Selected 3 Project

there is two phases into this project

phase 1

Project Requirements :

1- Make scrapping on the website and collect data

2- clean the scrapped data

3- auto correct the scrapped data

4- Save your data in csv file "Data sheet"

with the columns like " Author - title - publishing year - link for pdf - etc. "

5- cluster this data for topic modeling

The website for scrapping : https://www.arab-books.com

importing libraries

In [1]:
```
from bs4 import BeautifulSoup
import requests
import csv
from itertools import zip_longest
import re
import pandas as pd
import string
import nltk
from nltk.tokenize import word_tokenize , sent_tokenize
import numpy as np
```

1- Make scrapping on the website: www.arab-books.com and collect data

Request to get all books from the website

there are **200 pages** on the website Each page have on average **30 book**

In [2]:
```
main_src = []
for i in range(1,201):
  main_src.append(requests.get("https://www.arab-books.com//page/"+str(i)).text)
```

In [3]:
```
main_soup = []
for src in main_src:
  main_soup.append(BeautifulSoup(src, 'lxml'))
```

In [4]:
```
all_books_src = []
for soup in main_soup :
 all_books_src = all_books_src + soup.find_all('li',class_='post-item tie-standard')
```

In [5]:
```
all_books_title=[]
for soup in main_soup :
  all_books_title = all_books_title + soup.find_all("div",{"class":"excerpt-book"})
```

In [6]:
```
all_books_title[0:10]
```

In [7]:
```
books_titles=[]
for title in all_books_title :
  books_titles.append(title.text)
len(books_titles)
```

Out[7]:
```
5984
```

In [8]:
```
url_pattern = "http([^\"]+)"
URl_RegEx = re.compile(url_pattern)
Books_urls=[]
for Object in all_books_title :
  myResult = URl_RegEx.search(str(Object))
  Books_urls.append(myResult.group())
Books_urls[:10]
```

Out[8]:
```
['https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d8%a7%d9%84%d8%aa%d8%b6%d8%ad%d9%8a%d8%a9-%d8%b9%d9%86%d8%af-%d8%a7%d9%84%d8%ad%d9%8a%d9%88%d8%a7%d9%86-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d9%85%d8%b9%d8%ac%d8%b2%d8%a9-%d8%a7%d9%84%d8%b0%d8%b1%d8%a9-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d9%84%d8%a7-%d8%aa%d8%aa%d8%ac%d8%a7%d9%87%d9%84-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d8%a7%d9%84%d8%ad%d9%8a%d8%a7%d8%a9-%d9%81%d9%8a-%d8%b3%d8%a8%d9%8a%d9%84-%d8%a7%d9%84%d9%84%d9%87-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d8%a7%d9%84%d8%b9%d8%b8%d9%85%d8%a9-%d9%81%d9%8a-%d9%83%d9%84-%d9%85%d9%83%d8%a7%d9%86-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d9%85%d8%b9%d8%ac%d8%b2%d8%a9-%d8%a7%d9%84%d8%ac%d9%87%d8%a7%d8%b2-%d8%a7%d9%84%d9%85%d9%86%d8%a7%d8%b9%d9%8a-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d8%a7%d9%84%d8%b1%d9%88%d9%85%d9%86%d8%b3%d9%8a%d8%a9-%d8%b3%d9%84%d8%a7%d8%ad-%d8%a8%d9%8a%d8%af-%d8%a7%d9%84%d8%b4%d9%8a%d8%b7%d8%a7%d9%86-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d8%b1%d8%ad%d9%84%d8%a9-%d9%81%d9%8a-%d8%a7%d9%84%d9%83%d9%88%d9%86-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d8%a7%d9%84%d8%aa%d8%b5%d9%85%d9%8a%d9%85-%d9%81%d9%8a-%d8%a7%d9%84%d8%b7%d8%a8%d9%8a%d8%b9%d8%a9-pdf/',
 'https://www.arab-books.com/books/%d9%83%d8%aa%d8%a7%d8%a8-%d9%85%d8%b9%d8%ac%d8%b2%d8%a9-%d8%ae%d9%84%d9%82-%d8%a7%d9%84%d8%a5%d9%86%d8%b3%d8%a7%d9%86-pdf/']
```

In [9]:
```
len(Books_urls)
```

Out[9]:
```
5984
```

We have now the url and title of **5984 books**

Request to get more information about each book

In [10]:
```
i=0
scoup_src = []
for book in Books_urls :
    scoup_src.append(requests.get(book,allow_redirects=False).text)
    i+=1
    if(i%200==0):
        print(len(scoup_src))
```

```
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
```

```
4000
4200
4400
4600
4800
5000
5200
5400
5600
5800
```

In [11]:
```
scoop_soup = []
for page in scoup_src:
  scoop_soup.append(BeautifulSoup(page, 'lxml'))
len(scoop_soup)
```

Out[11]: 5984

In [12]:
```
scoop_books_info = []
scoop_books_summary=[]
for soop_page in scoop_soup :
    scoop_books_info = scoop_books_info + soop_page.find_all('div',class_="book-info")
    scoop_books_summary = scoop_books_summary + soop_page.find_all('div',class_="entry-content entry clearfix"
```

In [13]:
```
scoop_books_info[0]
```

In [14]:
```
scoop_books_summary[0]
```

In [15]:
```
books_info=[]
for book_info in scoop_books_info:
    new_book_info=[]
    for info in book_info.find_all("li"):
        new_book_info.append(info.text.split(":"))
    books_info.append(new_book_info)
```

In [16]:
```
books_info[0]
```

In [17]:
```
src_download_url=[]
for book_url in scoop_soup:
  src_download_url = src_download_url + book_url.find_all('div',class_="down-link-bottom")
```

In [17]:
```
src_download_url[:5]
```

In [18]:
```
Books_download_urls=[]
for src_url in src_download_url :
  myResult = URl_RegEx.search(str(src_url))
  Books_download_urls.append(myResult.group())
Books_download_urls[:10]
```

In [19]:
```
books_titles[0]
```

In [20]:
```
scoop_books_summary[0]
```

In [21]:
```
books_summary=[]
for text in scoop_books_summary:
    books_summary.append(text.find_all('p'))
books_summary
```

In [22]:
```
books_summary[0]
```

In [23]:
```
Books_download_urls[0]
```

In [26]:
```
books_Authors = []
books_Class = []
```

```
books_Language=[]
books_Pages =[]
books_publisher=[]
book_Size =[]
book_format =[]
for Book_Info in books_info :
  books_Authors.append(Book_Info[0][1])
  books_Class.append(Book_Info[1][1])
  books_Language.append(Book_Info[2][1])
  books_Pages.append(Book_Info[3][1])
  books_publisher.append(Book_Info[4][1])
  book_Size.append(Book_Info[5][1])
  book_format.append(Book_Info[6][1])
```

In [24]:
```
len(book_Size)
```

Now we have all information about our 5984 books

Lets save the data in a csv

In [28]:
```
data_list=[books_titles,books_Authors,books_summary,books_Class,books_Language,books_Pages,books_publisher,bool
exported=zip_longest(*data_list)
```

In [29]:
```
with open("Scrapped_Data.csv","w",encoding="utf-8") as Scrapped_Data:
  wr=csv.writer(Scrapped_Data)
  wr.writerow(["BookTitle","Author","BookSummary","Class","Language","Pages","publisher","BookSize","format","D
  wr.writerows(exported)
```

2-Clean Scrapped Data

In [25]:
```
#read Scrapped_Dataset to clean it
data=pd.read_csv("Scrapped_Data.csv")
data.head(2)
```

Out[25]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL
0	كتاب\r\n التضحية عند الحيوان PDF\r\n	كتب\r\n الكاتب هارون يحيى\r\nكتب\r\n الكاتب هارون ...	الروايات<p>] والكتب العربية تعتبر من الروابط ...ب	تحميل\r\n كتب الأدب العربي\r\nتحميل كتب متنوع...	عربي	160 صفحة	عدنان أوكطار	5 ميغابايت	pdf	https://www.arab-books.com/books/%d9%83%d8%aa%...
1	كتاب\r\n معجزة الذرة PDF\r\n	كتب\r\n الكاتب هارون يحيى\r\nكتب\r\n الكاتب هارون ...	الروايات<p>] والكتب العربية تعتبر من الروابط ...ب	تحميل\r\n كتب تطوير الذات\r\nتحميل كتب متنوعة...	عربي	136 صفحة	عدنان أوكطار	3.41 ميغابايت	pdf	https://www.arab-books.com/books/%d9%83%d8%aa%...

