Scrape Wikipedia Articles

Importing the packages

In [1]:
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import random
```

Create a function to clean up article text

In [2]:
```
from nltk.corpus import stopwords
import re
import string

def cleaning(text):
    # remove numbers
    text = re.sub("[0-9><]+", " ", text)
    # remove newlines
    text = re.sub(r"\n+", " ", text)
    # replace multiple spaces with one space
    text = re.sub(r"\s+", " ", text)
    # transfer text to lowercase
    text = text.lower()
    # remove punctuation
    text = re.sub(r'[^\w\s]','',text)

    return text
```

The main function for scraping the articles from Wikipedia

In [3]:
```
def scrapArticle(url):

    number_of_articles = 20 # set the number of articles
    list_links = []
    list_titles = []
    frame = []

    for i in np.arange(0, number_of_articles):

        r1 = requests.get(url)    # get the HTML content
        coverpage = r1.content    # coverpage variable contain the HTML content

        #create a soup in order to allow BeautifulSoup to work
        soup1 = BeautifulSoup(coverpage, 'html5lib')

        #locate the elements to find the links
        allLinks       = soup1.find(id="bodyContent").find_all("a") # we are looking for all links in the bo
        random.shuffle(allLinks)

        scrapedlink = 0
        for link in allLinks:
            # We are only interested in wiki articles
            if link['href'].find("/wiki/") == -1:
                continue

            if link['href'].find("Category:") != -1 or link['href'].find("http") != -1:
                continue

            title = link.get_text()    # get the title of the link
            list_titles.append(title)

            scrapedlink = link  # Use this link to scrape
            break


        FinalLink = "https://en.wikipedia.org" + scrapedlink['href']
        list_links.append(FinalLink)
        print(FinalLink)


        # Reading the content of article
        article = requests.get(FinalLink)
        article_content = article.content
        soup_article = BeautifulSoup(article_content, 'html5lib')
        body = soup_article.find_all('div', class_='mw-parser-output')
        x = body[0].find_all('p') # because articles are divided into paragraphs

        # collect all paragraphs
        list_paragraphs = []
        for p in np.arange(0, len(x)):
            paragraph = x[p].get_text()      # get the text of the paragraph
            paragraph = cleaning(paragraph)  # Make some cleaning
            list_paragraphs.append(paragraph)
            final_article = " ".join(list_paragraphs)

        frame.append([title,FinalLink,final_article])  # put all in a frame
    return frame
```

Testing using a link

In [4]:
```
# Link:
url = 'https://en.wikipedia.org/wiki/Category:Finance'
```

In [5]:
```
frame = scrapArticle(url)
```

