In []:

```
from bs4 import BeautifulSoup
import requests
url_1 = "http://theses.fr/fr/?q=&fq=dateSoutenance:(1965-01-01T23:59:59Z%2BTO%2B20
Outre & checkedfacets = & start = " URL_2 = 1
url_3 = " & sort = none & status = & access = & forecast = & filtrepersonne = & ar
ids = [ ]
for cpt , _ in enumerate ( range ( 300 ), start = 1 ):
    url = url_1 + str ( url_2 ) + url_3
    r = requests . get ( url )
    soup = BeautifulSoup ( r . text , "html.parser" )

    for k in soup . find_all ( "h2" ):
        ids . append ( k . find ( "a" ) . get ( "href" ))
    url_2 += 10
    # if (cpt% 1 == 0):
    print ( cpt , end = " \ r " )
```

In []:

```
url_1 = "https://theses.fr"
final = []
for cpt , i in enumerate ( ids ):
    url = url_1 + i
    r = requests . get ( url )
    soup = BeautifulSoup ( r . text , "html.parser" )
    title = ""
    if ( soup . find ( "meta" , attrs = { "name": "DC.title" })):
        title = ( soup . Find ( "meta" , attrs = { "name" : "DC.title" }) . Get
    elif ( soup . Find ( " meta " , attrs = { " name " : " title " })):
        title = ( soup . find ( " meta " , attrs = { "name" :"title" }) . get (
    else :
        title = ""
    if ( soup . find ( "meta" , attrs = { "name" : "DC.creator" })):
        author = soup . find ( "meta" , attrs = { "name" : "DC.creator" }) . ge
    else :

     soup . find ( "meta" , attrs = { "name" : "DC.subject" , "xml: lang" : "FR"
        fr_subject = soup . find ( "meta" , attrs = { "name" : "DC.subject" , "
    else :
        fr_subject = ""
      , author , fr_subject ]
    if ( cpt % 10 == 0 ):
        print ( cpt , end = " \ r " )
    final . append ( all )
```

In [5]:

```
import pandas as pd

scrapped_df = pd . DataFrame ( final , columns = [ "ID" , "Title" , "Author"
scrapped_df
```

Out [5]:

	ID	Title	Author	Subject
0	/ s230875	The development of short food circuits ...	Camille Horvath	Incentive actions Development of territories ...
1	/ s252662	Evaluation of the emotional states of the human being ...	Ines Elali	Emotions Wine Psychophysical Tasting Analyz ...
2	/ s299718	Approach to school architecture: study of ...	Sonia vermeulen	School architecture Well-being spaces through ...

	ID	Title	Author	Subject
3	/ s260915	Using and Interacting with AI-Based Intelligen ...	Giulia Pavone	in French
4	/ s299686	The management of natural resources in law ...	Thomas Abdo	Environmental protection Natural resources ...
...	...	...	...	...
2995	/ s294505	IMPACT OF VOLCANIC SULFUR EMISSIONS ON ...	Claire Lamotte	in French
2996	/ s294504	GENERATIONS OF VIRTUAL URBAN ENVIRONMENTS ...	Tiavina tantely Nivolala	in French
2997	/ s294503	GE REAL-TIME DISTRIBUTED OPTIMIZATION ...	Jean-baptiste Blanc - rouchosse	in French
2998	/ s294502	EPISTEMIC LOGICS AXIOMATIZATIONS AND FRAGM ...	Benito fabian Romero jimenez	in French
2999	/ s294501	BEHAVIORAL SINGULARITIES IN HABITAT I ...	Maxime Houssin	in French

3000 rows × 4 columns

In [7]:
```
scrapped_df . to_csv ( "data / scrapped_data.csv" )
```

In []:
```
```