In [26]:
```
data.shape
```

Out[26]:
```
(5984, 11)
```

In [27]:
```
data.tail(2)
```

Out[27]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL	
5982	القران\r\n الكريم باللغة الانجليزية - Quran wit...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	books.com/books/%
5983	\r\nLe Saint Coran en français\r\n	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	books.com/bo

In [28]:
```
#Remove dublicated and null rows
data=data.dropna()
data=data.drop_duplicates()
data.describe()
```

Out[28]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	Downloa
count	5898	5898	5898	5898	5898	5898	5898	5898	5898	
unique	5792	600	5370	290	21	791	784	1143	11	
top	كتاب\r\n نهاية العالم PDF\r\n	كتب\r\n أحمد خالد توفيق\r\n	[<p><span style="font-size: 24px; font-weight:...	تحميل\r\n كتب اسلامية\r\n	عربي	1 صفحة		1 ميغابايت	pdf	https://www. books.com/books/%d9%83%d8%a
freq	3	298	69	1230	3837	133	686	640	4048	

In [29]:
```
data.head(2)
```

Out[29]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL
0	كتاب\r\n التضحية عند الحيوان PDF\r\n	كتب\r\n الكاتب هارون يحيى\r\n كتب الكاتب هارون ...	[<p>الروايات والكتب العربية تعتبر من الروابط ب...	تحميل\r\n كتب الأدب العربي\r\n تحميل كتب متنوع...	عربي	160 صفحة	عدنان أوكطار	5 ميغابايت	pdf	https://www.arab-books.com/books/%d9%83%d8%aa%...
1	كتاب\r\n معجزة الذرة PDF\r\n	كتب\r\n الكاتب هارون يحيى\r\n كتب الكاتب هارون ...	[<p>الروايات والكتب العربية تعتبر من الروابط ب...	تحميل\r\n كتب تطوير الذات\r\n تحميل كتب متنوعة...	عربي	136 صفحة	عدنان أوكطار	3.41 ميغابايت	pdf	https://www.arab-books.com/books/%d9%83%d8%aa%...

In [30]:
```
data[["BookTitle","Author","BookSummary"]].head()
```

Out[30]:

	BookTitle	Author	BookSummary
0	كتاب التضحية عند الحيوان\r\n PDF\r\n	كتب الكاتب هارون\r\n كتب الكاتب هارون يحيى\r\n ...	[<p>الروايات والكتب العربية تعتبر من الروابط ب...
1	كتاب معجزة الذرة\r\n PDF\r\n	كتب الكاتب هارون\r\n كتب الكاتب هارون يحيى\r\n ...	[<p>الروايات والكتب العربية تعتبر من الروابط ب...
2	كتاب لا تتجاهل\r\n PDF\r\n	كتب الكاتب هارون\r\n كتب الكاتب هارون يحيى\r\n ...	[<p>الروايات والكتب العربية تعتبر من الروابط ب...
3	كتاب الحياة في سبيل الله\r\n PDF\r\n	كتب الكاتب هارون\r\n كتب الكاتب هارون يحيى\r\n ...	[<p>الروايات والكتب العربية تعتبر من الروابط ب...
4	كتاب العظمة في كل مكان\r\n PDF\r\n	كتب الكاتب هارون\r\n كتب الكاتب هارون يحيى\r\n ...	[<p>الروايات والكتب العربية تعتبر من الروابط ب...

In [31]:
```
#cleaning our data from any thing except arabic letters
pattern="[^ي-ء ]"
data["BookTitle"].replace(pattern,' ', regex=True,inplace=True)
data["Author"].replace(pattern,' ', regex=True,inplace=True)
data["Class"].replace(pattern,' ', regex=True,inplace=True)
data["BookSummary"].replace(pattern,' ', regex=True,inplace=True)
```

In [32]:
```
data["BookTitle"][0]
```

Out[32]:
```
'  كتاب التضحية عند الحيوان  '
```

In [33]:
```
#handle the start and end spaces
data["BookTitle"]=data["BookTitle"].apply(lambda title:title.strip())
data["BookTitle"][0]
```

Out[33]:
```
'كتاب التضحية عند الحيوان'
```

In [34]:
```
data["Author"][0]
```

Out[34]:
```
'  كتب الكاتب هارون يحيى  كتب الكاتب هارون يحيى  '
```

In [35]:
```
data["Class"][0]
```

Out[35]:
```
'  تحميل كتب الأدب العربي  تحميل كتب متنوعة  '
```

In [36]:
```
data["BookSummary"][0]
```

Out[36]:
```
'  الروايات والكتب العربية تعتبر من الروابط بيننا وبين تاريخنا كعرب او بيننا وبين اللغة العربية  وبين تاريخنا واحداث هذا التاريخ  الروايات العربية الي الان ومع
تقدم العصور والتكنولوجيا لازالت قائمة وحاضرة  كتاب التضحية عند الحيوان للكاتب
هارون يحيى  تعتبر صفحات التضحية والتعاون والرحمة والمحبة والتكافل من الأخلاق الحسنة التي يسعى كل مجتمع لنشرها بين أفراده  وهذه الميزات تسبغ على الإن
سان الذي يتحلى بها المحبة والاحترام في البيئة التي يعيش فيها  ولعل كل إنسان يسعى إلى أن يحيا وسط مجتمع تنتشر فيه هذه الفضائل  وفي هذا الكتاب وقع ت
ناول هذه الخصال والفضائل بالدرس بيد أن الذي يتصف بالتضحية والإيثار هنا  والذي يقوم بأعمال التعاون والتكافل ويفيض شفقة ومحبة ورحمة ليس كائنا  بشريا  إن ال
ذي نتناوله بالبحث هنا وتجتمع فيه جميع هذه الفضائل هو الحيوان  فثمة حيوانات تلقى نفسها إلى الهلاك والخطر في مواجهة أعدائها من أجل حماية حياة صغارها  وثمة ط
يور تعرض نفسها للخطر أيضا  من أجل حماية فراخها  وفي مناطق القطب المتجمد تمكث طيور البطريق لأشهر دون أكل ودون حركة حتى توفر الحرارة والدفء لفراخها
ونجد كذلك حيوانات تتحمل مسؤولية رعاية من انقطعت به السبل من صغار الحيوانات وبقي يتيما  بلا أم  ونجد من أمثال هذه النماذج الكثير الكثير  كتاب التضحية عند
لحيوان  عدنان أوكطار المعروف باسم هارون يحيى هو كاتب وباحث تركي مسلم  ولد في أنقرة عام  م وعاش فيها حتى عام  م عندما انتق
ل لإسطنبول حيث التحق بكلية الفنون الجميلة في جامعة المعمار سنان  وخلال سنواته الجامعية  قام ببحوث مفصلة في الفلسفة المادية والايديولوجية السائدة التي تحيط به  ق
ام بإنشاء مؤسسة البحث العلمي في تركيا  تركز كتاباته على تفنيد وتكذيب نظريات التطور والارتقاء والنشوء وبيان تناقضها حسب رأيه  كما يركز في كتاباته أيضا على م
وضوعات الماسونية والصهيونية والإلحاد  له أكثر من مئة كتاب حول قيم وأخلاقيات القرآن وحول مواضيع ايمانية عديدة ومختلفة وترجمت إلى العديد من اللغات ال
عالمية واسمه القلمي هارون يحيى  ومن أشهر كتبه كتاب  أطلس الخلق  الذي يقع في  صفحة ويتحدث عن رفض نظرية النشوء والارتقاء لداروين  ومن أشهر م
ؤلفاته  أطلس الخلق  نهاية الدارونية  معجزة الذر ة  الإعجاز في خلق النباتات  خلق الكون  المفاهيم الأساسية في
القرآن  هل فكرت في الحقيقة  الكوارث التي جلبتها الداروينية للإنسانية  معجزة النمل  معجزة النحل  معجزة الجهاز المناعي
استعمل عدنان أوكطار الاسم المستعار  جاويد يالجن  في بعض كتبه  ولكن القسم الأكبر منها نشره بالاسم المستعار هارون يحيى  وهذا الاسم المستعار  يتكون من اسمي
نبيين في إشارة إلى ذكرى النبي هارون والنبي يحيى الل ذي ن حاربا أفكار الإلحاد  كتاب التضحية عند الحيوان
ترون بأنفسكم نماذج من السلوك عند بعض الحيوانات يثير الاستغراب والدهشة ويبعث على التأمل  ذلك أنه سلوك ينطق بالحكمة والعقل ويبين عن وعي خارق وبعد نظر ع
ميق  وعند قراءة هذه النماذج سوف يقفز إلى أذهانكم السؤال التالي  كيف يمكن لحيوان أن يعرف كل هذا  كيف يمكنه أن يفكر في كل هذا  وسوف تجدون الإ
جابة على هذه التساؤلات في الكتاب الذي بين أيديكم  إن الذي خلقها جميعها هو الله الرحمن الرحيم  وهو الذي يلهمها طريقها فتسير فيه طائعة  فكل ما في السما
وات وما في الأرض طوع أمره سبحانه وتعالى  يمكنك أيضا قراءة وتحميل روايات عربي من خلال مكتبتكم
المكتبة العربية  مثل  '
```