```
https://en.wikipedia.org/wiki/Renting
https://en.wikipedia.org/wiki/Help:Categories
https://en.wikipedia.org/wiki/Non-financial_asset
https://en.wikipedia.org/wiki/Master_of_Applied_Finance
https://en.wikipedia.org/wiki/Finance
https://en.wikipedia.org/wiki/Capital_Markets_Union
https://en.wikipedia.org/wiki/Non-financial_asset
https://en.wikipedia.org/wiki/Capital_Markets_Union
https://en.wikipedia.org/wiki/P2F
https://en.wikipedia.org/wiki/Master_of_Applied_Finance
https://en.wikipedia.org/wiki/Designated_Professional_Body
https://en.wikipedia.org/wiki/Financial_stability
https://en.wikipedia.org/wiki/Real_bills_doctrine
https://en.wikipedia.org/wiki/Asset
https://en.wikipedia.org/wiki/Brattle_Group
https://en.wikipedia.org/wiki/P2F
https://en.wikipedia.org/wiki/Help:Categories
https://en.wikipedia.org/wiki/Trade_exchange
https://en.wikipedia.org/wiki/Shadow_Banking_in_China
https://en.wikipedia.org/wiki/Help:Category
```

In [6]:
```
data=pd.DataFrame(frame, columns=['title','link','article content'])
data
```

Out[6]:

	title	link	article content
0	Renting	https://en.wikipedia.org/wiki/Renting	renting also known as hiring or letting is an ...
1	category	https://en.wikipedia.org/wiki/Help:Categories	categories are used in wikipedia to link art...
2	Non-financial asset	https://en.wikipedia.org/wiki/Non-financial_asset	a nonfinancial asset is an asset that cannot b...
3	Master of Applied Finance	https://en.wikipedia.org/wiki/Master_of_Applie...	the master of finance is a masters degree aw...
4	Finance	https://en.wikipedia.org/wiki/Finance	finance is a term for matters regarding the ma...
5	Capital Markets Union	https://en.wikipedia.org/wiki/Capital_Markets_...	the capital markets union cmu is an ec...
6	Non-financial asset	https://en.wikipedia.org/wiki/Non-financial_asset	a nonfinancial asset is an asset that cannot b...
7	Capital Markets Union	https://en.wikipedia.org/wiki/Capital_Markets_...	the capital markets union cmu is an ec...
8	P2F	https://en.wikipedia.org/wiki/P2F	p f chinese 个人对金融机构 also known as p f model ...
9	Master of Applied Finance	https://en.wikipedia.org/wiki/Master_of_Applie...	the master of finance is a masters degree aw...
10	Designated Professional Body	https://en.wikipedia.org/wiki/Designated_Profe...	according to the uk financial conduct authorit...
11	Financial stability	https://en.wikipedia.org/wiki/Financial_stability	financial stability is a property of a financi...
12	Real bills doctrine	https://en.wikipedia.org/wiki/Real_bills_doctrine	the real bills doctrine says that as long as b...
13	Asset	https://en.wikipedia.org/wiki/Asset	in financial accounting an asset is any resour...
14	Brattle Group	https://en.wikipedia.org/wiki/Brattle_Group	the brattle group provides consulting services...
15	P2F	https://en.wikipedia.org/wiki/P2F	p f chinese 个人对金融机构 also known as p f model ...
16	category	https://en.wikipedia.org/wiki/Help:Categories	categories are used in wikipedia to link art...
17	Trade exchange	https://en.wikipedia.org/wiki/Trade_exchange	an association of businesses formed for the pu...
18	Shadow Banking in China	https://en.wikipedia.org/wiki/Shadow_Banking_i...	chinese shadow banking refers to underground f...
19	Categories	https://en.wikipedia.org/wiki/Help:Category	categories are intended to group together pa...

Example of article content

In [7]:
```
frame[0][2]
```

Out[7]:
```
'renting also known as hiring or letting is an agreement where a payment is made for the temporary use of a
good service or property owned by another a gross lease is when the tenant pays a flat rental amount and the
landlord pays for all property charges regularly incurred by the ownership an example of renting is equipmen
t rental renting can be an example of the sharing economy  there are many possible reasons for renting inste
ad of buying for example  shortterm rental of all sorts of products excluding real estate and holiday apartm
ents already represents an estimated  billion  billion annual market in europe and is expected to grow furth
er as the internet makes it easier to find specific items available for rent  according to a poll by yougov
of people looking to rent would go to the internet first to find what they need rising to  for those aged
  it has been widely reported that the financial crisis of  may have contributed to the rapid growth of onli
ne rental marketplaces such as erento as consumers are more likely to consider renting instead of buying in
times of financial hardship  environmental concerns fast depreciation of goods and a more transient workforc
