In [1]:
```
from bs4 import BeautifulSoup
import re
import unicodedata
import xml.etree.cElementTree as ET
import pandas as pd
```

In [15]:
```
repo = []
count  = 0
for i in range (0,1007):
    try:
        html_page = open("G:\\New folder\\springer_papers\\springer_" + str(i) + ".html", "r",encoding='utf-8')
        soup = BeautifulSoup(html_page, "html.parser")
        text1 = soup.find("meta", {"name" : "prism.doi"  }).attrs['content'] #storing doi
        text2 = soup.find("meta", {"name" : "dc.title"  }).attrs['content'] # title
        text3 = soup.find("meta", {"name" : "dc.description"  }).attrs['content'] # storing abstract
        paper = [text1, text2 ,text3]
        li = []  # for storing the paragraphs
        if len(soup.findAll("p"))>60:  # finding of para tags whick should be greater > 60 i.e. atleast 30 para tags
            for tag in soup.findAll("p"):
                if len(tag.get_text())<100: # if the words <100 then ignore.
                    continue
                text = tag.get_text() # extract the text
                text =  unicodedata.normalize("NFKD",text) # for removing  non-breaking space
                pattern = 'supported by|Department of|Département | London | Institute of| Springer | University  | School of |http|[A-Z]\.'  # for removing unnecesary links
                match= re.findall(pattern, text)
                if len(match)>0: # there is link in between the paragraph then ignore
                    continue
                li.append(text)
        paper.append(li)
        count += len(li)
        paper.append(len(li)) # append No. of para
        repo.append(paper)







    except:
        continue


print("lenght of repo is ", len(repo) , 'len of para: ', count)
```

```
lenght of repo is  987 len of para:  19289
```

In [16]:
```
df = pd.DataFrame(repo)
df.index.names = ["Serial_No"]
df.columns = ["DOI","Title","Astract","Para_list" , "Para_count"]
pd.set_option('display.max.rows',8580)
df.head(5)
```

Out[16]:

Serial_No	DOI	Title	Astract	Para_list	Para_count
0	doi:10.1557/mrc.2020.34	Robust resistive switching performance of puls...	In this work, the authors developed SiC(10 nm)...	[In this work, the authors developed SiC(10 nm...	1
1	doi:10.1557/mrc.2018.208	Influence of electrolyte substrates on the Sr-...	To systematically investigate the influence of...	[To systematically investigate the influence o...	2
2	doi:10.1557/mrc.2020.34	Robust resistive switching performance of puls...	In this work, the authors developed SiC(10 nm)...	[In this work, the authors developed SiC(10 nm...	1
3	doi:10.1557/mrc.2018.208	Influence of electrolyte substrates on the Sr-...	To systematically investigate the influence of...	[To systematically investigate the influence o...	2
4	doi:10.1557/jmr.2018.422	Misfit strain relaxations of (101)-oriented fe...	High-index ferroelectric thin films show excel...	[High-index ferroelectric thin films show exce...	23

In [17]:
```
df.to_csv("G:\\paper_to_para\\html_1007_para.csv",index = False )
```

In [18]:
```
paras = []
for para_list in df['Para_list']:
    for paragraphs in para_list:
        paras.append(paragraphs)


len(paras)
```

Out[18]:
```
19289
```

In [19]:
```
para_df = pd.DataFrame(paras)
para_df.index.names = ['Serial_No']
para_df.columns = ["Paragraph"]
para_df.head(10)
```

Out[19]:

Serial_No	Paragraph
0	In this work, the authors developed SiC(10 nm)...
1	To systematically investigate the influence of...
2	This work is supported in part by the New Ener...
3	In this work, the authors developed SiC(10 nm)...
4	To systematically investigate the influence of...
5	This work is supported in part by the New Ener...
6	High-index ferroelectric thin films show excel...
7	Recently, perovskite oxide thin films have att...
8	In the past decades, many studies indicated th...
9	For decades, MDs in perovskite ferroelectric t...

