Brief description:

This project is about collecting data from a books website by scrapping it, then clean the collected data and auto-correct it and save it as a csv file, then make a classifier that classifies the section of a book by its summary.

What did we do in this first phase ?

We **scrapped** the website to **collect** the needed data like book **names, author and download link** for each book.

and then we **cleaned** the collected data by **removing duplications** from book names, then we **removed the html tags** surrounding of link,

and then we **removed stop words** like "pdf" from books names and **autocorrected** the book names.

after all the previous phases of collecting and cleaning the data we convert it to data frame and **save it into csv file**.

We trained naïve bayes multinomial classifier its input is the summary of the book and its output is the section of the book, we trained it on 80% from the data and tested it on the other 20%.

What did we use?

sklearn  
requests  
itertools  
re  
nltk  
beautifulsoup  
selenium  
panads  
textblob\_ar