In [37]:
```
#remove all this unimportant spaces
space_patern="[ ]+"
data["BookSummary"].replace(space_patern,' ', regex=True,inplace=True)
```

In [38]:
```
data["BookSummary"][0]
```

Out[38]:
```
'  الروايات والكتب العربية تعتبر من الروابط بيننا وبين تاريخنا كعرب او بيننا وبين اللغة العربية وبين تاريخنا واحداث هذا التاريخ الروايات العربية الي الان ومع تقدم الع
صور والتكنولوجيا لازالت قائمة وحاضرة كتاب التضحية عند الحيوان للكاتب هارون يحيى تعتبر صفحات التضحية والتعاون والرحمة والمحبة والتكافل من الأخلاق الحسنة الت
ي يسعى كل مجتمع لنشرها بين أفراده وهذه الميزات تسبغ على الإنسان الذي يتحلى بها المحبة والاحترام في البيئة التي يعيش فيها ولعل كل إنسان يسعى إلى أن يحيا وس
ط مجتمع تنتشر فيه هذه الفضائل وفي هذا الكتاب وقع تناول هذه الخصال والفضائل بالدرس بيد أن الذي يتصف بالتضحية والإيثار هنا والذي يقوم بأعمال التعاون والتكافل
ويفيض شفقة ومحبة ورحمة ليس كائنا بشريا إن الذي نتناوله بالبحث هنا وتجتمع فيه جميع هذه الفضائل هو الحيوان فثمة حيوانات تلقى نفسها إلى الهلاك والخطر في مواجه
ة أعدائها من أجل حماية حياة صغارها وثمة طيور تعرض نفسها للخطر أيضا من أجل حماية فراخها وفي مناطق القطب المتجمد تمكث طيور البطريق لأشهر دون أكل ودو
ن حركة حتى توفر الحرارة والدفء لفراخها ونجد كذلك حيوانات تتحمل مسؤولية رعاية من انقطعت به السبل من صغار الحيوانات وبقي يتيما بلا أم ونجد من أمثال هذه ال
نماذج الكثير الكثير كتاب التضحية عند الحيوان عدنان أوكطار المعروف باسم هارون يحيى هو كاتب وباحث تركي مسلم ولد في أنقرة عام م وعاش فيها حتى عام م عندما
انتقل لإسطنبول حيث التحق بكلية الفنون الجميلة في جامعة المعمار سنان وخلال سنواته الجامعية قام ببحوث مفصلة في الفلسفة المادية والايديولوجية السائدة التي تحيط به قام
```

بإنشاء مؤسسة البحث العلمي في تركيا تركز كتاباته على تفنيد وتكذيب نظريات التطور والارتقاء والنشوء وبيان تناقضها حسب رأيه كما يركز في كتاباته أيضا على موضوع ات الماسونية والصهيونية والإلحاد له أكثر من مئة كتاب حول قيم وأخلاقيات القرآن وحول مواضيع ايمانية عديدة ومختلفة وترجمت إلى العديد من اللغات العالمية واسمه القلم ي هارون يحيى ومن أشهر كتبه كتاب أطلس الخلق الذي يقع في صفحة ويتحدث عن رفض نظرية النشوء والارتقاء لداروين ومن أشهر مؤلفاته أطلس الخلق نهاية الداروينية معجزة الذر ة الإعجاز في خلق النباتات خلق الكون المفاهيم الأساسية في القرآن هل فكرت في الحقيقة الكوارث التي جلبتها الداروينية للإنسانية معجزة النمل معجزة النحل مع جزة الجهاز المناعي استعمل عدنان أوكطار الاسم المستعار جاويد يالجن في بعض كتبه ولكن القسم الأكبر منها نشره بالاسم المستعار هارون يحيى وهذا الاسم المستعار يتكو ن من اسمي نبيين في إشارة إلى ذكرى النبي هارون والنبي يحيى الل ذي ن حاربا أفكار الإلحاد كتاب التضحية عند الحيوان ترون بأنفسكم نماذج من السلوك عند بعض ال حيوانات يثير الاستغراب والدهشة ويبعث على التأمل ذلك أنه سلوك ينطق بالحكمة والعقل ويبين عن وعي خارق وبعد نظر عميق وعند قراءة هذه النماذج سوف يقفز إلى أذ هانكم السؤال التالي كيف يمكن لحيوان أن يعرف كل هذا كيف يمكنه أن يفكر في كل هذا وسوف تجدون الإجابة على هذه التساؤلات في الكتاب الذي بين أيديكم إن الذي خ لقها جميعها هو الله الرحمن الرحيم وهو الذي يلهمها طريقها فتسير فيه طائعة فكل ما في السماوات وما في الأرض طوع أمره سبحانه وتعالى يمكنك أيضا قراءة وتحميل ر وايات عربي من خلال مكتبتكم المكتبة العربية مثل '

In [39]:
```
#remove redundant words unusefal
data["BookTitle"].replace(r"كتاب","",regex=True,inplace=True)
data["Author"].replace(r"الكاتب","",regex=True,inplace=True)
data["Class"].replace(r"كتب","",regex=True,inplace=True)
```

In [40]:
```
data[["BookTitle","Author","BookSummary","Class"]].head()
```

Out[40]:

	BookTitle	Author	BookSummary	Class
0	التضحية عند الحيوان	كتب هارون يحيى كتب هارون يحيى	...الروايات والكتب العربية تعتبر من الروابط بينن	تحميل الأدب العربي تحميل متنوعة
1	معجزة الذرة	كتب هارون يحيى كتب هارون يحيى	...الروايات والكتب العربية تعتبر من الروابط بينن	تحميل تطوير الذات تحميل متنوعة
2	لا تتجاهل	كتب هارون يحيى كتب هارون يحيى	...الروايات والكتب العربية تعتبر من الروابط بينن	تحميل الأدب العربي
3	الحياة في سبيل الله	كتب هارون يحيى كتب هارون يحيى	...الروايات والكتب العربية تعتبر من الروابط بينن	تحميل اسلامية
4	العظمة في كل مكان	كتب هارون يحيى كتب هارون يحيى	...الروايات والكتب العربية تعتبر من الروابط بينن	تحميل الأدب العربي

