Task Objective:

```
Your mission is to design and implement an Information Retrieval (IR) system capable of efficiently
retrieving relevant documents from a given dataset. Below are the key steps and details for this pro
ject:
```

Methodology

1. Dataset Selection and Preparation

2. Data Preprocessing

3. User Query Interface

4. Retrieval Algorithm

5. Query Processing

7. User Feedback (Optional)

6. Evaluation

8. Documentation and Presentation

1. Dataset Selection and Preparation:

- This Dataset is scraped from https://www.thenews.com.pk (https://www.thenews.com.pk) website. It has news articles from 2015 till date related to business and sports. It Contains the Heading of the particular Article, Its content and its date. The content also contains the place from where the statement or Article was published.

- you can download this dataset from kaggle(https://www.kaggle.com/datasets/asad1m9a9h6mood/news-articles (https://www.kaggle.com/datasets/asad1m9a9h6mood/news-articles))

About this csv file

Data set contains 4 columns

- Article : Text having the news article and the place where it was published from
- Heading : Text containing the heading of the news article.
- Date : Date when the article was published.
- NewsType : Type of Article i.e business or sports

Importing Libraries

In [64]:

```
## Importing necessary libraries
import pandas as pd
import numpy as np
import re
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords ,wordnet
from nltk.tokenize import word_tokenize
import matplotlib.pyplot as plt
```

In [65]:

```
## Import the dataframe using pandas
df = pd.read_csv("./dataset/Articles.csv", encoding='latin1')
df.head()
```

Out[65]:

	Article	Date	Heading	NewsType
0	KARACHI: The Sindh government has decided to b...	1/1/2015	sindh govt decides to cut public transport far...	business
1	HONG KONG: Asian markets started 2015 on an up...	1/2/2015	asia stocks up in new year trad	business
2	HONG KONG: Hong Kong shares opened 0.66 perce...	1/5/2015	hong kong stocks open 0.66 percent lower	business
3	HONG KONG: Asian markets tumbled Tuesday follo...	1/6/2015	asian stocks sink euro near nine year	business
4	NEW YORK: US oil prices Monday slipped below $...	1/6/2015	us oil prices slip below 50 a barr	business

In [66]:

```
## Take the rows only necessary
data = df.drop(["Date" ,"Heading"] ,axis = 1)
data.head()
```

Out[66]:

	Article	NewsType
0	KARACHI: The Sindh government has decided to b...	business
1	HONG KONG: Asian markets started 2015 on an up...	business
2	HONG KONG: Hong Kong shares opened 0.66 perce...	business
3	HONG KONG: Asian markets tumbled Tuesday follo...	business
4	NEW YORK: US oil prices Monday slipped below $...	business

2. Data Cleaning & Preprocessing

- Check for missing value
- Remove duplicates text
- Casing
- Removing puntuation

In [67]:

```
## Check for missing values
data.isna().sum()
```

Out[67]:

```
Article     0
NewsType    0
dtype: int64
```

In [68]:

```
## Check for duplicates
print("Duplicates ",data.duplicated().sum())

data.drop_duplicates(inplace = True) ## drop the duplicate

print("After dropping duplicates" , data.duplicated().sum())
```

```
Duplicates  108
After dropping duplicates 0
```

In [69]:

```
## Checking for duplicates in article Column
data.duplicated(subset = ["Article"]).sum()
data.head()
```

Out[69]:

	Article	NewsType
0	KARACHI: The Sindh government has decided to b...	business
1	HONG KONG: Asian markets started 2015 on an up...	business
2	HONG KONG: Hong Kong shares opened 0.66 perce...	business
3	HONG KONG: Asian markets tumbled Tuesday follo...	business
4	NEW YORK: US oil prices Monday slipped below $...	business

In [70]:

```
data.tail()
```

Out[70]:

	Article	NewsType
2669	strong>DUBAI: Dubai International Airport and ...	business
2670	strong>BEIJING: Former Prime Minister, Shaukat...	business
2671	strong>WASHINGTON: Uber has grounded its fleet...	business
2690	strong>BEIJING: The New Development Bank plans...	business
2691	strong>KARACHI: Karachi-based technology incub...	business

In [71]:

```
import nltk
nltk.download('stopwords')
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Barcha\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Barcha\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[71]:

```
True
```

In [72]:

```
## Create a preprocessing function
stops_word = set(stopwords.words("english")) ## Will contain stops words

def preprocessing(text):
    text = text.lower()
    text = re.sub(r'https\S+|www\S+https\S+', '', text, flags=re.MULTILINE)
    text = re.sub(r'\@\w+|\#', '', text)
    text = re.sub(r'[^\w\s\n]', '', text)
    text = re.sub(r'<br>|<strong>', '', text)

    lemitizer = WordNetLemmatizer()  ## this function converts the word to its base form
    words = word_tokenize(text) ## split the sentence into words/tokens
    lemitize_word = [lemitizer.lemmatize(word ,wordnet.VERB) for word in words]
    newArray = [stop_word for stop_word in lemitize_word if stop_word not in stops_word]

    return " ".join(newArray)
```

In [73]:

