.

[38]:
```python
# Dictionary to store book titles and their corresponding URLs
books = {
    "Jane Eyre": "https://www.gutenberg.org/cache/epub/1260/pg1260.txt",
    "A Doll's House": "https://www.gutenberg.org/cache/epub/2542/pg2542.txt",
    "Wuthering Heights": "https://www.gutenberg.org/cache/epub/768/pg768.txt"
}

# Loop through each item in the dictionary
for title, url in books.items():
    print(f"{title}: {url}")
```

```
Jane Eyre: https://www.gutenberg.org/cache/epub/1260/pg1260.txt
A Doll's House: https://www.gutenberg.org/cache/epub/2542/pg2542.txt
Wuthering Heights: https://www.gutenberg.org/cache/epub/768/pg768.txt
```

[62]:
```python
# Function to get text from a given URL
def get_text_from_url(url):
    response = requests.get(url)
    return response.text

results = []
```

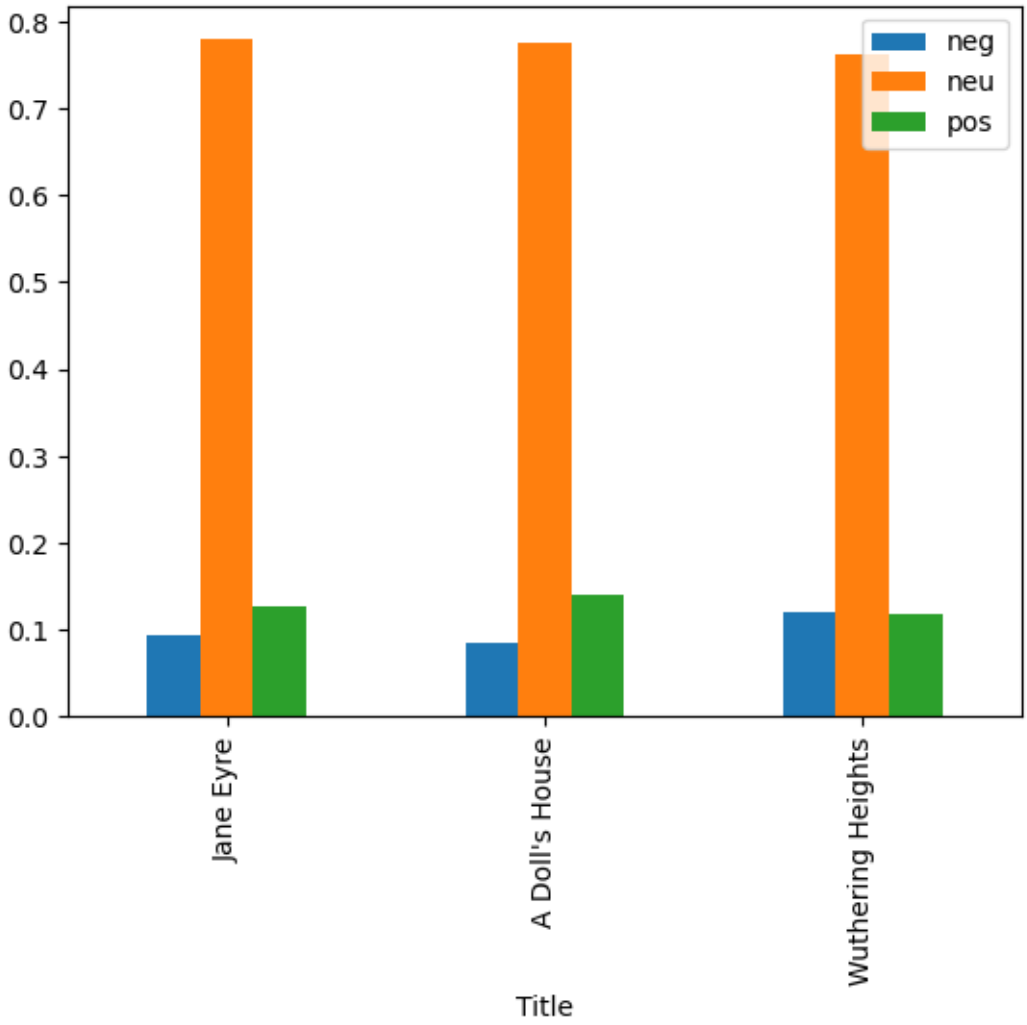
```
for title, url in books.items():
    text = get_text_from_url(url) # Retrieve text from the URL
    sentiment_scores = sid.polarity_scores(text) # Analyze sentiment of the text
    results.append({"Title": title, "URL": url, **sentiment_scores}) # Append a␣
 ↪dictionary with title, URL, and sentiment scores to results

# Create a DataFrame
df = pd.DataFrame(results).set_index('Title')

df
```

[62]:

Title	URL	neg	neu	pos	compound
Jane Eyre	https://www.gutenberg.org/cache/epub/1260/pg12…	0.094	0.779	0.127	1.0000
A Doll's House	https://www.gutenberg.org/cache/epub/2542/pg25…	0.084	0.775	0.141	1.0000
Wuthering Heights	https://www.gutenberg.org/cache/epub/768/pg768…	0.121	0.763	0.117	-0.9999

In addition to the DataFrame, I included a bar graph below in order to visualise the data better.

[63]:
```
df = df.drop(columns=['compound', 'URL']) # Drop the 'compound' and 'URL'␣
 ↪columns from the DataFrame
df.plot(kind = 'bar'); # Plot the remaining sentiment scores as a bar chart
```



The DataFrame displays the sentiment analysis for all three books in one table, allowing us to view the data and draw meaningful conclusions. From the sentiment analysis above, we can deduce that all three books have a similar proportion of positive words. However, we can see that 'A Doll's House' has the highest proportion of positive words while 'Wurthering Heights' has a the lowest proportion of positive words.

1. Highest proportion of positive words: A Doll's House - This play demonstrates the highest proportion of positive words, indicating a narrative that may convey themes of hope, empowerment, or resolution. The use of uplifting language could reflect the character's journey toward self-discovery and emancipation, particularly in the context of societal constraints.

2. Lowest proportion of positive words: Wuthering Heights - In contrast, this novel registers the lowest proportion of positive words. Its darker themes of obsession, revenge, and tumultuous relationships may contribute to this finding. The language used throughout the story likely reflects the emotional intensity and turmoil experienced by the characters, overshadowing more positive expressions.