In [41]:
```
#cleaning from redundant words unusefal
data["Class"]=data["Class"].apply(lambda text:text.replace("1,"","تحميل"))
data["Author"]=data["Author"].apply(lambda text:text.replace("1,"","كتب"))
#split Classes into multiclass & split Authers too
data["Class"]=data["Class"].apply(lambda row:row.split("تحميل"))
data["Author"]=data["Author"].apply(lambda row:row.split("كتب"))
```

In [42]:
```
data[["BookTitle","Author","BookSummary","Class"]].head()
```

Out[42]:

	BookTitle	Author	BookSummary	Class
0	التضحية عند الحيوان	[هارون يحيى , هارون يحيى]	...الروايات والكتب العربية تعتبر من الروابط بينن	[الأدب العربي , متنوعة]
1	معجزة الذرة	[هارون يحيى , هارون يحيى]	...الروايات والكتب العربية تعتبر من الروابط بينن	[تطوير الذات , متنوعة]
2	لا تتجاهل	[هارون يحيى , هارون يحيى]	...الروايات والكتب العربية تعتبر من الروابط بينن	[الأدب العربي]
3	الحياة في سبيل الله	[هارون يحيى , هارون يحيى]	...الروايات والكتب العربية تعتبر من الروابط بينن	[اسلامية]
4	العظمة في كل مكان	[هارون يحيى , هارون يحيى]	...الروايات والكتب العربية تعتبر من الروابط بينن	[الأدب العربي]

In [43]:
```
data["Author"][0]
```

Out[43]:
```
['    هارون يحيى  ' ,' هارون يحيى  ']
```

In [44]:
```
#handle the start and end spaces
data["Author"]=data["Author"].apply(lambda Authorlist:[Author.strip() for Author in Authorlist])
data["Author"][0]
```

Out[44]:
```
['هارون يحيى' ,' هارون يحيى']
```

In [45]:
```
data["Class"][0]
```

Out[45]:
```
['    متنوعة  ' ,' الأدب العربي  ']
```

In [46]:
```
#handle the start and end spaces
```

```
data["Class"]= data["Class"].apply(lambda classlist:[myclass.strip() for myclass in classlist])
data["Class"][0]
```

Out[46]:
```
['الأدب العربي', 'متنوعة']
```

In [47]:
```
#delete all dublicated values
data["Author"]=data["Author"].apply(lambda Authorlist:list(set(Authorlist)))
data["Class"] = data["Class"].apply(lambda classlist: (list(set(classlist))))
```

In [48]:
```
data.head(3)
```

Out[48]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL
0	التضحية عند الحيوان	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بين...	[الأدب العربي, متنوعة]	عربي	160 صفحة	عدنان أوكطار	5 ميغابايت	pdf	https://www.arab-books.com/books/%d9%83%d8%aa%...
1	معجزة الذرة	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بين...	[متنوعة, تطوير الذات]	عربي	136 صفحة	عدنان أوكطار	3.41 ميغابايت	pdf	https://www.arab-books.com/books/%d9%83%d8%aa%...
2	لا تتجاهل	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بين...	[الأدب العربي]	عربي	100 صفحة	عدنان أوكطار	3 ميغابايت	pdf	https://www.arab-books.com/books/%d9%83%d8%aa%...

In [49]:
```
data["format"].unique()
```

Out[49]:
```
array([' pdf', ' ', ' excel', ' word', ' PDF', ' 15.11', ' Pdf', ' حيب',
       ' P', ' PDF ZIP', ' PDf'], dtype=object)
```

In [50]:
```
data[data["format"]==' 15.11']
```

Out[50]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadU
3936	أحكام الأضحية والزكاة	[الشيخ محمد بن صالح العثيمين]	كتاب زاد الداعية إلى الله للشيخ العلامة محمد ...	[العقيدة]	العربية	34 صفحة	دار الثريا	15.11 ميغابايت	15.11	https://www.ara books.com/books/%d9%83%d8%aa

In [51]:
```
#cleaning
PDF=[" pdf"," PDF"," Pdf"," P"," PDF ZIP"," PDf"," 15.11 "," ","حيب"]
data["format"].replace(PDF,"PDF",inplace=True)
```

In [52]:
```
data["format"].unique()
```

Out[52]:
```
array(['PDF', ' excel', ' word'], dtype=object)
```

In [53]:
```
data["Language"].unique()
```

Out[53]:
```
array(['عربي', ' ', ' إنجليزي', ' العربية', ' العربيه', ' العربية ',
       ' الإنجليزية ', 'اللغة العربية', ' English', ' الانجليزية',
       ' الألمانية', ' عربية - المانية', ' العربية - الألمانية',
       'العربية - الالمانية', ' اللعة العربية', ' اللغه العربية',
       'اللغه العربيه', ' العربية', ' العربيية', ' اللغة العربية',
       ' العربيية'], dtype=object)
```

In [54]:
```
Arabic=[' العربية - الألمانية', ' العربية - الالمانية', ' اللعة العربية', ' اللغه العربية', ' العربية', ' العربيية', ' اللغة العربية', ' العربيية']
English=[' الإنجليزية ', 'إنجليزي', ' English', 'الانجليزية ']
German=' الألمانية'
```

In [55]:
```
#cleaning
data["Language"].replace(Arabic,"اللغة العربية",inplace=True)
data["Language"].replace(English,"اللغة الانجليزية",inplace=True)
```

```
data["Language"].replace(German,"اللغة الألمانية",inplace=True)
data["Language"].unique()
```

Out[55]:
```
array(['اللغة العربية', 'اللغة الانجليزية', 'اللغة الألمانية'],
      dtype=object)
```

In [56]:
```
data[["BookTitle","Author","BookSummary","Class"]].head()
```

Out[56]:

	BookTitle	Author	BookSummary	Class
0	التضحية عند الحيوان	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بين...	[الأدب العربي, متنوعة]
1	معجزة الذرة	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بين...	[متنوعة, تطوير الذات]
2	لا تتجاهل	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بين...	[الأدب العربي]
3	الحياة في سبيل الله	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بين...	[اسلامية]
4	العظمة في كل مكان	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بين...	[الأدب العربي]

In [63]:
```
Authors=[]
for Author in data["Author"]:
    if len(Author)==1:
        Authors.append(Author)
Authors[10:15]
```

Out[63]:
```
[['محمود سالم'],
 ['محمود سالم'],
 ['محمود سالم'],
 ['هارون يحيى'],
 ['هارون يحيى']]
```

In [64]:
```
Authors=[]
for Author in data["Author"]:
    if len(Author)>1:
        Authors.append(Author)
Authors[10:15]
```

Out[64]:
```
[['', 'الدكتور سليم حسن'],
 ['', 'الدكتور سليم حسن'],
 ['', 'الدكتور سليم حسن'],
 ['', 'الدكتور سليم حسن'],
 ['', 'الدكتور سليم حسن']]
```

In [65]:
```
#cleaning the column from fake values
data["Author"]= data["Author"].apply(lambda x:[item for item in x if item not in ['']])
```

In [66]:
```
Authors=[]
for Author in data["Author"]:
    if len(Author)>1:
        Authors.append(Author)
Authors[10:15]
```

Out[66]:
```
[['كوثر محمود عبد الرسول', 'محمد رياض'],
 ['ابن جزي الكلبي', 'ابن بطوطة'],
 ['جايا جرانت', 'أندرو جرانت'],
 ['تشستر إلتون', 'أدريان جوستيك'],
 ['بيتر سي كايرو', 'ديفيد إل دوتليتش', 'ستيفن إتش راينسميث']]
```

In [69]:
```
classs=[]
for bookclass in data["Class"]:
    if len(bookclass)==1:
        classs.append(bookclass)
classs[10:15]
```

Out[69]:
```
[['الأدب العربي'], ['اسلامية'], ['الفلسفة والمنطق'], ['اسلامية'], ['اسلامية']]
```

In [74]:
```
classs=[]
for bookclass in data["Class"]:
    if len(bookclass)>1:
        classs.append(bookclass)
classs[150:155]
```