```
## Apply preprocessing
data["ArticleCleaned"] = data["Article"].apply(preprocessing)
```

In [74]:

```
# check the preprocessed data
data["ArticleCleaned"]
```

Out[74]:

```
0       karachi sindh government decide bring public t...
1       hong kong asian market start 2015 upswing limi...
2       hong kong hong kong share open 066 percent low...
3       hong kong asian market tumble tuesday follow p...
4       new york us oil price monday slip 50 barrel fi...
                              ...
2669    strongdubai dubai international airport flag c...
2670    strongbeijing former prime minister shaukat az...
2671    strongwashington uber ground fleet selfdriving...
2690    strongbeijing new development bank plan cofina...
2691    strongkarachi karachibased technology incubato...
Name: ArticleCleaned, Length: 2584, dtype: object
```

In [75]:

```
data["NewsType"].value_counts()
```

Out[75]:

```
sports      1408
business    1176
Name: NewsType, dtype: int64
```

3. User Query Interface

- create a query function and preprocess it by using preprocess function from above

In [76]:

```
## create a input query function

def query():
    query = input("Write the query(Text) :\n ")
    
    query1 = preprocessing(query)
    
    return query1
```

In [ ]:

```

```

4. Retrieval Algorithm

- Use cosine_simularity to find the text for the query and print the most likely news article for it

In [77]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

def algo():
    vectorizer = TfidfVectorizer()   ## cont

    x = data["ArticleCleaned"]

    x_vectorize = vectorizer.fit_transform(x)

    query111 = query()

    query_vectorizer = vectorizer.transform([query111])


    similarity = cosine_similarity(query_vectorizer ,x_vectorize)
    return similarity ,query11
```

Run below cell to query

In [81]:

```python
## Search the text for the most relevant match
similarity, query11 = algo()
max_text_index = similarity[0].argmax()

print()
print(f"\nThe index is: {max_text_index}\n")
print(data["Article"].iloc[max_text_index])
```

```
Write the query(Text) :
 KARACHI: The Sindh government has decided to


The index is: 871

strong>ISLAMABAD: Sindh Chief Minister Syed Murad Ali Shah has said that World Bank has played an important
role in the development of education, health, agriculture and infrastructure sectors in Sindh. "The urge for
development of Sindh has brought me here at the WB Country office."</strongThis he said while talking to WB
Country Director Mr Illang Patchamuthu at WB Country office where he had a luncheon-meeting with him today.T
he chief minister said that the total development portfolio of WB in Sindh was $1.14 billion which include E
ducation, health, irrigation, Agriculture, skill development and other sectors.Syed Murad Ali Shah discussed
Karachi mega projects which include infrastructure development of the city, water supply and sanitation proj
ects. "Sindh government needs financial and technical assistance for implementation of these projects," he t
old the World Bank country director.He urged the World Bank country head to send his team to Sindh to discus
s these projects with the team of provincial government so that necessary paper work and legal formalities c
ould be done well in time.Presently, two World Bank projects in education sector, a $66m special grant for m
issing facilities and anothere $400 m Sindh Education Reform Program, are in progress.The chief minister als
o discussed Sukkur Barrage rehabilitation and canal lining projects with World Bank country chief.The World
Bank Country Director assured the chief minister that he would consider all the proposals he had proposed on
priority basis. He thanked the chief minister for visiting his office.
```

First 5 query matches

In [82]:

```
ranked_indices = np.argsort(similarity[0])[::-1]

ranked_documents = [data["Article"].iloc[idx] for idx in ranked_indices]
```

