Activity 9: Extracting the Top 100 eBooks from Gutenberg

Project Gutenberg encourages the creation and distribution of eBooks by encouraging volunteer efforts to digitize and archive cultural works. This activity aims to scrape the URL of Project Gutenberg's Top 100 eBooks to identify the eBooks' links. It uses BeautifulSoup4 to parse the HTML and regular expression code to identify the Top 100 eBook file numbers.

1. Import necessary libraries including regex, and beautifulsoup

In [1]:

```
import urllib.request, urllib.parse, urllib.error
import requests
from bs4 import BeautifulSoup
import ssl
import re
```

2. Check SSL certificate

In [2]:

```
# Ignore SSL certificate errors
ctx = ssl.create_default_context()
ctx.check_hostname = False
ctx.verify_mode = ssl.CERT_NONE
```

3. Read the HTML from the URL

In [3]:

```
# Read the HTML from the URL and pass on to BeautifulSoup
top100url = 'https://www.gutenberg.org/browse/scores/top'
response = requests.get(top100url)
```

4. Write a small function to check the status of web request

In [4]:

```
def status_check(r):
    if r.status_code==200:
        print("Success!")
        return 1
    else:
        print("Failed!")
        return -1
```

In [5]:
```
status_check(response)
```

```
Success!
```

Out[5]:
```
1
```

5. Decode the response and pass on to `BeautifulSoup` for HTML parsing

In [6]:
```
contents = response.content.decode(response.encoding)
```

In [7]:
```
soup = BeautifulSoup(contents, 'html.parser')
```

6. Find all the *href* tags and store them in the list of links. Check how the list looks like - print first 30 elements

In [8]:
```
# Empty list to hold all the http links in the HTML page
lst_links=[]
```

In [13]:
```
# Find all the href tags and store them in the list of links
# href tags are in the 'a' tag
for link in soup.find_all('a'):
    #print(link.get('href'))
    lst_links.append(link.get('href'))

len(lst_links)
```

Out[13]:
```
1310
```

In [22]:
```
soup.find_all('a')[:5]
```

Out[22]:
```
[<a class="logo" href="/wiki/Main_Page" tabindex="1" title="Go to Main Pa
ge"><img alt="" class="logo" height="80" src="/pics/pg-logo-002.png" widt
h="129"/></a>,
 <a accesskey="1" class="h1" href="/catalog/" tabindex="30" title="Go to
the online book catalog section - Accesskey=1">Online Book Catalog</a>,
 <a accesskey="s" href="/ebooks/" tabindex="30" title="Go to book search
page - Accesskey=s">Book  Search</a>,
 <a accesskey="r" href="/browse/recent/last1" tabindex="31" title="Go to
the Most recently posted books page - Accesskey=r">Recent  Books</a>,
 <a accesskey="p" href="/browse/scores/top" tabindex="32" title="Go to th
e Top 100 books and authors page - Accesskey=p">Top  100</a>]
```

In [10]:

```
lst_links[:30]
```

Out[10]:

```
['/wiki/Main_Page',
 '/catalog/',
 '/ebooks/',
 '/browse/recent/last1',
 '/browse/scores/top',
 '/wiki/Gutenberg:Offline_Catalogs',
 '/catalog/world/mybookmarks',
 '/wiki/Main_Page',
 'https://www.paypal.com/xclick/business=donate%40gutenberg.org&item_name
=Donation+to+Project+Gutenberg',
 '/wiki/Gutenberg:Project_Gutenberg_Needs_Your_Donation',
 'http://www.ibiblio.org',
 'http://www.pgdp.net/',
 'pretty-pictures',
 '#books-last1',
 '#authors-last1',
 '#books-last7',
 '#authors-last7',
 '#books-last30',
 '#authors-last30',
 '/ebooks/1342',
 '/ebooks/11',
 '/ebooks/84',
 '/ebooks/1952',
 '/ebooks/43',
 '/ebooks/844',
 '/ebooks/25525',
 '/ebooks/98',
 '/ebooks/2542',
 '/ebooks/74',
 '/ebooks/215']
```

7. Use regular expression to find the numeric digits in these links. These are the file number for the Top 100 books.

In [23]:

```
booknum=[]
for i in range(19,119):
    link=lst_links[i]
    link=link.strip()
    # Regular expression to find the numeric digits in the link (href) st
ring
    n=re.findall('[0-9]+',link)
    if len(n)==1:
        # Append the filenumber casted as integer
        booknum.append(int(n[0]))
```

Print the file numbers

In [14]:
```python
print ("\nThe file numbers for the top 100 ebooks on Gutenberg are shown
 below\n"+"-"*70)
print(booknum)
```

```
The file numbers for the top 100 ebooks on Gutenberg are shown below
----------------------------------------------------------------------
[1342, 84, 1080, 46, 219, 2542, 98, 345, 2701, 844, 11, 5200, 43, 16328,
76, 74, 1952, 6130, 2591, 1661, 41, 174, 23, 1260, 1497, 408, 3207, 1400,
30254, 58271, 1232, 25344, 58269, 158, 44881, 1322, 205, 2554, 1184, 260
0, 120, 16, 58276, 5740, 34901, 28054, 829, 33, 2814, 4300, 100, 55, 160,
1404, 786, 58267, 3600, 19942, 8800, 514, 244, 2500, 2852, 135, 768, 5826
3, 1251, 3825, 779, 58262, 203, 730, 20203, 35, 1250, 45, 161, 30360, 737
0, 58274, 209, 27827, 58256, 33283, 4363, 375, 996, 58270, 521, 58268, 3
6, 815, 1934, 3296, 58279, 105, 2148, 932, 1064, 13415]
```

9. What does the soup object's text look like? Use the .text method and print only the first 2,000 characters (do not print the whole thing, as it is too long).