e also mean that consumers are increasingly searching for rentals online   a us survey found  of renters pla
n to never buy a home   net income received or losses suffered by an investor from renting of one or two pro
perties is subject to idiosyncratic risk due to the numerous things that can happen to real property and var
iable behavior of tenants   there is typically an implied explicit or written rental agreement or contract i
nvolved to specify the terms of the rental which are regulated and managed under contract law  examples incl
ude letting out real estate real property for the purpose of housing tenure where the tenant rents a residen
ce to live in parking space for a vehicles storage space whole or portions of properties for business agricu
ltural institutional or government use or other reasons  when renting real estate the persons or party who l
ives in or occupies the real estate is often called a tenant paying rent to the owner of the property often
called a landlord or landlady the real estate rented may be all or part of almost any real estate such as an
apartment house building business offices or suite land farm or merely an inside or outside space to park a
vehicle or store things all under real estate law  the tenancy agreement for real estate is often called a l
ease and usually involves specific property rights in real property as opposed to chattels  in india the ren
tal income on property is taxed under the head income from house property a deduction of  is allowed from to
tal rent which is charged to tax  the time use of a chattel or other so called personal property is covered
under general contract law but the term lease also nowadays extends to long term rental contracts of more ex
pensive nonreal properties such as automobiles boats planes office equipment and so forth the distinction in
that case is long term versus short term rentals some nonreal properties commonly available for rent or leas
e are   in various degrees renting can involve buying services for various amounts of time such as staying i
n a hotel using a computer in an internet cafe or riding in a taxicab some forms of english use the term hir
ing for this activity  as seen from the examples some rented goods are used on the spot but usually they are
taken along to help guarantee that they are brought back one or more of the following applies  if the custom
er has a credit account with the rental company they may rent over several months or years and will receive
a recurring or continuation invoice each rental period until they return the equipment in this case deposits
are rarely required  in certain types of rental sometimes known as operated or wet rental the charge may be
calculated by the rental charge  timesheets of operators or drivers supplied by the rental company to operat
e the equipment this is particularly relevant for crane rental companies  sometimes the risk that the good i
s kept is reduced by it being a special model or having signs on it that cannot easily be removed making it
obvious that it is owned by the rental company this is especially effective for goods used in public places
but even when used at home it may help due to social control  persons and businesses that regularly rent goo
ds from a particular company generally have an account with that company which reduces the administrative pr
ocedure transaction costs on each occasion  signing out books from a library could be considered renting whe
n there is a fee per book however the term lending is more common  rental of personal property or real prope
rty for periods longer than a year which is governed by the signing of a lease is known as leasing leasing i
s usually used for highvalue capital equipment both in business and by consumers a lease in which the renter
benefits from an increase in value of the asset is known as a finance lease a leasing agreement which is not
a finance lease is known as an operating lease  a rental agreement may provide for the renter or lessee to b
ecome the owner of the asset at the end of the rental period usually at the renters option on payment of a n
ominal fee such arrangements may be known as '
```

Testing other links

In [8]:
```
print('Finance: ')
Finance_frame = scrapArticle("https://en.wikipedia.org/wiki/Category:Finance")
print('Mathematics: ')
Mathematics_frame = scrapArticle("https://en.wikipedia.org/wiki/Category:Mathematics")
print('Sports: ')
Sports_frame = scrapArticle("https://en.wikipedia.org/wiki/Category:Sports")
```

