

# data extraction final

February 12, 2023

```
[1]: #importing library
import pandas as pd
import requests
from bs4 import BeautifulSoup
import numpy as np

[2]: #importing input file
df=pd.read_csv('Input.csv')[['URL_ID','URL','title']]

[3]: df=df.iloc[0:114]

[4]: df.drop('URL_ID',axis=1,inplace=True)

[5]: my_data = []
#extracting text from all the url
for i in range(0,114):
    #j=df.iloc[i].values

    df=pd.read_csv('Input.csv')[['URL_ID','URL','title']]
    site = str(df['URL'][i])
    page = requests.get(site)
    if(page.status_code==200):
        print('data fetched succesfully',i)
        page = requests.get(site)
        soup = BeautifulSoup(page.content, 'html.parser')
        content = soup.findAll(attrs = {'class' : 'td-post-content'})
        content1=content[0].text.replace('\xa0'," ").replace('\n'," ")
        title = df['title']
        data = [[title[i] , content1]]
        df1 = pd.DataFrame(data, columns = [ 'title', 'text'])

        solution = ([title[i],content1])
        #y = data.insert(i,[title[i],[content1]])
        my_data.append(solution)
```

```
else:
    print('wrong',i)
df2 = pd.DataFrame(my_data, columns = [ 'title', 'text'])
df2.to_csv('pagedata.csv')
```

```
data fetched succesfully 0
data fetched succesfully 1
data fetched succesfully 2
data fetched succesfully 3
data fetched succesfully 4
data fetched succesfully 5
data fetched succesfully 6
wrong 7
data fetched succesfully 8
data fetched succesfully 9
data fetched succesfully 10
data fetched succesfully 11
data fetched succesfully 12
data fetched succesfully 13
data fetched succesfully 14
data fetched succesfully 15
data fetched succesfully 16
data fetched succesfully 17
data fetched succesfully 18
data fetched succesfully 19
wrong 20
data fetched succesfully 21
data fetched succesfully 22
data fetched succesfully 23
data fetched succesfully 24
data fetched succesfully 25
data fetched succesfully 26
data fetched succesfully 27
data fetched succesfully 28
data fetched succesfully 29
data fetched succesfully 30
data fetched succesfully 31
data fetched succesfully 32
data fetched succesfully 33
data fetched succesfully 34
data fetched succesfully 35
data fetched succesfully 36
data fetched succesfully 37
data fetched succesfully 38
data fetched succesfully 39
data fetched succesfully 40
data fetched succesfully 41
```

data fetched successfully 42  
data fetched successfully 43  
data fetched successfully 44  
data fetched successfully 45  
data fetched successfully 46  
data fetched successfully 47  
data fetched successfully 48  
data fetched successfully 49  
data fetched successfully 50  
data fetched successfully 51  
data fetched successfully 52  
data fetched successfully 53  
data fetched successfully 54  
data fetched successfully 55  
data fetched successfully 56  
data fetched successfully 57  
data fetched successfully 58  
data fetched successfully 59  
data fetched successfully 60  
data fetched successfully 61  
data fetched successfully 62  
data fetched successfully 63  
data fetched successfully 64  
data fetched successfully 65  
data fetched successfully 66  
data fetched successfully 67  
data fetched successfully 68  
data fetched successfully 69  
data fetched successfully 70  
data fetched successfully 71  
data fetched successfully 72  
data fetched successfully 73  
data fetched successfully 74  
data fetched successfully 75  
data fetched successfully 76  
data fetched successfully 77  
data fetched successfully 78  
data fetched successfully 79  
data fetched successfully 80  
data fetched successfully 81  
data fetched successfully 82  
data fetched successfully 83  
data fetched successfully 84  
data fetched successfully 85  
data fetched successfully 86  
data fetched successfully 87  
data fetched successfully 88  
data fetched successfully 89

```

data fetched succesfully 90
data fetched succesfully 91
data fetched succesfully 92
data fetched succesfully 93
data fetched succesfully 94
data fetched succesfully 95
data fetched succesfully 96
data fetched succesfully 97
data fetched succesfully 98
data fetched succesfully 99
data fetched succesfully 100
data fetched succesfully 101
data fetched succesfully 102
data fetched succesfully 103
data fetched succesfully 104
data fetched succesfully 105
data fetched succesfully 106
wrong 107
data fetched succesfully 108
data fetched succesfully 109
data fetched succesfully 110
data fetched succesfully 111
data fetched succesfully 112
data fetched succesfully 113

```

#Three URL Not found That why we drop them

```
[7]: df2
```

```

[7]:
0          ai-in-healthcare-to-improve-patient-outcomes
1    what-if-the-creation-is-taking-over-the-creator
2    what-jobs-will-robots-take-from-humans-in-the-...
3    will-machine-replace-the-human-in-the-future-o...
4          will-ai-replace-us-or-work-with-us
..
106          blockchain-for-payments
107          the-future-of-investing
108          big-data-analytics-in-healthcare
109    business-analytics-in-the-healthcare-industry
110 challenges-and-opportunities-of-big-data-in-he...

          title \
0    Introduction  "If anything kills over 10 mil...
1    Human minds, a fascination in itself carryin...
2    Introduction  AI is rapidly evolving in the ...
3    "Anything that could give rise to smarter-th...
4    "Machine intelligence is the last invention ...

```

```
..
106   Reconciling with the financial realities of ...
107   What Is an Investment?  An investment is a r...
108   Quality and affordable healthcare is a visio...
109   Analytics is a statistical scientific proces...
110   Big Data  To begin with I shall first like t...
```

```
[111 rows x 2 columns]
```

```
[ ]:
```