In [32]:
```python
print(soup.text[:2000])
```

Out[32]:
```
26025
```

10. Search in the extracted text (using regular expression) from the `soup` object to find the names of top 100 Ebooks (Yesterday's rank)

In [34]:
```python
# Temp empty list of Ebook names
lst_titles_temp=[]
```

11. Create a starting index. It should point at the text *"Top 100 Ebooks yesterday"*.

In [35]:
```python
start_idx=soup.text.splitlines().index('Top 100 EBooks yesterday')
```

12. Loop 1-100 to add the strings of next 100 lines to this temporary list. Hint: `splitlines()` method

In [36]:
```python
for i in range(100):
    lst_titles_temp.append(soup.text.splitlines()[start_idx+2+i])
```

13. Use regular expression to extract only text from the name strings and append to an empty list

In [48]:

```
lst_titles=[]
for i in range(100):
    id1,id2=re.match('^[a-zA-Z ]*',lst_titles_temp[i]).span()
    lst_titles.append(lst_titles_temp[i][id1:id2])

# Print the list of titles
rank = 1
for l in lst_titles:
    print(str(rank) + ' ' + l)
    rank += 1
```

1 Pride and Prejudice by Jane Austen
2 Alice
3 Frankenstein
4 The Yellow Wallpaper by Charlotte Perkins Gilman
5 The Strange Case of Dr
6 The Importance of Being Earnest
7 The Works of Edgar Allan Poe
8 A Tale of Two Cities by Charles Dickens
9 Et dukkehjem
10 The Adventures of Tom Sawyer by Mark Twain
11 The Call of the Wild by Jack London
12 Ion by Plato
13 The Adventures of Sherlock Holmes by Arthur Conan Doyle
14 A Modest Proposal by Jonathan Swift
15 Moby Dick
16 Treasure Island by Robert Louis Stevenson
17 Anthem by Ayn Rand
18 The Picture of Dorian Gray by Oscar Wilde
19 Adventures of Huckleberry Finn by Mark Twain
20 Little Women by Louisa May Alcott
21 Great Expectations by Charles Dickens
22 Peter Pan by J
23 A Journal of the Plague Year by Daniel Defoe
24 A Christmas Carol in Prose
25 Metamorphosis by Franz Kafka
26 The Wonderful Wizard of Oz by L
27 The Hound of the Baskervilles by Arthur Conan Doyle
28 Dracula by Bram Stoker
29 Walden
30 Beowulf
31 Jane Eyre
32 Siddhartha by Hermann Hesse
33 Dubliners by James Joyce
34 Grimms
35 War and Peace by graf Leo Tolstoy
36 Ulysses by James Joyce
37 The Secret Garden by Frances Hodgson Burnett
38 Anne of Green Gables by L
39 Emma by Jane Austen
40 The Scarlet Letter by Nathaniel Hawthorne
41 Il Principe
42 Tractatus Logico
43 Bunner Sisters by Edith Wharton
44 The Last Martian by Ray Van Houten
45 Wuthering Heights by Emily Bront
46 The Awakening
47 Heart of Darkness by Joseph Conrad
48 The Count of Monte Cristo
49 Uncle Tom
50 The Masque of the Red Death by Edgar Allan Poe
51 The War of the Worlds by H
52 The Wonderful Wizard of Oz by L
53 An Autobiography by Elizabeth Butler
54 Frankenstein
55 Pygmalion by Bernard Shaw
56 Prestuplenie i nakazanie
57 A History of Epidemics in Britain

```
58 The Brothers Karamazov by Fyodor Dostoyevsky
59 The Iliad by Homer
60 The Kama Sutra of Vatsyayana by Vatsyayana
61 The Republic by Plato
62 The Life and Adventures of Robinson Crusoe by Daniel Defoe
63 The Sign of the Four by Arthur Conan Doyle
64 A Study in Scarlet by Arthur Conan Doyle
65 Also sprach Zarathustra
66 The Prophet by Kahlil Gibran
67 Candide by Voltaire
68 Essays of Michel de Montaigne
69 The Souls of Black Folk by W
70 The Star Mouse by Fredric Brown
71 The Odyssey by Homer
72 The Jungle Book by Rudyard Kipling
73 Les Mis
74 The Decameron of Giovanni Boccaccio by Giovanni Boccaccio
75 Don Quixote by Miguel de Cervantes Saavedra
76 A Pickle for the Knowing Ones by Timothy Dexter
77 The Secret Garden by Frances Hodgson Burnett
78 The Time Machine by H
79 An Index of The Divine Comedy by Dante by Dante Alighieri
80 Oliver Twist by Charles Dickens
81 The Ballad of Venus Nell by Nelson S
82 David Copperfield by Charles Dickens
83 The Jungle by Upton Sinclair
84 The Mysterious Affair at Styles by Agatha Christie
85 Beyond Good and Evil by Friedrich Wilhelm Nietzsche
86 Sense and Sensibility by Jane Austen
87 Index of Project Gutenberg Works on Black History by Various
88 The Voyage Out by Virginia Woolf
89 Narrative of the Life of Frederick Douglass
90 Meditations by Emperor of Rome Marcus Aurelius
91 Leviathan by Thomas Hobbes
92 Anna Karenina by graf Leo Tolstoy
93 The Roman Index of Forbidden Books by Francis S
94 Gulliver
95 The Tale of Peter Rabbit by Beatrix Potter
96 Calculus Made Easy by Silvanus P
97 The Memoirs
98 The Devil
99 The Works of Edgar Allan Poe
100 Songs of Innocence
```