Book Summary Extraction

In [19]:
```
import requests
import pandas as pd
import re
from bs4 import BeautifulSoup


import os
import urllib.request as request

import numpy as np

import warnings
warnings.filterwarnings("ignore")
```

In [117]:
```
def find_book_url(page_url):
    with request.urlopen(page_url) as res:
        print(page_url," parsing.....")
        source_data = res.read()

    # parse html content
    soup = BeautifulSoup(source_data , 'html.parser')

    goodreads_url = "https://www.goodreads.com/"
    table_class = soup.find_all( class_ = "js-tooltipTrigger tooltipTrigger" )


    for tag in table_class:
        raw_data_ = tag.find("a")
        title_ = str(raw_data_).split(">")[0].split(" title=")[1].replace('"','')
        href_ = goodreads_url + str(raw_data_).split(">")[0].split(" ")[1].split("href=")[1].replace('"','')
        if title_ and href_:
            book_title.append(title_)
            book_link.append(href_)
        else:
            book_title.append("")
            book_link.append("")
```

In [118]:
```
book_title=[]
book_link=[]
for i in range(1,101):
    url_= "https://www.goodreads.com/list/show/1.Best_Books_Ever?page=" + str(i)
    find_book_url(url_)
```

```
oks_Ever?page=51)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=52 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=52)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=53 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=53)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=54 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=54)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=55 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=55)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=56 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=56)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=57 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=57)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=58 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=58)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=59 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=59)  parsing.....
https://www.goodreads.com/list/show/1.Best_Books_Ever?page=60 (https://www.goodreads.com/list/show/1.Best_Bo
oks_Ever?page=60)  parsing.....
```

In [121]:
```
df = pd.DataFrame(list(zip(book_title, book_link)),
                  columns =['Title', 'book_url'])
```

In [122]:
```
df.to_csv("Goodreads_books_links.csv")
```

In [14]:
```
def get_summary(book_url):
    with request.urlopen(book_url) as res:
        source_data_ = res.read()
    soup = BeautifulSoup(source_data_ , 'html.parser')

    # get summary
    summary_class = soup.find("div", {"id": "description"})
    book_sum= summary_class.text

    # get author name
    author_class = soup.find(class_ ="authorName")

    # get bookcover image
    bookcover_class = soup.find(class_ ="bookCoverPrimary")
    img_data = bookcover_class.find_all("a")
    if img_data:
        cover_img_ = str(img_data).split("src=")[1].split("jpg")[0].replace('"','')+"jpg"
    else:
        cover_img_=""




    book_summaries.append(book_sum.replace("\n","")[:-7])
    book_cover.append(cover_img_)
    author_name.append(author_class.text)
```

In [20]:
```
import pandas as pd
data_= pd.read_csv("Goodreads_books_links.csv")
book_title=[]
book_link=[]
for i,j in zip(data_["Title"].values,data_["book_url"].values):
    book_title.append(i)
    book_link.append(j)
```

In [ ]:
```
book_summaries =[]
book_cover=[]
author_name=[]

#Do the batch scraping to avoid server side errors

for count,link in enumerate(book_link[1000:1100],1000):
    #print(count," :",link," collecting data....")
    get_summary(link)
```

In [63]:
```
books_df = pd.DataFrame(list(zip(book_title[1000:1100],author_name, book_summaries, book_cover, book_link[1000:1
                columns =['title', 'author','summary','bookcover','book_url'])
```

In [64]:
```
books_df.to_csv("goodreads_book_summaries_13.csv")
```

In [68]:
```
# merging all csv files
csv_files= ['goodreads_book_summaries_1.csv','goodreads_book_summaries_2.csv', 'goodreads_book_summaries_3.csv',
            'goodreads_book_summaries_4.csv','goodreads_book_summaries_5.csv', 'goodreads_book_summaries_6.csv',
            'goodreads_book_summaries_7.csv','goodreads_book_summaries_8.csv', 'goodreads_book_summaries_9.csv',
            'goodreads_book_summaries_10.csv','goodreads_book_summaries_11.csv', 'goodreads_book_summaries_12.csv
df = pd.concat(
    map(pd.read_csv, csv_files), ignore_index=True)
df
```

In [69]:
```
df.head()
```

Out[69]:

	title	author	summary	bookcover	book_url
0	The Hunger Games	Suzanne Collins	Could you survive on your own in the wild, wit...	https://i.gr-assets.com/images/S/compressed.ph...	https://www.goodreads.com//book/show/2767052-t...
1	Harry Potter and the Order of the Phoenix	J.K. Rowling	There is a door at the end of a silent corrido...	https://i.gr-assets.com/images/S/compressed.ph...	https://www.goodreads.com//book/show/2.Harry_P...
2	To Kill a Mockingbird	Harper Lee	The unforgettable novel of a childhood in a sl...	https://i.gr-assets.com/images/S/compressed.ph...	https://www.goodreads.com//book/show/2657.To_K...
3	Pride and Prejudice	Jane Austen	Alternate cover edition of ISBN 9780679783268S...	https://i.gr-assets.com/images/S/compressed.ph...	https://www.goodreads.com//book/show/1885.Prid...
4	The Book Thief	Markus Zusak	Librarian's note: An alternate cover edition c...	https://i.gr-assets.com/images/S/compressed.ph...	https://www.goodreads.com//book/show/19063.The...

In [67]:
```
df.to_csv("goodreads_book_summaries.csv")
```