```
Finance:
https://en.wikipedia.org/wiki/Finance
https://en.wikipedia.org/wiki/Designated_Professional_Body
https://en.wikipedia.org/wiki/Designated_Professional_Body
https://en.wikipedia.org/wiki/Request_for_quote
https://en.wikipedia.org/wiki/Real_bills_doctrine
https://en.wikipedia.org/wiki/Asset
https://en.wikipedia.org/wiki/P2F
https://en.wikipedia.org/wiki/Brattle_Group
https://en.wikipedia.org/wiki/Shadow_Banking_in_China
https://en.wikipedia.org/wiki/JEL_classification_codes
https://en.wikipedia.org/wiki/Finance
https://en.wikipedia.org/wiki/Wikipedia:FAQ/Categorization#Why_might_a_category_list_not_be_up_to_date?
https://en.wikipedia.org/wiki/Help:Categories
https://en.wikipedia.org/wiki/JEL_classification_codes
https://en.wikipedia.org/wiki/Approved_Publication_Arrangement
https://en.wikipedia.org/wiki/Trade_exchange
https://en.wikipedia.org/wiki/Trade_exchange
https://en.wikipedia.org/wiki/Asset
https://en.wikipedia.org/wiki/Master_of_Applied_Finance
https://en.wikipedia.org/wiki/Renting
Mathematics:
https://en.wikipedia.org/wiki/Quota_rule
https://en.wikipedia.org/wiki/Composite_methods_for_structural_dynamics
https://en.wikipedia.org/wiki/Space
https://en.wikipedia.org/wiki/Colon_classification
https://en.wikipedia.org/wiki/Archives_of_American_Mathematics
https://en.wikipedia.org/wiki/Peano_kernel_theorem
https://en.wikipedia.org/wiki/Composite_methods_for_structural_dynamics
https://en.wikipedia.org/wiki/Pseudorandom_graph
https://en.wikipedia.org/wiki/Mathematics_in_Nepal
https://en.wikipedia.org/wiki/Space
https://en.wikipedia.org/wiki/File:Nuvola_apps_edu_mathematics_blue-p.svg
https://en.wikipedia.org/wiki/Analysis_of_Boolean_functions
https://en.wikipedia.org/wiki/Colon_classification
https://en.wikipedia.org/wiki/Mathematics
https://en.wikipedia.org/wiki/Archives_of_American_Mathematics
https://en.wikipedia.org/wiki/Help:Category
https://en.wikipedia.org/wiki/Pseudorandom_graph
https://en.wikipedia.org/wiki/Quantity
https://en.wikipedia.org/wiki/Library_classification
https://en.wikipedia.org/wiki/Change_(mathematics)
Sports:
https://en.wikipedia.org/wiki/Help:Category
https://en.wikipedia.org/wiki/Height_in_sports
https://en.wikipedia.org/wiki/Help:Category
https://en.wikipedia.org/wiki/Library_of_Congress_Classification
https://en.wikipedia.org/wiki/Portal:Sports
https://en.wikipedia.org/wiki/List_of_sports
https://en.wikipedia.org/wiki/Portal:Sports
https://en.wikipedia.org/wiki/Pacha_Nobin_Jomoh
https://en.wikipedia.org/wiki/Sport
https://en.wikipedia.org/wiki/Sport
https://en.wikipedia.org/wiki/Christine_Giampaoli_Zonca
https://en.wikipedia.org/wiki/List_of_professional_sports
https://en.wikipedia.org/wiki/Library_catalog
https://en.wikipedia.org/wiki/Universal_Decimal_Classification
https://en.wikipedia.org/wiki/Library_classification
https://en.wikipedia.org/wiki/House_of_Highlights
https://en.wikipedia.org/wiki/Library_of_Congress_Classification
https://en.wikipedia.org/wiki/Pre-game_ceremony
https://en.wikipedia.org/wiki/House_of_Highlights
https://en.wikipedia.org/wiki/House_of_Highlights
```

In [9]:
```
Mathematics_articles = pd.DataFrame(Mathematics_frame , columns=['title','link','article content'])
Mathematics_articles
```

Out[9]:

	title	link	article content
0	Quota rule	https://en.wikipedia.org/wiki/Quota_rule	in mathematics and political science the quota...
1	Composite methods for structural dynamics	https://en.wikipedia.org/wiki/Composite_method...	composite methods are an approach applied in s...
2	space	https://en.wikipedia.org/wiki/Space	space is the boundless threedimensional ex...
3	Colon	https://en.wikipedia.org/wiki/Colon_classifica...	colon classification cc is a system of library...
4	Archives of American Mathematics	https://en.wikipedia.org/wiki/Archives_of_Amer...	the archives of american mathematics located a...
5	Peano kernel theorem	https://en.wikipedia.org/wiki/Peano_kernel_the...	in numerical analysis the peano kernel theorem...
6	Composite methods for structural dynamics	https://en.wikipedia.org/wiki/Composite_method...	composite methods are an approach applied in s...