In [83]:

```
for i ,news in enumerate(ranked_documents):
    print(i)
    print(news)
    print(100*"-")
    print("\n")
    if i == 4:
        break
```

```
0
strong>ISLAMABAD: Sindh Chief Minister Syed Murad Ali Shah has said that World Bank has played an important
role in the development of education, health, agriculture and infrastructure sectors in Sindh. "The urge for
development of Sindh has brought me here at the WB Country office."</strongThis he said while talking to WB
Country Director Mr Illang Patchamuthu at WB Country office where he had a luncheon-meeting with him today.T
he chief minister said that the total development portfolio of WB in Sindh was $1.14 billion which include E
ducation, health, irrigation, Agriculture, skill development and other sectors.Syed Murad Ali Shah discussed
Karachi mega projects which include infrastructure development of the city, water supply and sanitation proj
ects. "Sindh government needs financial and technical assistance for implementation of these projects," he t
old the World Bank country director.He urged the World Bank country head to send his team to Sindh to discus
s these projects with the team of provincial government so that necessary paper work and legal formalities c
ould be done well in time.Presently, two World Bank projects in education sector, a $66m special grant for m
issing facilities and anothere $400 m Sindh Education Reform Program, are in progress.The chief minister als
o discussed Sukkur Barrage rehabilitation and canal lining projects with World Bank country chief.The World
Bank Country Director assured the chief minister that he would consider all the proposals he had proposed on
priority basis. He thanked the chief minister for visiting his office.
----------------------------------------------------------------------------------------------------


1
KARACHI: The Sindh government has decided to bring down public transport fares by 7 per cent due to massive
reduction in petroleum product prices by the federal government, Geo News reported.Sources said reduction in
fares will be applicable on public transport, rickshaw, taxi and other means of traveling.Meanwhile, Karachi
Transport Ittehad (KTI) has refused to abide by the government decision.KTI President Irshad Bukhari said th
e commuters are charged the lowest fares in Karachi as compare to other parts of the country, adding that 80
pc vehicles run on Compressed Natural Gas (CNG). Bukhari said Karachi transporters will cut fares when decre
ase in CNG prices will be made.










----------------------------------------------------------------------------------------------------


2
KARACHI: Governor Sindh Dr. Ishrat-ul-Ebad Khan has said exporters of various goods have played a vital role
```

```
in economy of Pakistan and due to their efforts valuable addition is witnessed in national exchequer every y
ear.This he said while talking to a 9-members delegation of Rice Exporters Association of Pakistan (REAP) at
Governor House here on Monday.Principal Secretary to Governor Muhammad Hussain Syed was also present on the 
occasion.Dr. Ebad said that agriculture was the back-bone of Pakistans economy as majority population is en
gaged with this sector.Cotton, rice, sugarcane, mango, citrus fruits and other crops have a pivotal contribu
tion in Gross Domestic Product (GDP) of the country as they employ millions of people, he observed.Governor 
Sindh said that rice is an important part of exports of Pakistan and basmati rice of Pakistan is renowned fo
r its quality and taste worldwide.<br/> It also counts for sizeable amount of foreign exchange, he opined.On
pointing of dormant state of Rice Research Institute (RRI), Dokri district, Larkana, Governor Sindh assured 
that all concerned would be called soon to know the reasons behind its ineffectiveness.The RRI has a very im
portant role in producing new varieties of rice which are not only cost effective but also have visible cons
umption due to their quality, he added.On complaint of harassment from market committees, Governor Sindh ask
ed Principal Secretary to examine the matter and resolve the same in consultation of all stake holders.He sa
id that after improvement of law and order situation in Karachi, business community was engaged in their eco
nomic activities without any fear.Exporters would be provided all possible help and assistance to continue t
heir exports, he assured.Governor Sindh commended the idea of holding a Biryani Festival and said that it wo
uld help in increasing rice exports.The Chief Patron of  REAP Abdul Rahim Janoo informed Governor Sindh that
the Association has 1600 members from which 850 belong to Sindh.Pakistani rice is exported to 117 countries 
of the world including China, he said and added that Punjab produces Basmati while Sindh has Irri rice in ab
undance.He lauded the efforts of Governor Sindh in maintaining law &amp; order in Sindh and providing every 
possible facilities to business community.The delegation members included Senior Vice Chairman REAP Nauman A
hmed Shaikh, members managing Committee Javed Jilani, Inder Lal, Hamid Qureshi, Latif Paracha, Wajid Parach
a, Rauf Aziz and Secretary Altaf Hussain.
----------------------------------------------------------------------------------------------------



3
strong>KARACHI: Pakistan Tehreek-e-Insaf MPA Khurram Sherzaman has submitted a resolution against K-Electri
c's overbilling and load-shedding issues in the Sindh Assembly.</strongThe resolution asked Sindh government
to take action against the power utility for overcharging its consumers through overbilling. This act of th
e company has put extra financial burden on the masses, it said.Sindh government must take up these problems
with the federal government, the motion demanded.
----------------------------------------------------------------------------------------------------



4
strong>KARACHI: Sindh Chief Minister Syed Murad Ali Shah has directed Sindh Revenue Board (SRB) to take nece
ssary measures to bring in the services of insurances in sales tax net.</strongThis he said while presiding 
over a meeting with Chairman SRB Khalid Mahmood here at the CM House on Tuesday. Secretary Finance Secretary
Hassan Naqvi was also present.Murad Ali Shah discussed various issues relating to Sindh Revenue Board and Si
ndh Sales Tax on Services, including the issues of sales tax on the services of life insurance, health insur
ance, restaurants, etc.Appreciating the performance of Sindh Revenue Board, the chief minister emphasized th
e need of making relentless efforts to achieve the assigned target of Rs 78 billion for the current financia
```

```
l year. On this, Chairman SRB Khalid Mahmood said that the SRB has surpassed its six months target by 21 per
cent.
----------------------------------------------------------------------------------------------------
```

```
In [ ]:
```

Conclusion

- As you cn see from above that successful query and related documents are shown above.

```
In [ ]:
```