ScanlanMatthew_project3_Epub_to_txt_

March 4, 2022

[1]:

```
from pathlib import PureWindowsPath
from random import shuffle
from re import sub
from glob import glob
from os.path import basename, mkdir
from bs4 import BeautifulSoup as bs
```

[ ]:

```
#This ipynb processes Epub files that are not included in the GitHub Repo as I␣
 ↪do not have permission to republish the original works
#The files included have been converted to .txt and cleaned
#the content was shuffled to simulate real life text documents
```

[ ]:

```
def extract_content(fullpath,bookname,chapter,shufflefilepathlist):
    with open(fullpath,'r',encoding="UTF-8") as f:
        a = f.read()
        soup = bs(a)
    content=''
    with open('extractandshuffle/'+bookname+chapter,'w+',encoding="UTF-8") as n:
        shufflefilepathlist.append('extractandshuffle/'+bookname+chapter)
        for a in soup.find_all("p"):
            data=a.get_text()
            data=sub('[^a-zA-Z \']'," ",data)
            data=list(data.split(' '))
            shuffle(data)
            content =content+' '.join(word for word in data)
        n.write(content)
    return shufflefilepathlist
```

[ ]:

```
files = glob.glob("extractandshuffle1")
files.pop(0)
for x in range(len(files)):
    f = PureWindowsPath(files[x])
    files[x] = f.as_posix()
```

[ ]:

```
filenames=[]
for x in range(len(files)):
    fn=basename(files[x])
    fn=fn.replace('.xhtml','.txt')
    filenames.append(fn)
filenames[0]
```

[ ]:

```
bookname=[]
b=1
bn=[]
for x in files:
    bn.append(x.split('/')[1])

for x in range(len(bn)):
    if bn[x]== bn[x-1]:
        book="_book_"+str(b)+'_'
    else:
        b+=1
        book="_book_"+str(b)+'_'
    bookname.append(book)
```

[ ]:

```
try:
    mkdir('extractandshuffle')
except:
    pass
```

[ ]:

```
shufflefilepathlist=[]
from itertools import zip_longest as zipp
for (x,y,z) in zipp(files,bookname,filenames):
    shufflefilepathlist = extract_content(x,y,z,shufflefilepathlist)
```