7	Pseudorandom graph	https://en.wikipedia.org/wiki/Pseudorandom_graph	in graph theory a graph is said to be a pseudo...
8	Mathematics in Nepal	https://en.wikipedia.org/wiki/Mathematics_in_N...	mathematics has been used in nepal for measure...
9	space	https://en.wikipedia.org/wiki/Space	space is the boundless threedimensional ex...
10		https://en.wikipedia.org/wiki/File:Nuvola_apps...	where x is greater than or equal to if the sq...
11	Analysis of Boolean functions	https://en.wikipedia.org/wiki/Analysis_of_Bool...	in mathematics and theoretical computer scienc...
12	Colon	https://en.wikipedia.org/wiki/Colon_classifica...	colon classification cc is a system of library...
13	Mathematics	https://en.wikipedia.org/wiki/Mathematics	mathematics from greek μάθημα máthēma know...
14	Archives of American Mathematics	https://en.wikipedia.org/wiki/Archives_of_Amer...	the archives of american mathematics located a...
15	Categories	https://en.wikipedia.org/wiki/Help:Category	categories are intended to group together pa...
16	Pseudorandom graph	https://en.wikipedia.org/wiki/Pseudorandom_graph	in graph theory a graph is said to be a pseudo...
17	quantity	https://en.wikipedia.org/wiki/Quantity	quantity is a property that can exist as a m...
18	classification	https://en.wikipedia.org/wiki/Library_classifi...	a library classification is a system of knowle...
19	change	https://en.wikipedia.org/wiki/Change_(mathemat...	mathematics from greek μάθημα máthēma know...

In [10]:
```
Sports_articles = pd.DataFrame(Sports_frame , columns=['title','link','article content'])
Sports_articles
```

Out[10]:

	title	link	article content
0	Categories	https://en.wikipedia.org/wiki/Help:Category	categories are intended to group together pa...
1	Height in sports	https://en.wikipedia.org/wiki/Height_in_sports	height can significantly influence success in ...
2	Categories	https://en.wikipedia.org/wiki/Help:Category	categories are intended to group together pa...
3	Library of Congress	https://en.wikipedia.org/wiki/Library_of_Congr...	the library of congress classification lcc is...
4	Sports portal	https://en.wikipedia.org/wiki/Portal:Sports	sport includes all forms of competitive physic...
5	List of sports	https://en.wikipedia.org/wiki/List_of_sports	the following is a list of sportsgames divided...
6	Portal:Sports	https://en.wikipedia.org/wiki/Portal:Sports	sport includes all forms of competitive physic...
7	Pacha Nobin Jomoh	https://en.wikipedia.org/wiki/Pacha_Nobin_Jomoh	pacha nobin jomoh born st november is an ind...
8	Sport	https://en.wikipedia.org/wiki/Sport	sport includes all forms of competitive ph...
9	sport	https://en.wikipedia.org/wiki/Sport	sport includes all forms of competitive ph...
10	Christine Giampaoli Zonca	https://en.wikipedia.org/wiki/Christine_Giampa...	christine giampaoli zonca born july is an ita...
11	List of professional sports	https://en.wikipedia.org/wiki/List_of_professi...	this is a list of professional sports that is...
12	Library cataloging	https://en.wikipedia.org/wiki/Library_catalog	a library catalog or library catalogue in br...
13	Universal Decimal	https://en.wikipedia.org/wiki/Universal_Decima...	the universal decimal classification udc is a ...
14	classification	https://en.wikipedia.org/wiki/Library_classifi...	a library classification is a system of knowle...
15	House of Highlights	https://en.wikipedia.org/wiki/House_of_Highlights	house of highlights often abbreviated as hoh i...
16	Library of Congress	https://en.wikipedia.org/wiki/Library_of_Congr...	the library of congress classification lcc is...
17	Pre-game ceremony	https://en.wikipedia.org/wiki/Pre-game_ceremony	a pregame ceremony or prematch ceremony is an ...
18	House of Highlights	https://en.wikipedia.org/wiki/House_of_Highlights	house of highlights often abbreviated as hoh i...
19	House of Highlights	https://en.wikipedia.org/wiki/House_of_Highlights	house of highlights often abbreviated as hoh i...

Links

- E-mail : zakaria.abbou199434@gmail.com
- GitHub : github.com/ZakariaAABBOU

- LinkedIn : linkedin.com/in/zakaria-aabbou/