Out[74]:
```
[['الأدب العربي', 'روايات عربية', 'الروايات العالمية المترجمة'],
 ['الأدب العربي', 'روايات عربية', 'الروايات العالمية المترجمة'],
 ['الأدب العربي', 'روايات عربية', 'الروايات العالمية المترجمة'],
 ['الأدب العربي', 'روايات عربية', 'الروايات العالمية المترجمة'],
 ['الأدب العربي', 'روايات عربية', 'الروايات العالمية المترجمة']]
```

In [75]:
```
#cleaning the column from fake values
data["Class"]= data["Class"].apply(lambda x:[item for item in x if item not in ['']])
```

In [76]:
```
data.head(3)
```

Out[76]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL
0	التضحية عند الحيوان	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بينن...	[الأدب العربي، متنوعة]	اللغة العربية	160 صفحة	عدنان أوكطار	5 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
1	معجزة الذرة	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بينن...	[متنوعة، تطوير الذات]	اللغة العربية	136 صفحة	عدنان أوكطار	3.41 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
2	لا تتجاهل	[هارون يحيى]	الروايات والكتب العربية تعتبر من الروابط بينن...	[الأدب العربي]	اللغة العربية	100 صفحة	عدنان أوكطار	3 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...

now our dataset is clean

lets save it in CSV

In [77]:
```
data.to_csv("CleanScrappedDataset.csv")
data.to_json("CleanScrappedDataset.json")
```

3-auto correct the scrapped data

3.1-Auto contextual_correct

In [78]:
```
from ar_corrector.corrector import Corrector
corr = Corrector()
```

In [79]:
```
dataset=pd.read_json("CleanScrappedDataset.json")
```

In [80]:
```
#lenth befor context correction
len(dataset["BookSummary"][0])
```

Out[80]:
```
2831
```

In [81]:
```
#len after context correction
len(dataset["BookSummary"].iloc[0:1].apply(lambda row:corr.contextual_correct(row))[0])
```

Out[81]:
```
2824
```

we notice the diferrent betwen two lengths

lets apply to all data

In [82]:
```
#context correction this take too time
dataset["BookSummary"]= dataset["BookSummary"].apply(lambda row:corr.contextual_correct(row))
```

Tokenization to make Auto correction on words

In [84]:
```
dataset['BookSummary'] = dataset.apply(lambda row: nltk.word_tokenize(row['BookSummary']), axis=1)
```

In [86]:
```
dataset["BookSummary"].iloc[0][:10]
```

Out[86]:
```
['الروايات',
 'والكتب',
 'العربية',
 'تعتبر',
 'من',
 'الروابط',
 'بيننا',
 'وبين',
 'تاريخنا',
 'كعرب']
```

3.2-Auto word correct the scrapped data

In [88]:
```
dataset.head(3)
```

Out[88]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL
0	التضحية عند الحيوان	هارون يحيى]	الروايات, والكتب, العربية, تعتبر, من, الروابط...	الأدب العربي, متنوعة]	اللغة العربية	160 صفحة	عدنان أوكطار	5 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
1	معجزة الذرة	هارون يحيى]	الروايات, والكتب, العربية, تعتبر, من, الروابط...	متنوعة, تطوير الذات]	اللغة العربية	136 صفحة	عدنان أوكطار	3.41 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
2	لا تتجاهل	هارون يحيى]	الروايات, والكتب, العربية, تعتبر, من, الروابط...	الأدب العربي]	اللغة العربية	100 صفحة	عدنان أوكطار	3 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...

In [89]:
```
check=lambda word : type(corr.spell_correct(word,1)) is type([])
get =lambda word : corr.spell_correct(word,1)[0][0]
```

In [90]:
```
#this take alot of time to correct all words
dataset["BookSummary"]=dataset["BookSummary"].apply(lambda row: [get(word) if check(word) else word for word in
```

Remove Aabic Stop words

In [92]:
```
import arabicstopwords.arabicstopwords as stp
stp.classed_stopwords_list()
```

Out[92]:
```
dict_keys(['حم', 'غداة', 'جنوب', 'ذواتا', 'حي', 'لازلنا', 'زمان', 'عوض', 'بنا', 'أجمع', 'ؤ', 'إيه', 'لدن', 'ها', 'ش', 'غرب',
'لازلتم', 'هج', 'هب', 'عسى', 'ل', 'آنفا', 'لازلتن', 'هن', 'هم', 'هل', 'هاؤم', 'ارتد', 'هي', 'هو', 'لازالتا', 'تلكما', 'وقت', 'أولئك',
'م', 'نحو', 'حسب', 'نحن', 'لئن', 'اللذين', 'أب', 'أخ', 'قبل', 'بدون', 'مادامتا', 'بئس', 'د', 'ذا', 'لعمر', 'هاهنا', 'كما', 'لستم', 'ل
ستن', 'حتى', 'لدى', 'ذه', 'ذي', 'ذو', 'أي', 'أو', 'أف', 'أن', 'أم', 'نعما', 'هيت', 'هيا', 'مابرح', 'حينما', 'هلا', 'إنما', 'جع
ل', 'كخ', 'ت', 'عما', 'بكن', 'بكم', 'مازلنا', 'غ', 'رويدك', 'دون', 'أولئك', 'كي', 'هؤلاء', 'لها', 'ي', 'مكانكما', 'كم', 'كل', 'م
افتئتن', 'ماانفككتما', 'لعل', 'إنا', 'لازال', 'متى', 'مابرحوا', 'خلال', 'مازلتن', 'مازلتم', 'ثنا', 'راح', 'هلم', 'لولا', 'مابرحتما', 'مافتئت',
'ماقتئتم', 'تان', 'مرة', 'أصلا', 'وشكان', 'كأنما', 'إياهما', 'أسفل', 'ص', 'أمامك', 'م', 'مادمتما', 'إياهم', 'ماانفكوا', 'استحال', 'إليك',
'أمسى', 'انبرى', 'مافتئتا', 'أنتم', 'لازالوا', 'أنتن', 'ليستا', 'ماانفككن', 'أوشك', 'طفق', 'لازالت', 'صه', 'لازالا', 'هناك', 'ليس', 'هاذي',
ن', 'مازالوا', 'إلا', 'سوف', 'ب', 'آمين', 'آناء', 'هاتين', 'ليت', 'و', 'أكثر', 'إلى', 'أثناء', 'ابن', 'بعدئذ', 'فلان', 'ليل', 'تين', 'أ
صبح', 'لكما', 'كأين', 'قد', 'ماانفككت', 'قط', 'لحظة', 'شمال', 'ثم', 'أجل', 'عليك', 'دونك', 'ذلكما', 'حما', 'أنت', 'التي', 'أوه', 'أن
ا', 'مثل', 'هذين', 'س', 'لازلتما', 'ك', 'أنى', 'حمو', 'حمي', 'هما', 'جير', 'أبدا', 'اللذان', 'عند', 'ظل', 'سرعان', 'خلف', 'تحو
ل', 'هاته', 'هاتي', 'حوالى', 'جلل', 'تينك', 'أولاء', 'آنذاك', 'بعد', 'حين', 'ذيت', 'أيان', 'قلما', 'مازالا', 'خ', 'إزاء', 'بعض', 'فيم
ا', 'لنا', 'مكانكن', 'تلكم', 'ذين', 'اللائي', 'تلقاء', 'خلا', 'حيث', 'مافتئوا', 'صار', 'أينما', 'يمين', 'رب', 'ماذا', 'لست', 'مافتئن', 'ذا
ن', 'إياه', 'ذاك', 'ة', 'ن', 'حذار', 'ماانفكتا', 'ع', 'مادمتم', 'مابرحنا', 'ى', 'ذات', 'واها', 'أما', 'جدا', 'أمد', 'اللواتي',
'أمس', 'لسن', 'معاذ', 'الذين', 'حيثما', 'تحت', 'غير', 'حاشا', 'مافتئ', 'مابرحتن', 'مابرحتم', 'أمام', 'ج', 'أضحى', 'اللتيا', 'هاتان',
'هكذا', 'مابرحتا', 'كلتا', 'بضع', 'انقلب', 'اللتين', 'ساء', 'لستما', 'إليكن', 'إليكم', 'غدا', 'حاي', 'كلما', 'تبدل', 'مكانكم', 'كذلك', 'شتا
ن', 'الآن', 'أولالك', 'إياي', 'ممن', 'إياك', 'كيت', 'بعدما', 'اخلولق', 'حرى', 'مادمنا', 'ط', 'ه', 'عاد', 'كيف', 'إياكم', 'إياكن', 'سب
```

```
ان', 'هذان', 'أقبل', 'الألاء', 'مما', 'ليسوا', 'لما', 'ضحوة', 'مابرحا', 'آها', 'ماداموا', 'آ', 'كاد', 'عل', 'عن', 'مافتئتما', 'ز', 'لازل
ت', 'بيد', 'ق', 'ظ', 'ماداما', 'كان', 'كلاهما', 'بين', 'هيهات', 'إذا', 'إياكما', 'الذي', 'بل', 'نخ', 'طاق', 'أبو', 'أبي', 'هذا', 'لاس
يما', 'هنا', 'ساءما', 'ذينك', 'آه', 'ساعة', 'ح', 'ابتدأ', 'صباح', 'ضمن', 'لكيلا', 'بمن', 'كيفما', 'سوى', 'إياها', 'تانك', 'أقل', 'إليكم
ا', 'لئلا', 'كذا', 'أض', 'بما', 'مانفكا', 'فيم', 'إياهن', 'فا', 'هذي', 'أبا', 'هذه', 'تجاه', 'شطر', 'حينئذ', 'أيضا', 'شرق', 'رجع',
'عامة', 'إذن', 'غروب', 'عندما', 'وراءك', 'إذ', 'ألا', 'لكي', 'مابرحت', 'لكن', 'لكم', 'يومئذ', 'لكنما', 'قبلما', 'بي', 'ذ', 'مانفككنا',
'بك', 'بن', 'به', 'اللتان', 'مانفككتم', 'تارة', 'بس', 'مابرحن', 'فقط', 'أنشأ', 'مانفككتن', 'بخ', 'مثلما', 'شبه', 'بها', 'مازلتما', 'فوق',
'شهر', 'مانفكت', 'إي', 'اللاتي', 'مازال', 'إن', 'كأن', 'كليهما', 'مافتئا', 'كليكما', 'شرع', 'ليسا', 'مكانك', 'ليست', 'وي', 'أعلى', 'لس
نا', 'ثمة', 'لات', 'حسبما', 'ث', 'طالما', 'ذوا', 'مع', 'مانفك', 'نفس', 'مذ', 'بكما', 'ذوي', 'ذوو', 'مه', 'من', 'حبذا', 'مادام',
'عدا', 'بينما', 'وراء', 'بات', 'عدس', 'وا', 'كلا', 'على', 'علق', 'مازلن', 'حار', 'وراءكن', 'مازالتا', 'وراءكم', 'مادمت', 'أخي', 'أخ
و', 'قام', 'ئ', 'إذما', 'أين', 'بطآن', 'الألى', 'ض', 'ذلكم', 'ذلكن', 'أية', 'نعم', 'أيا', 'هاك', 'ذواتي', 'مازلت', 'آنئذ', 'أخذ', 'ذل
ك', 'عندئذ', 'أخا', 'ويكأن', 'إيانا', 'كرب', 'لي', 'لو', 'لن', 'له', 'لم', 'لك', 'بلا', 'مادمن', 'ذانك', 'إما', 'هنالك', 'مادامت',
'ء', 'ريث', 'طق', 'جميع', 'حول', 'ر', 'مهما', 'تي', 'كأي', 'لوما', 'ته', 'ف', 'فو', 'في', 'أ', 'أول', 'مافتئنا', 'مساء', 'بماذ
'!', 'عين', 'لا', 'جنب', 'تلك', 'بله', 'منذ', 'أنتما', 'وراءكما', 'لازلن', 'ما', 'بلى'])
```