In [20]:
```
para_df.to_csv("G:\\paper_to_para\\html_1007_only_para.csv",index = False )
```

Electrochemical papers paragraph Extraction

In [21]:
```
paper_count = 0  # counts total number of papers from 10K publisherless dois
paper_read = 0
repo1 = []
for i in range (0,7689):
    try:
        html_page = open("G:\\New folder\\Electrochemical_society\\random_" + str(i) + ".html", "r",encoding='utf-8')
        soup = BeautifulSoup(html_page, "html.parser")
        paper_count = paper_count + 1
        text1 = soup.find("meta", {"name" : "prism.doi"  }).attrs['content'] #storing doi
        text2 = soup.find("meta", {"name" : "dc.title"  }).attrs['content'] # title
        text3 = soup.find("meta", {"name" : "dc.description"  }).attrs['content'] # storing abstract
        paper_read = paper_read + 1
        paper = [text1, text2 ,text3]
        li = []  # for storing the paragraphs
        if len(soup.findAll("p"))>60:  # finding of para tags whick should be greater > 60 i.e. atleast 30 para tags
            for tag in soup.findAll("p"):
                if len(tag.get_text())<100: # if the words <100 then ignore.
                    continue
                text = tag.get_text() # extract the text
                text =  unicodedata.normalize("NFKD",text) # for removing  non-breaking space
                pattern = 'supported by|Department of|Département | London | Institute of| Springer | University  | School of |http|[A-Z]\.'    # for removing unnecesary links
                match= re.findall(pattern, text)
                if len(match)>0: # there is link in between the paragraph then ignore
                    continue
                li.append(text)
        paper.append(li)
        paper.append(len(li)) # append No. of para

        repo1.append(paper)







    except:
        continue

print("lenght of repo is ", len(repo1))
```

```
lenght of repo is  1187
```

In [22]:
```
df = pd.DataFrame(repo1)
df.index.names = ["Serial_No"]
df.columns = ["DOI","Title","Astract","Para_list" , "Para_count"]
pd.set_option('display.max.rows',8580)
df.head(5)
```

Out[22]:

Serial_No	DOI	Title	Astract	Para_list	Para_count
0	doi:10.1007/s00170-012-4402-y	Spatter reduction in nanosecond fibre laser dr...	Pulsed wave fibre lasers are becoming a popula...	[\nThe International Journal of Advanced Manuf...	23
1	doi:10.1007/s00339-012-7211-0	Laser ablation in a running hall effect thrust...	Hall Effect Thrusters (HETs) are promising ele...	[Hall Effect Thrusters (HETs) are promising el...	22
2	doi:10.1007/s00339-012-7216-8	Study of optical properties and biocompatibili...	Optical and biomedical properties of diamond-l...	[Optical and biomedical properties of diamond-...	20
3	doi:10.1007/s00339-012-7223-9	Correlation of plume dynamics and oxygen press...	Vanadium dioxide thin films have been deposite...	[Vanadium dioxide thin films have been deposit...	19
4	doi:10.1007/s00339-012-7324-5	Electron-beam deposition of vanadium dioxide t...	Developing a reliable and efficient fabricatio...	[Developing a reliable and efficient fabricati...	22

In [23]:
```
df.to_csv("G:\\paper_to_para\\random_para.csv",index = False )
```

In [24]:
```
paras = []
for para_list in df['Para_list']:
    for paragraphs in para_list:
        paras.append(paragraphs)


len(paras)
```

Out[24]:
```
9703
```

In [25]:
```
para_df = pd.DataFrame(paras)
para_df.index.names = ['Serial_No']
para_df.columns = ["Paragraph"]
para_df.head(10)
```

Out[25]:

Serial_No	Paragraph
0	\nThe International Journal of Advanced Manufa...
1	Pulsed wave fibre lasers are becoming a popula...
2	Chichkov BN, Momma C, Nolte S, Von Alvensleben...
3	Pronko PP, Dutta SK, Squier J, Rudd JV, Du D, ...
4	Tam SC, Willams R, Yang LJ, Jana S, Lim LEN, L...
5	Dijk MHHV, Vlieger GD, Brouwer JE(1989) Laser ...
6	Corfe AG (1983) Laser drilling of aero-engine ...
7	Naeem M, Chinn J (2008) Advancement in laser d...
8	Hainsey RF, Hooper AE, Swenson EJ, Nashner MS ...
9	Meijer J, Du K, Gilner A, Hoffmann D, Kovalenk...

In [26]:
```
para_df.to_csv("G:\\paper_to_para\\random_only_para.csv",index = False )
```