Brief description:

This project is about making information retrieval system using clustering of the data that we should collect from books website and clean it.

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**.

What did we use?

sklearn  
requests  
itertools  
re  
nltk  
beautifulsoup  
selenium  
panads  
textblob\_ar