In [93]:
```
#remove stop words
dataset['BookSummary'] = dataset['BookSummary'].apply(lambda row: [word for word in row if not stp.is_stop(wor
```

In [94]:
```
dataset[['BookTitle','Author','Class',"BookSummary"]].head()
```

Out[94]:

	BookTitle	Author	Class	BookSummary
0	التضحية عند الحيوان	[هارون, يحيى]	[الأدب العربي, متنوعة]	[الروايات, والكتب, العربية, تعتبر, الروابط, تا...
1	معجزة الذرة	[هارون, يحيى]	[متنوعة, تطوير الذات]	[الروايات, والكتب, العربية, تعتبر, الروابط, تا...
2	لا تتجاهل	[هارون, يحيى]	[الأدب العربي]	[الروايات, والكتب, العربية, تعتبر, الروابط, تا...
3	الحياة في سبيل الله	[هارون, يحيى]	[اسلامية]	[الروايات, والكتب, العربية, تعتبر, الروابط, تا...
4	العظمة في كل مكان	[هارون, يحيى]	[الأدب العربي]	[الروايات, والكتب, العربية, تعتبر, الروابط, تا...

Stemming

In [95]:
```
from nltk.stem.isri import ISRIStemmer
st = ISRIStemmer()
```

In [96]:
```
dataset['BookSummary']=dataset['BookSummary'].apply(lambda x: [st.stem(item) for item in x])
```

In [97]:
```
dataset['BookSummary'].iloc[0][:10]
```

Out[97]:
```
['روي', 'كتب', 'عرب', 'عبر', 'ربط', 'ارخ', 'كعرب', 'او', 'لغة', 'عرب']
```

In [99]:
```
dataset.head(3)
```

Out[99]:

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL
0	التضحية عند الحيوان	[هارون, يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, ... ,كعرب, او, لغة]	[الأدب العربي, متنوعة]	اللغة العربية	160 صفحة	عدنان أوكطار	5 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
1	معجزة الذرة	[هارون, يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, ... ,كعرب, او, لغة]	[متنوعة, تطوير الذات]	اللغة العربية	136 صفحة	عدنان أوكطار	3.41 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
2	لا تتجاهل	[هارون, يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, ... ,كعرب, او, لغة]	[الأدب العربي]	اللغة العربية	100 صفحة	عدنان أوكطار	3 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...

4-Save your data in csv file "Data sheet"

with the following columns " Author - title - publishing year - link for pdf - etc. "

In [100…
```
dataset.to_csv("finalCleanedData.csv",header=True)
```

Save your data in json file to save our lists as it is

In [101…
```
dataset.to_json("finalCleanedData.csv")
```

5-Cluster this data for topic modeling

In [102…
```
mydata=pd.read_json("finalCleanedData.csv")
```

In [103…
```
dataset.head()
```

Out[103…

	BookTitle	Author	BookSummary	Class	Language	Pages	publisher	BookSize	format	DownloadURL
0	التضحية عند الحيوان	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]	[الأدب العربي, متنوعة]	اللغة العربية	160 صفحة	عدنان أوكطار	5 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
1	معجزة الذرة	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]	[متنوعة, تطوير الذات]	اللغة العربية	136 صفحة	عدنان أوكطار	3.41 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
2	لا تتجاهل	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]	[الأدب العربي]	اللغة العربية	100 صفحة	عدنان أوكطار	3 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
3	الحياة في سبيل الله	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]	[اسلامية]	اللغة العربية	100 صفحة	عدنان أوكطار	2.11 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...
4	العظمة في كل مكان	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]	[الأدب العربي]	اللغة العربية	140 صفحة	عدنان أوكطار	5.3 ميغابايت	PDF	https://www.arab-books.com/books/%d9%83%d8%aa%...

In [104…
```
#clustering by BookSummary base
mydata=mydata[["BookTitle","Author","BookSummary"]]
mydata.head()
```

Out[104…

	BookTitle	Author	BookSummary
0	التضحية عند الحيوان	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]
1	معجزة الذرة	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]
2	لا تتجاهل	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]
3	الحياة في سبيل الله	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]
4	العظمة في كل مكان	[هارون يحيى]	[روي, كتب, عرب, عبر, ربط, ارخ, كعرب, او, لغة ...]

In [105…
```
mydata["BookSummary"][0][:10]
```

Out[105…
```
['روي', 'كتب', 'عرب', 'عبر', 'ربط', 'ارخ', 'كعرب', 'او', 'لغة', 'عرب']
```

In [106…]
```
mydata["BookSummary"]=mydata["BookSummary"].apply(lambda row: ' '.join(row))
```

In [107…]
```
mydata["BookSummary"][0]
```

Out[107…]
```
'روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حدث ارخ روي عرب الي الن قدم عصر كنولوج قئم حضر كتب ضحي حيو كتب هار يحى عبر صفح ضح
ي تعا رحم حبة كفل خلق حسن سعى جمع نشر فرد ميز سبغ انس حلى حبة حرم بيئ يعش انس سعى يحا وسط جمع نشر فضل كتب وقع نول خصل فضل درس تص
ف ضحي يثر يقم أعمال تعا كفل يفض شفق محب رحم كائ بشر نول بحث جمع فضل حيو حيو لقى هلك خطر وجه عدئ حمي حية صغر طير عرض خطر حمي فرخ
نطق قطب تجمد مكث طير طرق أشهر حرك وفر حرر دفء فرخ نجد حيو حمل سؤل رعي قطع سبل صغر حيو وبق يتي نجد مثل اذج كثر كثر كتب ضحي حيو عدن
كطر عرف بسم هار يحى كتب بحث ترك سلم ولد نقر عام وعش عام نقل لإسطنبول تحق بكل فنن جمل جمع عمر سنن سنا جمع بحث فصل لسف ادة ايديولوجية سئد ت
حط إنشاء وسس بحث علم ترك ركز كتب فند كذب نظر تطر رقء نشء وبي نقض رأه ركز كتب وضع اسن صهو لحد مئة كتب قيم أخلاق قرأ مواضيع يمن عدد خلف
رجم عدد لغت علم وسم قلم هار يحى شهر كتب كتب طلس خلق يقع صفح حدث رفض نظر نشء رقء درو شهر وَلف طلس خلق نهي درن عجز لذر عجز خلق نبت
خلق كون مفاهيم سسي قرأ فكر حقق كرث جلب دارووينية سان عجز نمل عجز نحل عجز جهز نعي عمل عدن كطر اسم عار جاويد يلج كتب قسم كبر نشر اسم عار هار
يحى اسم عار يتك اسم نبي شرة ذكرى نبي هار نبي يحى الل حرب فكر لحد كتب ضحي حيو ترن أنفس اذج سلك حيو يثر غرب دهش بعث أمل سلك نطق حكم عقل
ويب وعي خرق نظر عمق قرء اذج قفز ذهن سؤل تلي يمكن لحو عرف يمك فكر تجد جبة سؤل كتب ايد خلق الل رحم رحم لهم طرق تسر طئع سمو ارض طوع امر
على يمك قرء حمل روي عرب كبت كتب عرب'
```

In [108…]
```
mydata.head()
```

Out[108…]

	BookTitle	Author	BookSummary
0	التضحية عند الحيوان	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...
1	معجزة الذرة	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...
2	لا تتجاهل	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...
3	الحياة في سبيل الله	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...
4	العظمة في كل مكان	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...

kmean clustering using CountVectorizer

In [109…]
```
from sklearn.feature_extraction.text import CountVectorizer
count_vect = CountVectorizer()
bow = count_vect.fit_transform(mydata['BookSummary'].values)
bow
```

Out[109…]
```
<5898x15872 sparse matrix of type '<class 'numpy.int64'>'
        with 663630 stored elements in Compressed Sparse Row format>
```

In [110…]
```
terms = count_vect.get_feature_names()
```

In [113…]
```
from sklearn.cluster import KMeans
model = KMeans(n_clusters = 20,init='k-means++',random_state=99)
model.fit(bow)
```

Out[113…]
```
KMeans(n_clusters=20, random_state=99)
```

In [114…]
```
labels = model.labels_
cluster_center=model.cluster_centers_
```

In [115…]
```
cluster_center
```

Out[115…]
```
array([[0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
```

```
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.00355872, 0.00355872, 0.00355872, ..., 0.        , 0.        ,
        0.        ]])
```

In [116…
```
terms[1:10]
```

Out[116…
```
['عات', 'عاتت', 'عاذ', 'عام', 'عامنـ', 'عة', 'عله', 'عمر', 'عمن']
```

In [117…
```
from sklearn import metrics
silhouette_score = metrics.silhouette_score(bow, labels, metric='euclidean')
```

In [118…
```
# which tells us that clusters are far away from each other
silhouette_score
```

Out[118…
```
0.18003436494177047
```

In [119…
```
# Giving Labels/assigning a cluster to each point/text
mydata['Bow Class Labal'] = model.labels_ # the last column you can see the label numebers
mydata.head(2)
```

Out[119…

	BookTitle	Author	BookSummary	Bow Class Labal
0	التضحية عند الحيوان	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...	9
1	معجزة الذرة	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...	9

In [120…
```
mydata.groupby(['Bow Class Labal'])['BookSummary'].count()
```

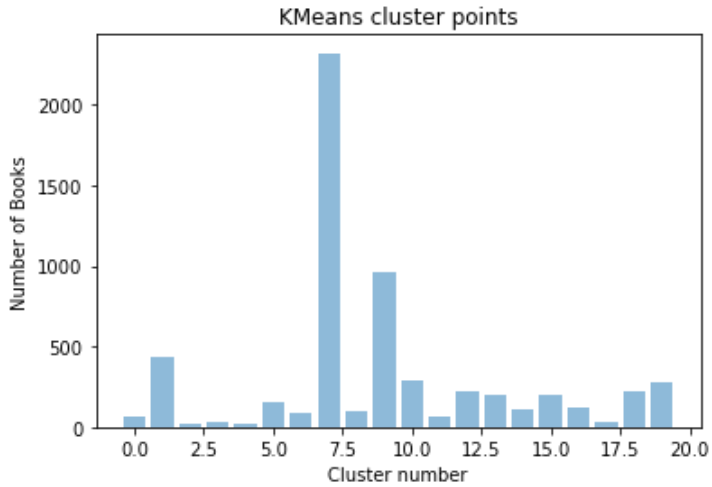
Out[120…
```
Bow Class Labal
0       64
1      430
2       17
3       33
4       16
5      151
6       84
7     2321
8       96
9      961
10     287
11      62
12     223
13     194
14     110
15     197
16     116
17      28
18     227
19     281
Name: BookSummary, dtype: int64
```

In [124…
```
#Refrence credit - to find the top 10 features of cluster centriod
print("Top terms per cluster:")
order_centroids = model.cluster_centers_.argsort()[:, ::-1]
terms = count_vect.get_feature_names()
for i in range(20):
    print("Cluster %d:" % i, end='')
    for ind in order_centroids[i, :10]:
        print(' %s-' % terms[ind], end='')
    print()
```

```
Top terms per cluster:
Cluster 0: كتب- انس- نصر- لسف- الي- حفظ- عرب- قال- ولد- ادب-
Cluster 1: روي- كتب- عرب- ادب- شهر- حمد- حفظ- بدي- خلد- وفق-
Cluster 2: ون- ين- ال- لل- ان- يم- اب- لر- ور- يه-
```

```
Cluster 3: -كتب- قطب- سيد- سلم- فكر- رحل- ادب- جمع- شهد- اخو
Cluster 4: -لل- ين- ون- يم- ان- ام- ال- ول- اب- يل
Cluster 5: -سلم- كتب- فكر- حمد- عمر- دكتور- علم- انه- جمع- حدث
Cluster 6: -علم- كتب- نفس- درس- عام- بشر- رود- عقل- عرف- جمع
Cluster 7: -كتب- علم- قرء- حمل- حمد- جمع- سلم- عام- يمك- عمل
Cluster 8: -كتب- عرب- عام- الل- ارخ- حمد- روي- رسل- قصص- سلم
Cluster 9: -كتب- عرب- ارخ- علم- عام- جمع- سلم- روي- حمد- درس
Cluster 10: -كتب- سلم- حدث- جمع- حمد- علم- الل- كثر- دمشق- عبد
Cluster 11: -حكم- كتب- سرح- وفق- حيت- شهر- ادب- دبي- عمل- كانت
Cluster 12: -يكي- عدد- ومك- طفل- لتر- ديز- بطط- كتب- تحر- جال
Cluster 13: -كتب- دين- علم- جلل- سيط- علماء- حفظ- طلب- فسر- كثر
Cluster 14: -كتب- شيخ- قرضاو- سلم- وسف- علم- حصل- تمك- سنة- عصر
Cluster 15: -رمض- يوم- شهر- سلم- مسك- بدأ- فطر- وقت- وفق- رؤة
Cluster 16: -كتب- غزل- حمد- علم- ولد- هـ- طوس- دين- فقه- صوف
Cluster 17: -ون- ين- لل- ال- ان- يم- اب- نت- لس- يه
Cluster 18: -كتب- حمد- سلم- علم- لغز- لسل- نظر- عدد- خنخ- نقل
Cluster 19: -الل- سلم- سعد- علم- كتب- ؤلف- عبد- دكتور- دين- قال
```

In [132…
```
# visually how points or reviews are distributed across 10 clusters
import matplotlib.pyplot as plt
plt.bar([x for x in range(20)], mydata.groupby(['Bow Class Labal'])['BookSummary'].count(), alpha = 0.5)
plt.title('KMeans cluster points')
plt.xlabel("Cluster number")
plt.ylabel("Number of Books")
plt.show()
```



kmean clustering using TF-IDF

In [128…
```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_vect = TfidfVectorizer()
tfidf = tfidf_vect.fit_transform(mydata['BookSummary'].values)
tfidf.shape
```

Out[128… (5898, 15872)

In [129…
```
from sklearn.cluster import KMeans
model_tf = KMeans(n_clusters = 20,random_state=99)
model_tf.fit(tfidf)
```

Out[129… KMeans(n_clusters=20, random_state=99)

In [130…
```
labels_tf = model_tf.labels_
cluster_center_tf=model_tf.cluster_centers_
```

In [131…
```
cluster_center_tf
```

Out[131…
```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
```

```
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

In [133…
```
# to understand what kind of words generated as columns by BOW
terms1 = tfidf_vect.get_feature_names()
```

In [134…
```
terms1[1:10]
```

Out[134…
```
['عات', 'عاتت', 'عاذ', 'عام', 'عامنـ', 'عة', 'عله', 'عمر', 'عمن']
```

In [135…
```
from sklearn import metrics
silhouette_score_tf = metrics.silhouette_score(tfidf, labels_tf, metric='euclidean')
```

In [136…
```
silhouette_score_tf
```

Out[136…
```
0.14474513124824395
```

In [137…
```
mydata1 = mydata
mydata1['TF-IDF class label'] = model_tf.labels_
mydata1.head(5)
```

Out[137…

	BookTitle	Author	BookSummary	Bow Class Labal	TF-IDF class label
0	التضحية عند الحيوان	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...	9	3
1	معجزة الذرة	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...	9	3
2	لا تتجاهل	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...	9	3
3	الحياة في سبيل الله	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...	9	3
4	العظمة في كل مكان	[هارون يحيى]	روي كتب عرب عبر ربط ارخ كعرب او لغة عرب ارخ حد...	9	3

In [138…
```
# How many points belong to each cluster ->

mydata1.groupby(['TF-IDF class label'])['BookSummary'].count()
```

Out[138…
```
TF-IDF class label
0      119
1      369
2      116
3     2332
4       79
5      293
6      224
7       62
8      288
9       96
10     194
11     131
12      98
13     100
14     141
15     147
16      36
17     627
18     248
19     198
Name: BookSummary, dtype: int64
```

In [139…
```
#Refrence credit - to find the top 10 features of cluster centriod
print("Top terms per cluster:")
order_centroids = model_tf.cluster_centers_.argsort()[:, ::-1]
```

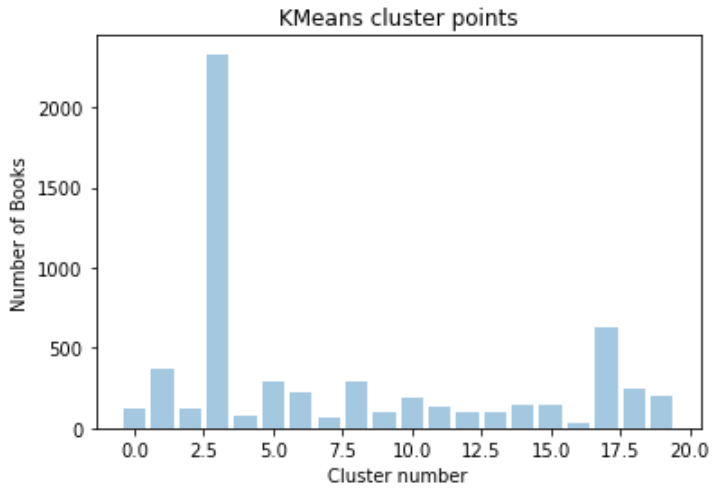
```
for i in range(20):
    print("Cluster %d:" % i, end='')
    for ind in order_centroids[i, :10]:
        print(' %s-' % terms1[ind], end='')
    print()
```

```
Top terms per cluster:
Cluster 0: دمشق- يمي- فتى- سنة- بلغ- عصر- تقي- شيخ- حنبل- جمع-
Cluster 1: روي- خلد- بدي- ادب- وفق- كتب- عرب- لسل- نجب- الك-
Cluster 2: غزل- طوس- صوف- هـ- حمد- اسماعيلي- شفع- علم- فقه- مات-
Cluster 3: كتب- علم- سلم- حمد- روي- الل- جمع- ارخ- قرأ- شيخ-
Cluster 4: كتب- فقي- راهيم- حسن- ندي- كبت- كند- ندق- لخص- عرب-
Cluster 5: دكتور- نبل- سلم- فكر- فرق- لطب- كتب- عمر- عام- انه-
Cluster 6: يكي- ديز- ومك- لتر- بطط- تحر- طفل- كة- رسم- جال-
Cluster 7: سرح- حكم- وفق- درا- حيت- دبي- بصف- كتب- ثوب- قلد-
Cluster 8: لغز- ختخ- حمد- سحل- برا- لسل- كتب- سلم- حرب- اسكندرية-
Cluster 9: الب- حدث- نبي- ضعف- صحح- حاديث- كتب- امم- نصر- حمد-
Cluster 10: جلل- سيط- دين- كتب- علم- فسر- علماء- طلب- حفظ- اعت-
Cluster 11: ين- ون- حنبل- لل- اب- ذهب- جهد- يم- وب- هـ-
Cluster 12: فى- وسع- سلم- أحكام- نبي- عجز- ليل- صلى- فقه- وقع-
Cluster 13: صلب- كتب- عقد- حرب- ارخ- زكر- عبس- عضو- وهى- حدث-
Cluster 14: يمك- حمل- قرء- وكل- قلب- ثعب- ياب- كتب- عمل- عجز-
Cluster 15: نفس- علم- كتب- انس- درس- نصر- رود- لسف- سلك- فرع-
Cluster 16: نجي- شرح- يدو- دكتور- ورق- تجد- تيب- فيسيولوجي- قنت- بشت-
Cluster 17: كتب- عرب- ارخ- روي- درس- عام- سلم- جمع- فكر- حمد-
Cluster 18: فسر- سور- كثر- دمشق- كتب- الل- كشك- سلم- هـ- بصر-
Cluster 19: رمض- مسك- يوم- فطر- شهر- رؤة- ايو- توقع- بدأ- وقت-
```

In [140…

```
# visually how points or reviews are distributed across 10 clusters

plt.bar([x for x in range(20)], mydata1.groupby(['TF-IDF class label'])['BookSummary'].count(), alpha = 0.4)
plt.title('KMeans cluster points')
plt.xlabel("Cluster number")
plt.ylabel("Number of Books")
plt.show()
```



In [ ]:

In [ ]: