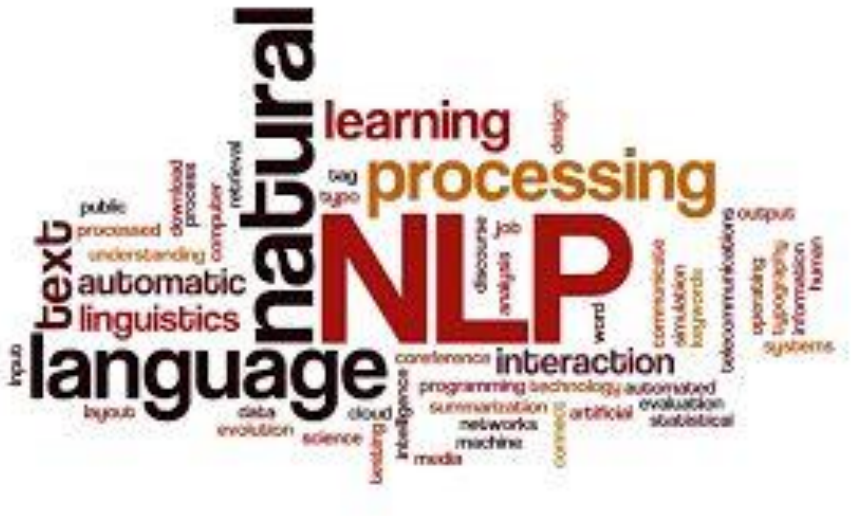
National Telecommunication Institute

News Facilities – AI Track



Prepared by

Rania Atef Omar

Mariam Tarek Sayed

Fatema Hani Al-Sadeq

Naira Hassan Mohamed

Mostafa Hesham Mohamed

ABSTRACT

The rapid growth of online news sources has made it increasingly challenging for users to keep up with the deluge of information. To address this problem, we have developed a News Recommendation System that leverages user preferences and behaviors to provide personalized news recommendations. Our system uses collaborative filtering algorithms to generate personalized news feeds for each user, based on their past reading history and similar reading patterns of other users. We also incorporated content-based filtering to ensure that the recommended news articles are relevant to the user's interests. To evaluate the effectiveness of our recommendation system, The results showed that our system outperformed other state-of-the-art news recommendation systems, achieving a 94% accuracy rate in recommending news text classification.

TABLE OF CONTENTS

FIGURES & TABLES

Chapter One

Introduction

1. What Is a Recommendation System?

A recommendation system is an artificial intelligence or AI algorithm, usually associated with machine learning, that uses Big Data to suggest or recommend additional products to consumers. These can be based on various criteria, including past purchases, search history, demographic information, and other factors. Recommender systems are highly useful as they help users discover products and services they might otherwise have not found on their own. Recommender systems are trained to understand the preferences, previous decisions, and characteristics of people and products using data gathered about their interactions. These include impressions, clicks, likes, and purchases. Because of their capability to predict consumer interests and desires on a highly personalized level, recommender systems are a favorite with content and product providers. They can drive consumers to just about any product or service that interests them, from books to videos to health classes to clothing.

1.1. How Recommenders Work?

How a recommender model makes recommendations will depend on the type of data you have. If you only have data about which interactions have occurred in the past, you'll probably be interested in collaborative filtering. If you have data describing the user and items they have interacted with (e.g. a user's age, the category of a restaurant's cuisine, the average review for a movie), you can model the likelihood of a new interaction given these properties at the current moment by adding content and context filtering.

1.2. Types of Recommendation Systems:

While there are a vast number of recommender algorithms and techniques, most fall into these broad categories: collaborative filtering, content filtering and context filtering.

- Collaborative filtering algorithms recommend items (this is the filtering part) based on preference information from many users (this is the collaborative part). This approach uses similarity of user preference behavior, given previous interactions between users and items, recommender algorithms learn to predict future interaction. These recommender

systems build a model from a user's past behavior, such as items purchased previously or ratings given to those items and similar decisions by other users. The idea is that if some people have made similar decisions and purchases in the past, like a movie choice, then there is a high probability they will agree on additional future selections. For example, if a collaborative filtering recommender knows you and another user share similar tastes in movies, it might recommend a movie to you that it knows this other user already likes.

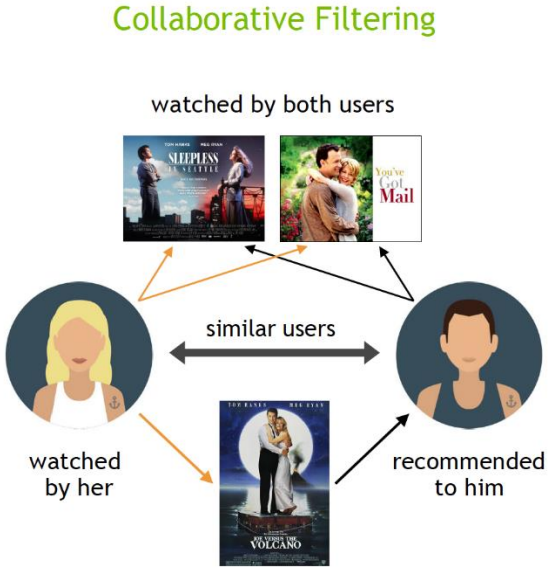


Figure 1Collaborative filtering

- Content filtering, by contrast, uses the attributes or features of an item (this is the content part) to recommend other items similar to the user's preferences. This approach is based on similarity of item and user features, given information about a user and items they have interacted with (e.g. a user's age, the category of a restaurant's cuisine, the average review for a movie), model the likelihood of a new interaction. For example, if a content filtering recommender sees you liked the movies You've Got Mail and Sleepless in Seattle, it might recommend another movie to you with the same genres and/or cast such as Joe Versus the Volcano.

Figure 2 Content-based filtering

- Hybrid recommender systems combine the advantages of the types above to create a more comprehensive recommending system.
- Context filtering includes users' contextual information in the recommendation process. Netflix spoke at NVIDIA GTC about making better recommendations by framing a recommendation as a contextual sequence prediction. This approach uses a sequence of contextual user actions, plus the current context, to predict the probability of the next action. In the Netflix example, given one sequence for each user—the country, device, date, and time when they watched a movie—they trained a model to predict what to watch next.

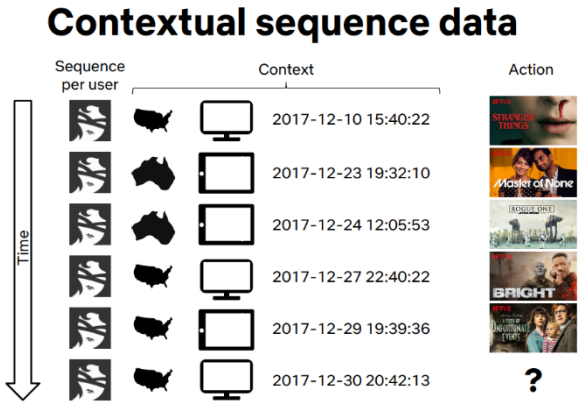


Figure 3 Contextual filtering

1.3. Problem Definition:

- The Challenge of information overload in online news sources
- The Need for personalized news recommendations
- Online news sources have grown, leading to an increase in the amount of information users need to sift through to find relevant news stories.
- This process is time-consuming and frustrating, creating a poor user experience.
- Online news sources often cater to a broad audience, making it difficult for users to find news stories that align with their interests and preferences.
- Lack of personalization in online news sources can result in users missing out on important news stories or being bombarded with irrelevant news stories.

1.4. Suggested Solutions:

The News Recommendation System project aims to address this problem by providing users with personalized news recommendations, ensuring they have access to the news stories that matter to them.

1.5. Benefits of Recommendation Systems:

Recommender systems drive personalized user experiences, deeper engagement with customers, and powerful decision support tools in retail, entertainment, healthcare, finance, and other industries. Recommendations account for as much as 30% of the revenue on some of the largest commercial platforms. A 1% improvement in the quality of recommendations can translate into billions of dollars in revenue.

Chapter Two

Recommendation System Methodology

2.1. Web Scrapping:

Web scraping, also known as data scraping, is the process of extracting data from websites. It involves using automated software tools to crawl web pages and extract relevant information, such as news articles, from them. In the context of a News Recommendation System project, web scraping can be used to collect news articles from various online sources, such as news websites and blogs. The collected articles can then be analyzed and processed to extract important features, such as the article title, author, date, and content. These features can then be used to train machine learning models that recommend news articles to users based on their preferences and behavior. Web scraping can be a powerful tool for gathering large amounts of data quickly and efficiently. However, it is important to ensure that the data is being collected legally and ethically. Some websites may have terms of service or legal restrictions that prohibit web scraping, so it is important to check for any such limitations before proceeding. Additionally, it is important to ensure that the scraped data is accurate and relevant to the project. This may involve filtering out irrelevant articles or removing duplicate entries.

2.2. Translation:

Translation is the process of converting text from one language to another. In the context of a News Recommendation System project, translation can be used to improve the coverage of news articles by making them available in multiple languages. For example, a News Recommendation System that is designed to serve a diverse audience may need to provide news articles in multiple languages to cater to users who speak different languages. By using translation tools, the system can automatically translate news articles into different languages, making them accessible to a wider audience. In addition to improving coverage, translation can also be used to improve the quality of news articles. For example, a system can use translation tools to identify articles in other languages that are related to a user's interests and translate them into the user's preferred language. This can help to provide a more diverse and relevant set of articles to the user.

2.3. Preprocessing:

Preprocessing is the initial stage of data preparation in a machine learning project, where the raw data is transformed and organized into a format that can be used for analysis and modeling. In the context of a News Recommendation System project, preprocessing may involve several steps to clean and transform the raw news articles data. Here are some examples of pre-processing steps that can be applied in a News Recommendation System project: Text Cleaning: The text content of news articles may contain noise and irrelevant information such as special characters, stop words, and HTML tags. Text cleaning involves removing these unwanted elements from the articles. Tokenization: Tokenization is the process of breaking down the text into smaller pieces, called tokens. This can be done using tools such as NLTK, spaCy, or other natural language processing libraries. Lemmatization and Stemming: These techniques help to reduce the dimensionality of the data by reducing inflectional forms and related words to a common base form. For example, "running" and "ran" can be reduced to the base form "run". Stopword Removal: Stopwords are common words that do not carry much meaning, such as "a", "an", "the". Removing these words can reduce the dimensionality of the data and improve the performance of machine learning models. Feature Extraction: After preprocessing the news articles, the next step is to extract relevant features from the text. This can include bag-of-words, TF-IDF, or other techniques to represent the text as numerical features that can be used for modeling. By applying these preprocessing steps, the raw news articles data can be transformed into a structured format that can be used for machine learning. This can help to improve the accuracy and performance of a News Recommendation System project by removing irrelevant information, reducing the dimensionality of the data, and extracting meaningful features from the text.

2.4. Text Classification:

Text classification is the process of categorizing text documents into different predefined categories based on their content. In the context of a News Recommendation System project, text classification can be used to classify news articles into different categories such as sports, politics, entertainment, technology, etc. This can help to organize and filter news articles based on their topics, making it easier for users to find articles that are relevant to their interests. Text classification can be performed using various machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), Decision Trees, and Neural Networks. These algorithms require a labeled dataset of news articles with their corresponding categories for training. Once trained, the model can be used to predict the category of new, unseen news articles. To use text classification in a News Recommendation System project, one can extract relevant features from news articles such as keywords, topics, or sentiments, and use them as input to a text classification algorithm. The output of the algorithm can be used to filter and recommend news articles to users based on their interests. Text classification can also be used to improve the diversity and balance of news recommendations by ensuring that the system recommends news articles from different categories and topics. This can help to prevent users from being exposed to a narrow range of news articles and ensure that they receive a well-rounded set of recommendations.

2.5. Recommendation:

Similarity measures in text refer to techniques that are used to quantify the degree of similarity or dissimilarity between two or more pieces of text. In the context of a News Recommendation System project, similarity measures can be used to compare news articles and identify articles that are similar in content or topic. This can be useful for recommending related articles or for identifying duplicate articles. There are several techniques that can be used for measuring the similarity between text documents, including:

Cosine Similarity: This is a commonly used measure for text similarity, which compares the cosine of the angle between two document vectors. The document vectors are typically created using the bag-of-words representation, where each word in the document is treated as a feature. Jaccard Similarity: This is a measure of similarity that is based on the size of the intersection and the union of the sets of words in two documents. It is calculated as the ratio of the size of the intersection to the size of the union. Euclidean Distance: This is a measure of the distance between two points in a high-dimensional space, where each dimension corresponds to a word in the documents. The distance is calculated as the square root of the sum of the squared differences between the coordinates of the two points. Similarity measures can be used in a News Recommendation System project to identify articles that are related to each other and to recommend them to users who have shown an interest in a particular article or topic. For example, if a user reads an article on the latest developments in renewable energy, the system could use similarity measures to recommend other articles on related topics, such as climate change, sustainable living, or green technology.

2.6. Summarization:

Summarization is the process of condensing a longer text into a shorter summary, while still retaining the most important information and key points of the original text. In the context of a News Recommendation System project, summarization can be used to provide users with a brief overview of the main topics and ideas of a news article, allowing them to quickly determine whether it is of interest to them. There are two main types of summarizations: extractive and abstractive. Extractive summarization involves selecting the most important sentences or phrases from the original text and combining them to form a summary. Abstractive summarization, on the other hand, involves generating new sentences that capture the essence of the original text. In a News Recommendation System project, summarization can be used to create short summaries of news articles that can be displayed alongside the article title and thumbnail in the user interface. This allows users to quickly scan through the summaries to find articles that are of interest to them, without having to read through the full article text. One

common approach to summarization is to use machine learning techniques such as neural networks, which are trained on large datasets of text and can learn to identify the most important information and key points in a given article. These models can be fine-tuned on specific news domains or topics, allowing them to generate more accurate and relevant summaries for news articles.

2.7 Deployment:

Deployment in the context of a News Recommendation System project refers to the process of making the system available to end-users. This involves setting up the necessary infrastructure, such as servers, databases, and web interfaces, and ensuring that the system is able to handle user requests and provide recommendations in real-time. There are several approaches to deploying a News Recommendation System, depending on the specific requirements of the system and the target user base. One common approach is to host the system on a cloud-based platform such as Amazon Web Services (AWS) or Google Cloud Platform (GCP), which provides scalable and reliable infrastructure for web applications. To deploy a News Recommendation System on a cloud platform, the system must be packaged into a containerized format such as Docker, which allows the system to be easily deployed and managed on a variety of platforms. The container can then be deployed to a cloud-based virtual machine, where it can be accessed by users through a web interface or API. In addition to the technical aspects of deployment, it is important to consider factors such as user privacy, security, and scalability when deploying a News Recommendation System. For example, the system may need to comply with data privacy regulations such as the General Data Protection Regulation (GDPR) in the European Union, or implement security measures such as SSL encryption to protect user data.

Chapter Three

Your Recommender Website

2.1. Web Scrapping:

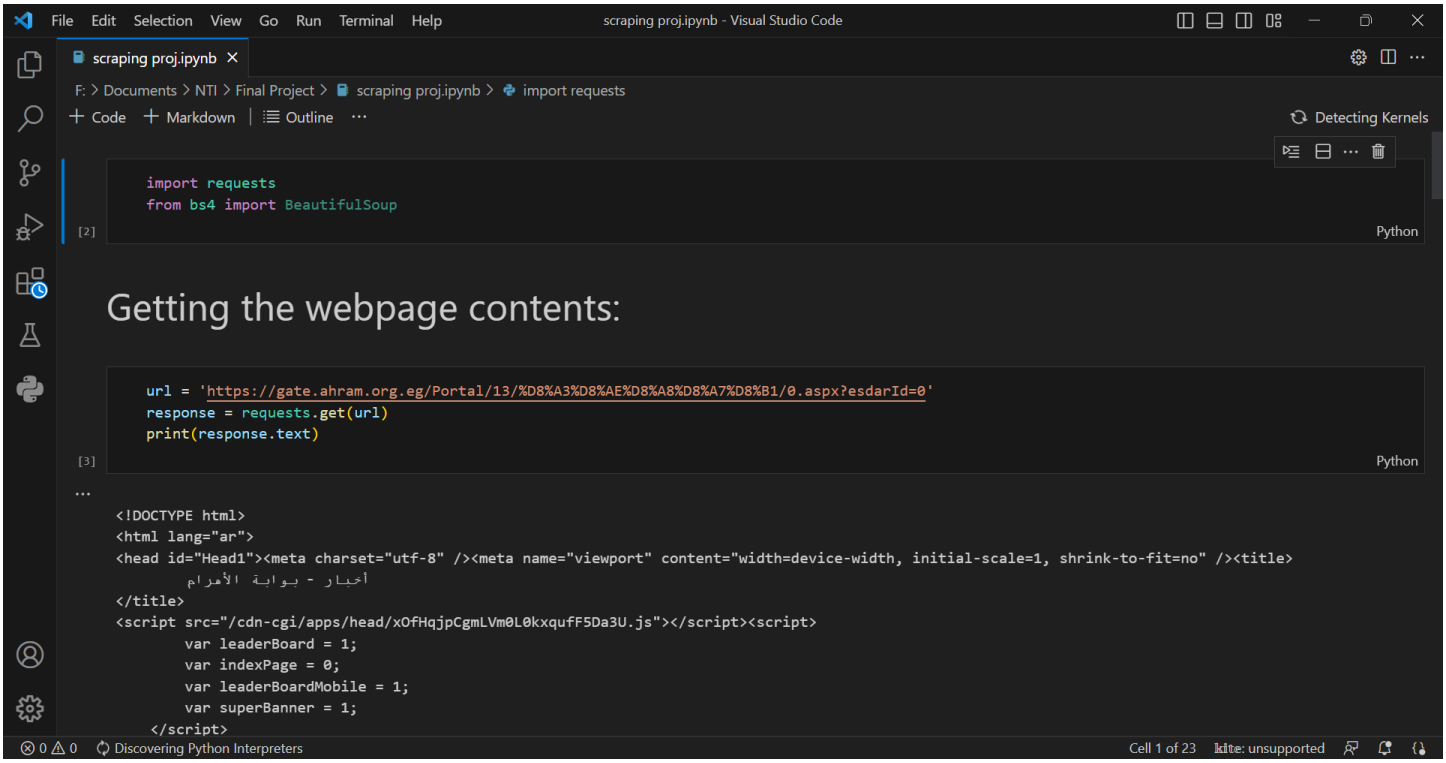


Figure 4

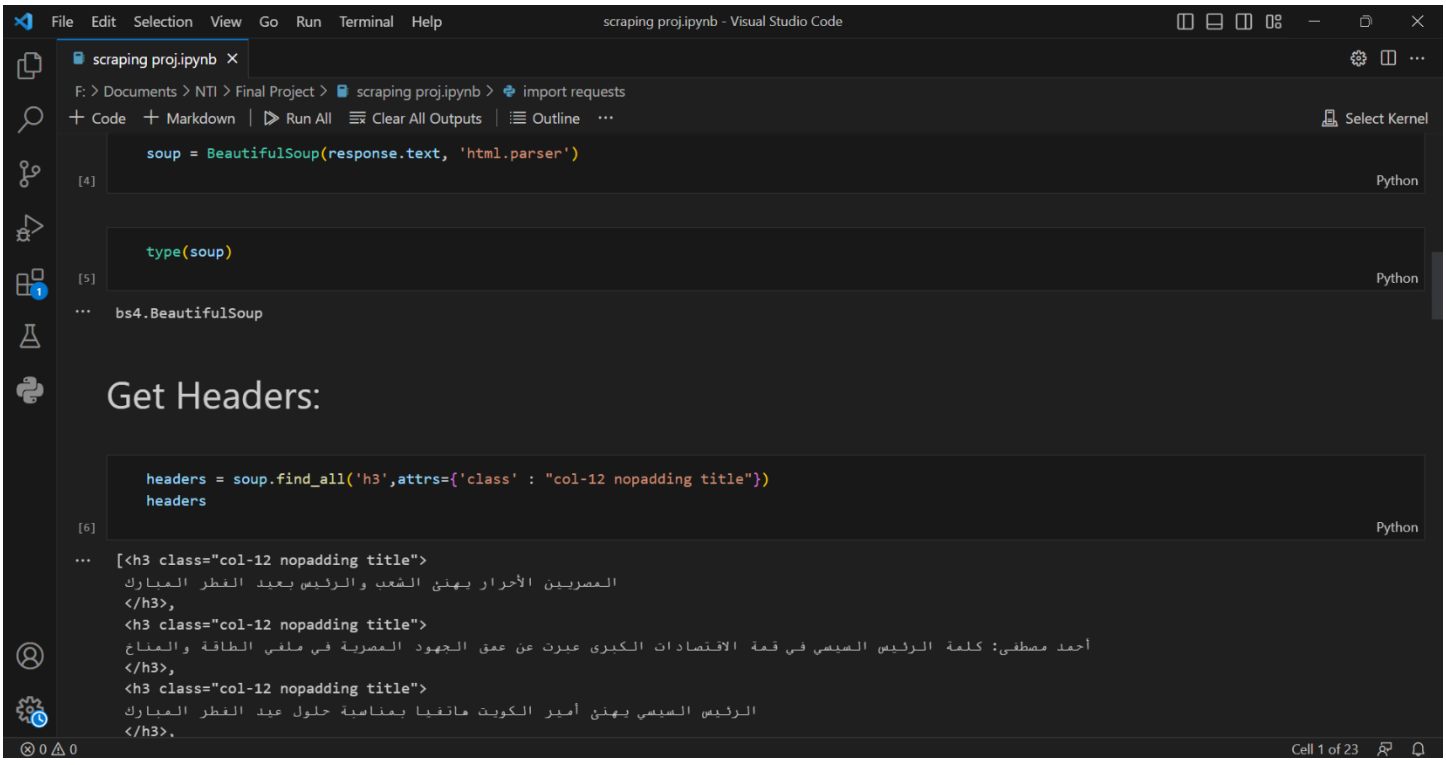


Figure 5

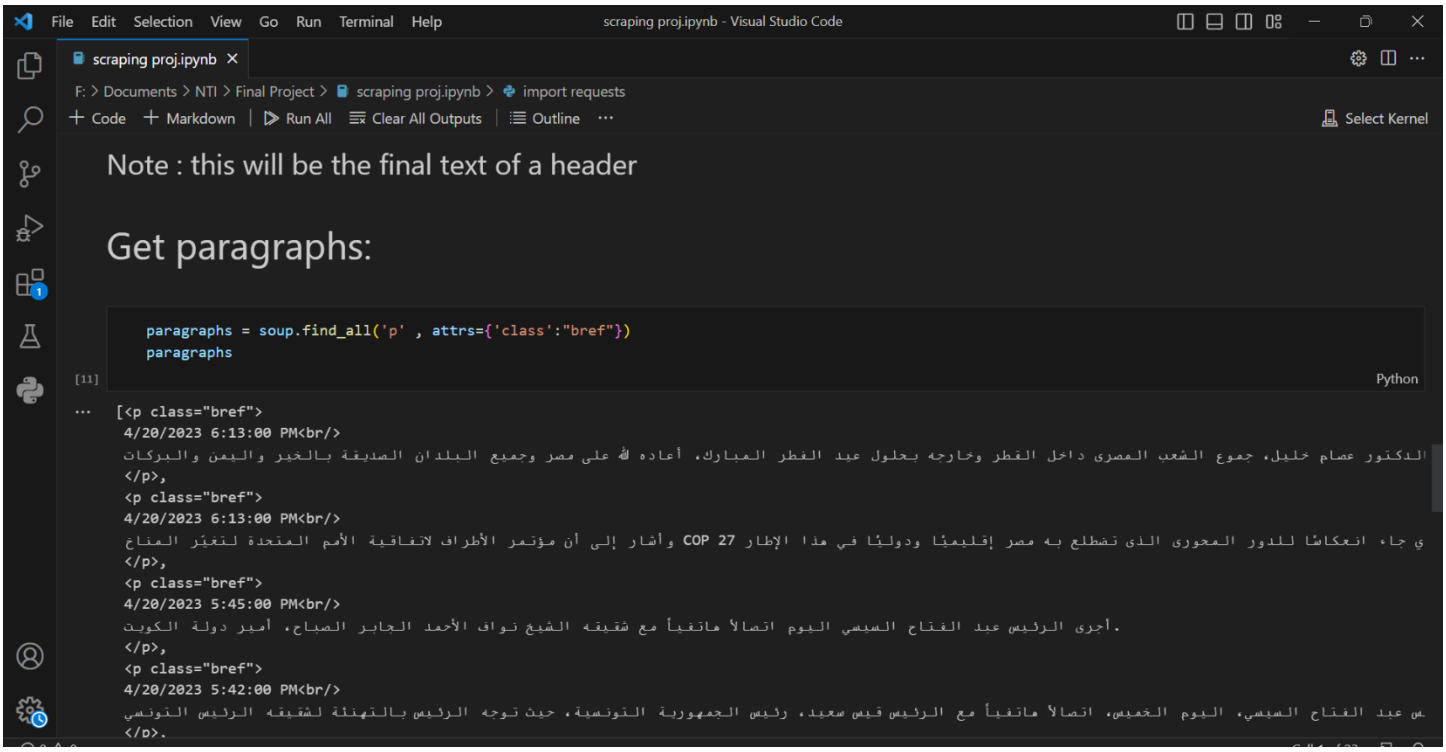


Figure 6

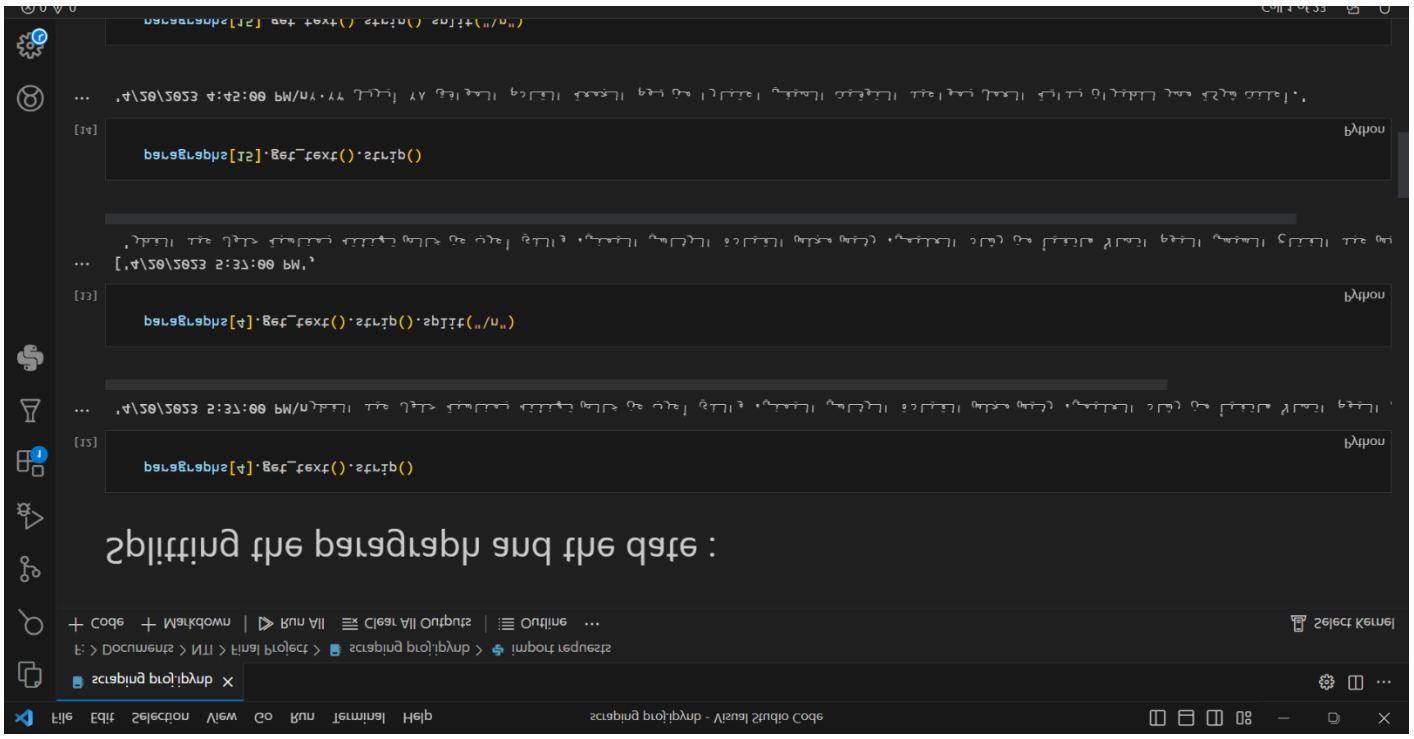


Figure 7

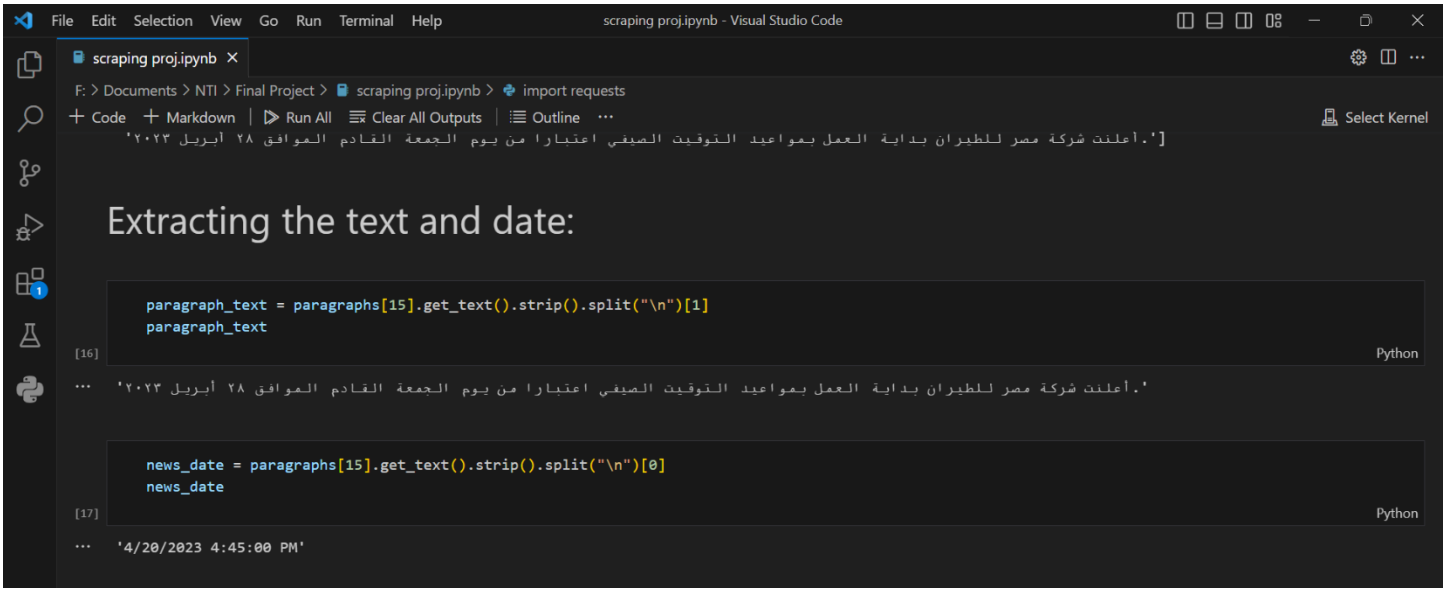


Figure 8

2.2. Translation:

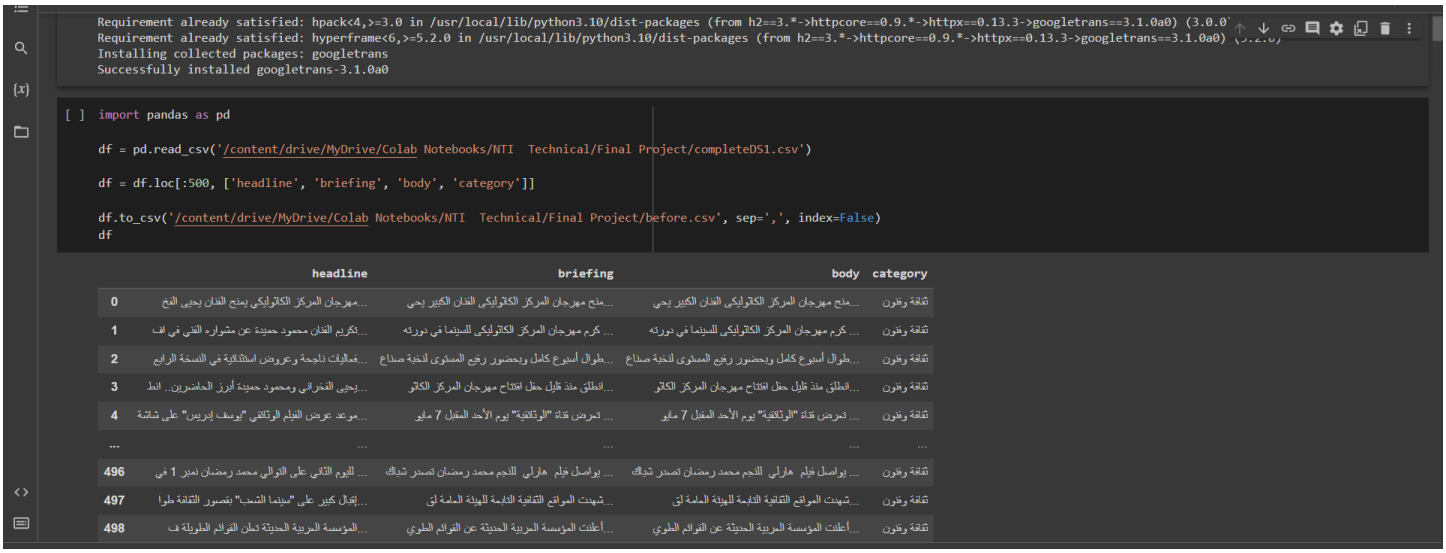


Figure 9

```python
import csv
from googletrans import Translator

# Set up the translator
translator = Translator()

# Specify the input and output CSV file paths
input_csv_file = '/content/drive/MyDrive/Colab Notebooks/NTI  Technical/Final Project/before.csv'
output_csv_file = '/content/drive/MyDrive/Colab Notebooks/NTI  Technical/Final Project/new_csv/after.csv'

# Open the input CSV file and read its contents
with open(input_csv_file, 'r', encoding='utf-8-sig') as f:
    reader = csv.reader(f)
    data = [row for row in reader]

# Translate the data
for row in data:
    translated_row = []
    for cell in row:
        if cell:
            translation = translator.translate(cell, src='ar', dest='en')
            translated_row.append(translation.text)
        else:
            translated_row.append('')
    row[:] = translated_row

# Write the translated data to the output CSV file
with open(output_csv_file, 'w', encoding='utf-8-sig', newline='') as f:
    writer = csv.writer(f)
    writer.writerows(data)
```

Figure 10

2.3. Preprocessing:

```python
# ----------------------------------------------------------------------------------------
# defining a fuction to remove punctuations
def remove_punc(news_text):
    result = re.sub(r'[0-9]+', '', news_text)
    text = re.compile('[%s]' % re.escape(string.punctuation)).sub('', result)
    return text

# defining a function to remove stop words
def remove_stop(the_text):
  # Tokenize the text into words
  words = word_tokenize(the_text)

  # Filter out the stop words from the text
  filtered_words = [word for word in words if not word in stop_words]

  # Join the filtered words into a string
  filtered_text = ' '.join(filtered_words)

  # return the filtered text
  return(filtered_text)

# prepare spacy model for lemmatization
nlp = spacy.load('en_core_web_lg')
nlp = en_core_web_lg.load()


# defining lemmatization function
def lemm(text):
    lemme=[]
    for token in nlp(text):
        lemme.append(token.lemma_)

    return " ".join(lemme)


# defining a fuction that combines all preprocessing steps needed
def preprocessing_text(text):
    text = remove_punc(text)
    text = remove_stop(text)
    text = lemm(text)
    return text
```

Figure 11

2.4. Text Classification:

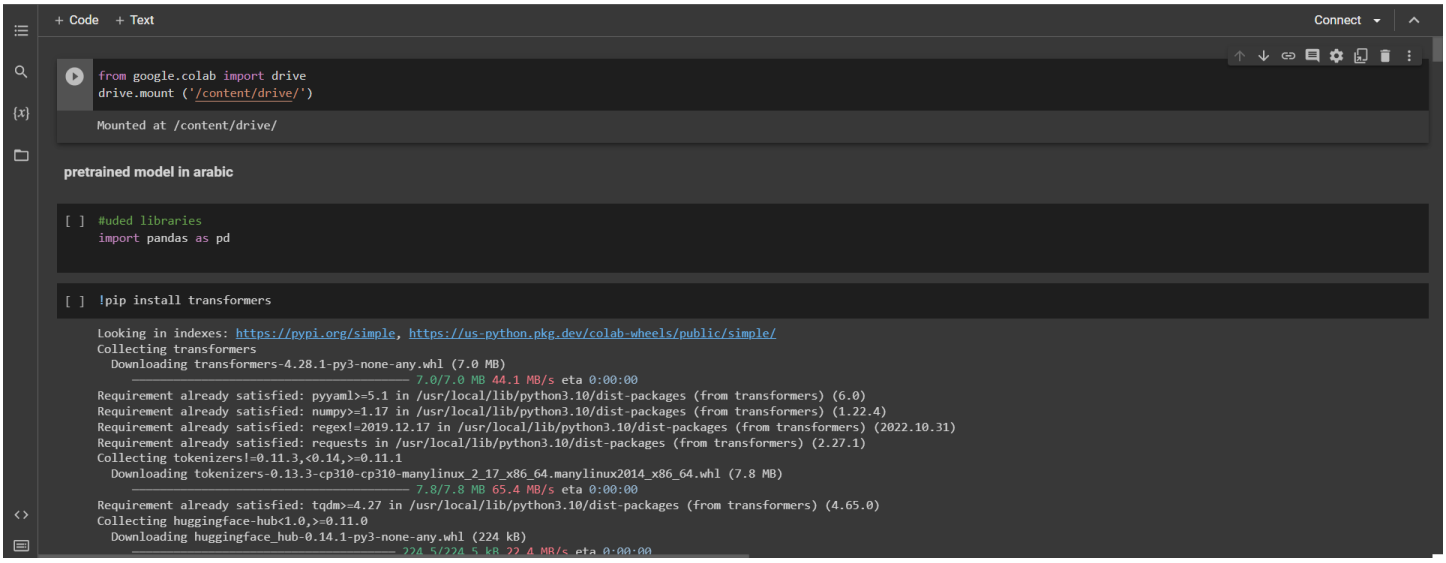


Figure 12

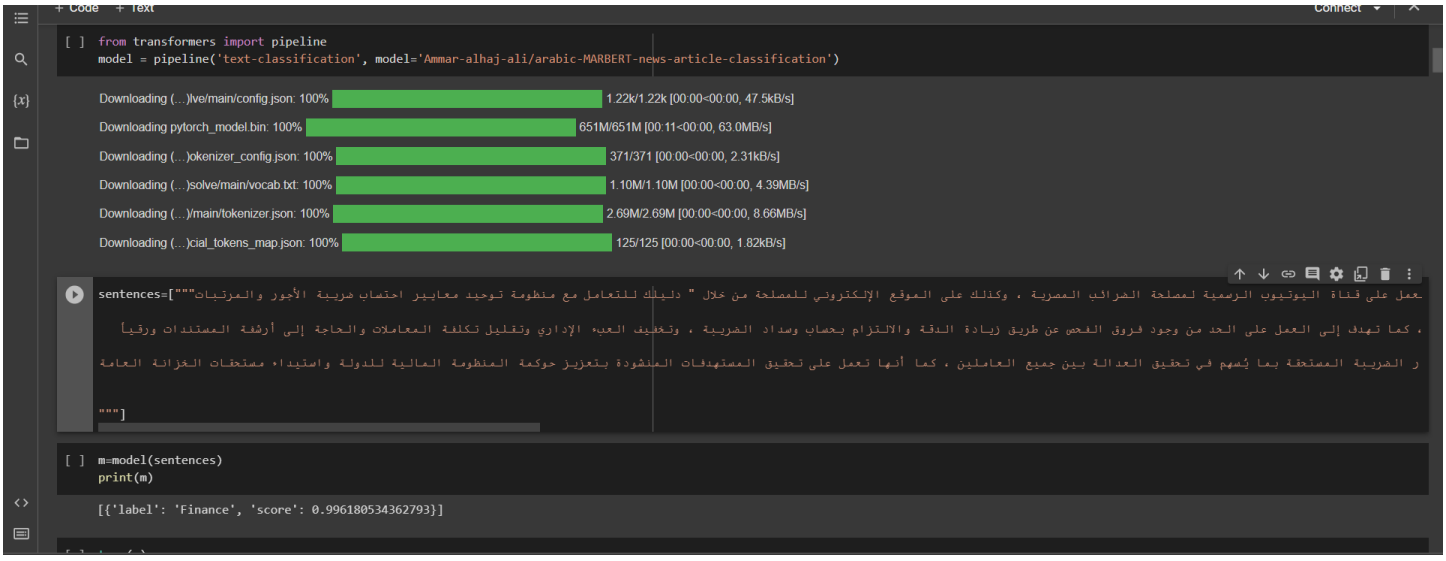


Figure 13

trained model

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    import re
    import nltk
    from nltk.corpus import stopwords
    nltk.download('stopwords')
    from nltk.stem import PorterStemmer
    from nltk.stem import WordNetLemmatizer
    nltk.download('wordnet')
    from nltk.tokenize import word_tokenize
    from nltk.tokenize import sent_tokenize
    nltk.download('punkt')
    from wordcloud import WordCloud
    from sklearn.feature_extraction.text import CountVectorizer
    from sklearn.feature_extraction.text import TfidfVectorizer
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import accuracy_score
    from sklearn.metrics import confusion_matrix
    from sklearn.metrics import classification_report
    from sklearn.metrics import make_scorer, roc_curve, roc_auc_score
    from sklearn.metrics import precision_recall_fscore_support as score
    from sklearn.metrics.pairwise import cosine_similarity
    from sklearn.multiclass import OneVsRestClassifier
    from sklearn.linear_model import LogisticRegression
    from sklearn.svm import SVC, LinearSVC
    from sklearn.tree import DecisionTreeClassifier
```

Figure 14

```
[ ] dataset=pd.read_csv("/content/drive/MyDrive/final project_NTI/final_data_yarab_b2a")
```

```
[ ] dataset=dataset[['body', 'category']]
    dataset
```

	body	category
0	Dr. Tariq Rahmi, Governor of Gharbia, met with...	محافظات
1	The Department of Youth Development at the Dir...	محافظات
2	Today, Friday, the Follow-up Department of the...	محافظات
3	Ahmed Samir, President of the Egyptian Federat...	محافظات
4	The Civil Protection Forces at Al-Hamoul Cente...	محافظات
...	...	...
20225	The price of gold recorded today, Monday, in E...	اقتصاد
20226	Today, Monday, the price of the euro witnessed...	اقتصاد
20227	Minister of Manpower, Hassan Shehata, said tha...	اقتصاد
20228	Mokhtar Tawfiq, Head of the Egyptian Tax Autho...	اقتصاد
20229	The Sugar and Integrated Industries Company of...	اقتصاد

20230 rows × 2 columns

Figure 15

```
[ ] dataset.shape

    (20230, 2)

[ ] target_category = dataset['category'].unique()
    print(target_category)

    ['محافظات' 'رياضة' 'عرب وعالم' 'أخبار' 'ثقافة وفنون' 'حوادث' 'اقتصاد']

[ ] dataset['CategoryId'] = dataset['category'].factorize()[0]

[ ] dataset
```

	body	category	CategoryId
0	Dr. Tariq Rahmi, Governor of Gharbia, met with...	محافظات	0
1	The Department of Youth Development at the Dir...	محافظات	0
2	Today, Friday, the Follow-up Department of the...	محافظات	0
3	Ahmed Samir, President of the Egyptian Federat...	محافظات	0
4	The Civil Protection Forces at Al-Hamoul Cente...	محافظات	0
...	...	...	...

Figure 16

```
[ ] # Create a new pandas dataframe "category", which only has unique Categories, also sorting this list in order of CategoryId values
    category = dataset[['category', 'CategoryId']].drop_duplicates().sort_values('CategoryId')
    category
```

	category	CategoryId
0	محافظات	0
2996	رياضة	1
5966	عرب وعالم	2
8937	أخبار	3
11663	ثقافة وفنون	4
14494	حوادث	5
17484	اقتصاد	6

```
[ ] dataset['category'].value_counts()

    2996     محافظات
    2990      حوادث
    2971    عرب وعالم
    2970      رياضة
    2831    ثقافة وفنون
    2746      اقتصاد
    2726      أخبار
    Name: category, dtype: int64
```

Figure 17

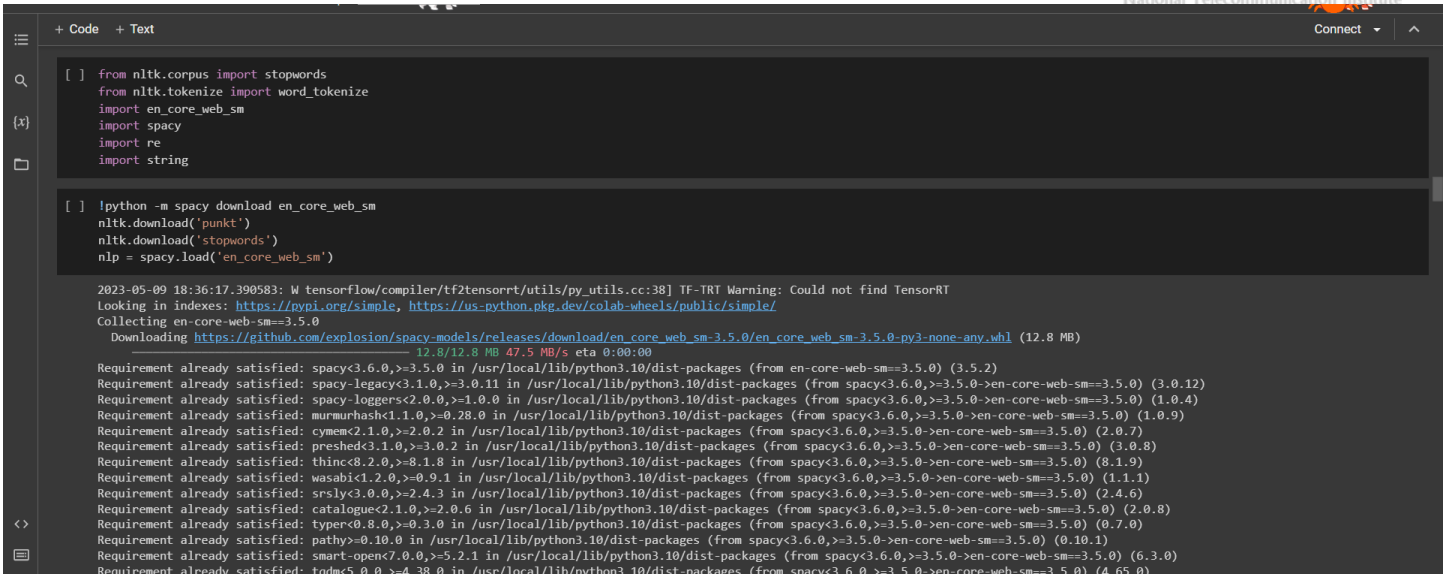


Figure 18

Figure 19

```
[ ] nlp = spacy.load('en_core_web_sm')

    nlp = en_core_web_sm.load()
```

```
[ ] def lemm(text):
        lemme=[]
        for token in nlp(text):
            lemme.append(token.lemma_)

        return " ".join(lemme)
```

```
[ ] def preprocessing_text(text):
        text = remove_punc(text)
        text = remove_stop(text)
        text = lemm(text)
        return text
```

```
[ ] dataset['text']=dataset['body'].apply(preprocessing_text)
```

```
[ ] dataset['text']

    0        Dr Tariq Rahmi Governor Gharbia meet Eng Nasse...
    1        the Department Youth Development Directorate Y...
    2        today Friday Followup Department Directorate H...
    3        Ahmed Samir President Egyptian Federation Mini...
```

Figure 20

```
[ ] data_prprocessed=dataset[['text','category','CategoryId']]
    data_prprocessed.to_csv('/content/drive/MyDrive/final project_NTI/Data/final data/preprocessed_final_insh2llah.csv')
```

Vectorization

```
[ ] !pip install gensim

    Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
    Requirement already satisfied: gensim in /usr/local/lib/python3.10/dist-packages (4.3.1)
    Requirement already satisfied: numpy>=1.18.5 in /usr/local/lib/python3.10/dist-packages (from gensim) (1.22.4)
    Requirement already satisfied: scipy>=1.7.0 in /usr/local/lib/python3.10/dist-packages (from gensim) (1.10.1)
    Requirement already satisfied: smart-open>=1.8.1 in /usr/local/lib/python3.10/dist-packages (from gensim) (6.3.0)
```

```
[ ] import gensim
    from gensim.models.doc2vec import Doc2Vec, TaggedDocument
    from nltk.tokenize import word_tokenize
```

```
[ ] # Tokenize documents and create TaggedDocuments for Doc2Vec model
    tagged_data = [TaggedDocument(words=word_tokenize(doc), tags=[str(i)]) for i, doc in enumerate(data_prprocessed['text'])]

    # Train Doc2Vec model
    model = Doc2Vec(tagged_data, vector_size=100, window=5, min_count=1, workers=4, epochs=20)
```

```
[ ] # Obtain vector representations for each document
    vectors = [model.docvecs[str(i)] for i in range(len(data_prprocessed))]
```

Figure 21

```
[ ] # Tokenize documents and create TaggedDocuments for Doc2Vec model
    tagged_data = [TaggedDocument(words=word_tokenize(doc), tags=[str(i)]) for i, doc in enumerate(data_prprocessed['text'])]

    # Train Doc2Vec model
    model = Doc2Vec(tagged_data, vector_size=100, window=5, min_count=1, workers=4, epochs=20)

[ ] # Obtain vector representations for each document
    vectors = [model.docvecs[str(i)] for i in range(len(data_prprocessed))]

    <ipython-input-32-c5b1bbefd54e>:2: DeprecationWarning: Call to deprecated `docvecs` (The `docvecs` property has been renamed `dv`.).
      vectors = [model.docvecs[str(i)] for i in range(len(data_prprocessed))]

[ ] vectors

    [array([-1.2097434e+00,  8.7511396e-01,  9.1965395e-01, -1.1184933e+00,
             4.1932475e-02, -3.5396212e-01, -6.5525037e-01,  7.5402391e-01,
            -4.4432843e-01,  1.8710719e-01,  7.1630347e-01,  5.8010572e-01,
             1.4167716e-01,  1.9786741e-01,  2.7465883e-01,  6.7876995e-02,
            -1.1639152e+00,  3.9724788e-01,  4.7962978e-01,  2.7686346e-01,
            -1.3112479e+00,  5.0851792e-01, -2.5025785e-01,  7.0076525e-01,
            -6.4901423e-01,  1.1748614e+00, -1.0963533e+00,  8.2570189e-01,
            -9.4997311e-01, -1.8077277e-01, -2.8372517e-01, -1.2950180e-01,
             7.0535280e-02,  7.5022814e-05,  1.2101258e-01,  9.2961574e-01,
            -1.2946986e+00,  1.7754875e-01, -2.3727182e-01,  2.9393286e-01,
             4.7066835e-01, -3.5687590e-01, -3.5261530e-01, -4.2689260e-02,
             6.1662191e-01,  5.0619924e-01,  7.1438199e-01, -6.9174469e-02,
             7.1927524e-01, -4.6188533e-02, -1.4313592e+00,  3.6452678e-01,
            -3.2074571e-01,  9.8570549e-01,  9.4615698e-01,  8.0956198e-02,
             1.1537598e+00, -4.1953057e-01,  4.7480330e-01, -7.1571702e-01,
            -5.3070241e-01,  6.7072433e-01, -3.9268789e-01, -2.1185681e-01,
```

Figure 22

```
[ ] data_prprocessed['vector'] = vectors
    data_prprocessed
```

	text	category	CategoryId	vector
0	Dr Tariq Rahmi Governor Gharbia meet Eng Nasse...	محافظات	0	[-1.2097434, 0.87511396, 0.91965395, -1.118493...
1	the Department Youth Development Directorate Y...	محافظات	0	[-0.26734543, 0.043513965, -0.0036009243, -0.3...
2	today Friday Followup Department Directorate H...	محافظات	0	[-0.39624828, 0.4585688, 0.6427528, -0.2907614...
3	Ahmed Samir President Egyptian Federation Mini...	محافظات	0	[-0.4989984, -0.3932304, -0.10849956, -0.3112...
4	the Civil Protection Forces AlHamoul Center Ka...	محافظات	0	[-0.5460894, -0.2578218, 0.8180921, 0.25256616...
...	...	...	...	...
20225	the price gold record today Monday Egypt stabl...	اقتصاد	6	[-0.2861491, -0.18736519, 0.07955681, -0.61001...
20226	today Monday price euro witness remarkable sta...	اقتصاد	6	[-0.3785579, 0.034202, 0.54200995, -0.24133717...
20227	Minister Manpower Hassan Shehata say sign join...	اقتصاد	6	[-0.31237072, 1.4915228, 0.96611536, 0.1823597...
20228	Mokhtar Tawfiq Head Egyptian Tax Authority ann...	اقتصاد	6	[-1.133717, -0.5342979, 0.85990304, -0.7248993...
20229	the Sugar Integrated Industries Company Holdin...	اقتصاد	6	[0.30140892, -0.24901362, 0.1288036, -0.527487...

20230 rows × 4 columns

```
[ ] x=np.array(data_prprocessed['vector'])
    y=np.array(data_prprocessed['CategoryId'])
```

Figure 23

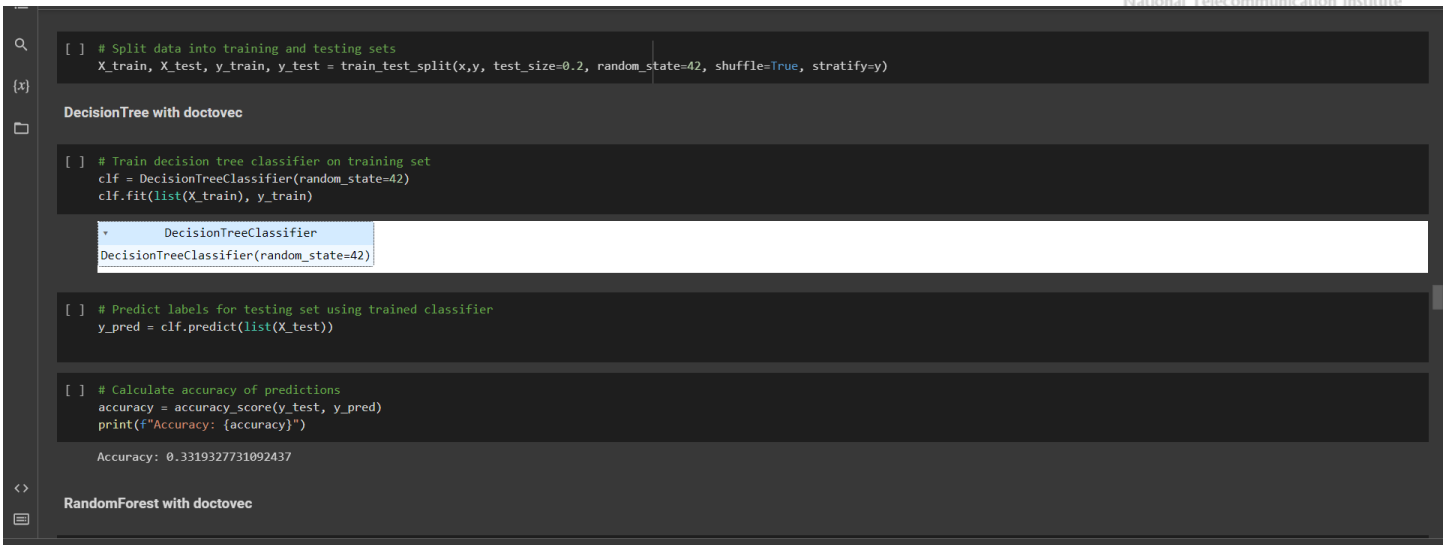


Figure 24

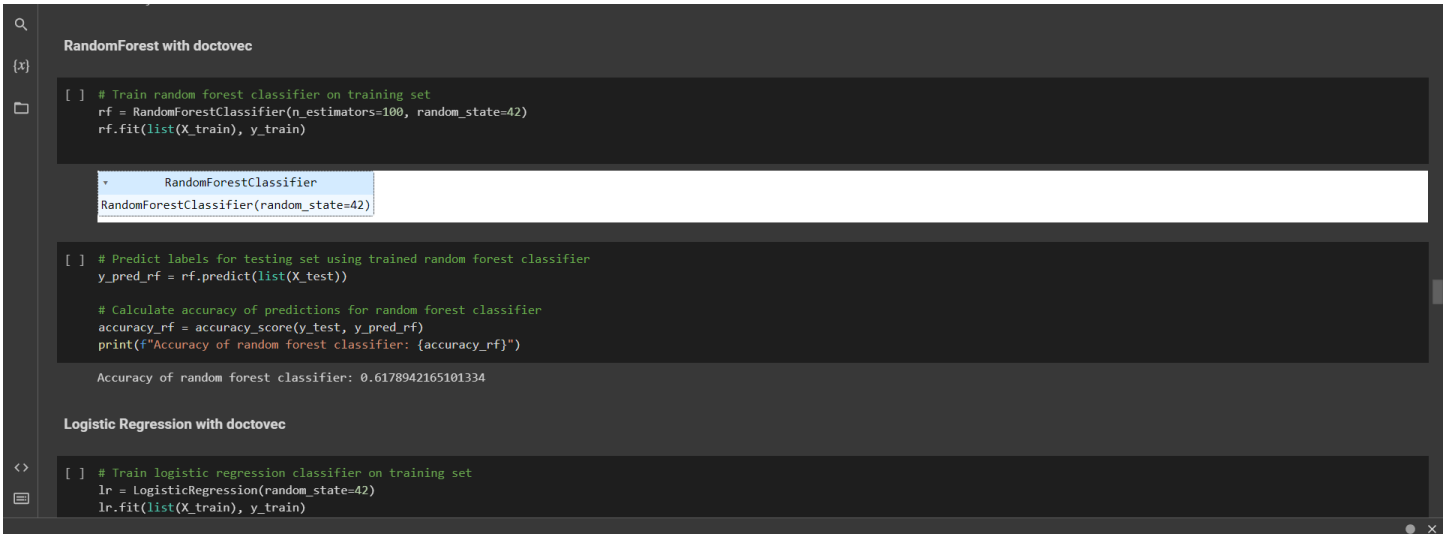


Figure 25

Figure 26

TFI-DF Embedding

```
[ ] from sklearn.feature_extraction.text import TfidfVectorizer
```

```
[ ] preprocess_df=pd.read_csv('/content/drive/MyDrive/final project_NTI/Data/final data/preprocessed_final_insh2llah.csv')
```

```
[ ] preprocess_df.drop('Unnamed: 0', axis=1,inplace=True)
```

```
[ ] preprocess_df
```

	text	category	CategoryId
0	Dr Tariq Rahmi Governor Gharbia meet Eng Nasse...	محافظات	0
1	the Department Youth Development Directorate Y...	محافظات	0
2	today Friday Followup Department Directorate H...	محافظات	0
3	Ahmed Samir President Egyptian Federation Mini...	محافظات	0
4	the Civil Protection Forces AlHamoul Center Ka...	محافظات	0
...	...	...	...
20225	the price gold record today Monday Egypt stabl...	اقتصاد	6
20226	today Monday price euro witness remarkable sta...	اقتصاد	6

Figure 27

```
[ ] # Preprocess data using TfidfVectorizer
    vectorizer = TfidfVectorizer()
    x_tf = vectorizer.fit_transform(preprocess_df['text'])

    y_tf=preprocess_df['CategoryId']
```

```
[ ] # Split data into training and testing sets
    X_trainn, X_testt, y_trainn, y_testt = train_test_split(x_tf, y_tf, test_size=0.2, random_state=42,shuffle=True, stratify=y_tf)
```

```
[ ] # Train random forest classifier on training set
    rf_tf = RandomForestClassifier(n_estimators=100, random_state=42)
    rf_tf.fit(X_trainn, y_trainn)
```

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
[ ] # Predict labels for testing set using trained random forest classifier
    y_pred_TF_Random = rf_tf.predict(X_testt)
```

```
[ ] # Calculate accuracy of predictions
    accuracy = accuracy_score(y_testt, y_pred_TF_Random)
    print(f"Accuracy: {accuracy}")
```

Figure 28

XGBOOST with TFIDF

```
[ ] df_preprocessed_edit=pd.read_csv('/content/drive/MyDrive/final project_NTI/Data/final data/preprocessed_final_insh2llah.csv')
```

```
[ ] df_preprocessed_edit
```

	Unnamed: 0	text	category	CategoryId
0	0	Dr Tariq Rahmi Governor Gharbia meet Eng Nasse...	محافظات	0
1	1	the Department Youth Development Directorate Y...	محافظات	0
2	2	today Friday Followup Department Directorate H...	محافظات	0
3	3	Ahmed Samir President Egyptian Federation Mini...	محافظات	0
4	4	the Civil Protection Forces AlHamoul Center Ka...	محافظات	0
...	...	...	...	...
20225	20225	the price gold record today Monday Egypt stabl...	اقتصاد	6
20226	20226	today Monday price euro witness remarkable sta...	اقتصاد	6
20227	20227	Minister Manpower Hassan Shehata say sign join...	اقتصاد	6
20228	20228	Mokhtar Tawfiq Head Egyptian Tax Authority ann...	اقتصاد	6
20229	20229	the Sugar Integrated Industries Company Holdin...	اقتصاد	6

20230 rows × 4 columns

Figure 29

```
[ ] x_g=df_preprocessed_edit['text']
    y_g=df_preprocessed_edit['CategoryId']
```

```
[ ] # Split data into training and testing sets
    X_trainxg, X_testxg, y_traingxg, y_testgxg = train_test_split(x_g, y_g, test_size=0.05, random_state=42,shuffle=True, stratify=y_g)
```

```
[ ] # Preprocess data using TfidfVectorizer
    vectorizer_tf = TfidfVectorizer(max_features=5000,ngram_range=(1,3))
    X_train_gb = vectorizer_tf.fit_transform(X_trainxg)
    X_testg_gb = vectorizer_tf.transform(X_testxg )
```

```
[ ] import xgboost as xgb
```

```
[ ] # Initialize XGBClassifier object
    xg = xgb.XGBClassifier()

    # Train the XGBoost model
    xg.fit(X_train_gb, y_traingxg)
```

```
XGBClassifier
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, early_stopping_rounds=None,
```

Figure 30

```
[ ] # Preprocess data using TfidfVectorizer
    vectorizer = TfidfVectorizer(max_features=5000,ngram_range=(1,3))
    x = vectorizer.fit_transform(x_g)

[ ] xg.fit(x, y_g)
```

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=0.8, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='mlogloss',
              feature_types=None, gamma=None, gpu_id=None, grow_policy=None,
              importance_type=None, interaction_constraints=None,
              learning_rate=0.1, max_bin=None, max_cat_threshold=None,
              max_cat_to_onehot=None, max_delta_step=None, max_depth=6,
              max_leaves=None, min_child_weight=None, missing=nan,
              monotone_constraints=None, n_estimators=100, n_jobs=None,
              num_parallel_tree=None, objective='multi:softmax', predictor=None, ...)
```

```
[ ] xg = xgb.XGBClassifier(
        max_depth=6,
        learning_rate=0.1,
        n_estimators=100,
        subsample=0.8,
        colsample_bytree=0.8,
        objective='multi:softmax',
        eval_metric='mlogloss')
```

Figure 31

```
[ ]     n_estimators=100,
        subsample=0.8,
        colsample_bytree=0.8,
        objective='multi:softmax',
        eval_metric='mlogloss')

[ ] xg.fit(X_train_gb, y_traingxg)
```

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=0.8, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='mlogloss',
              feature_types=None, gamma=None, gpu_id=None, grow_policy=None,
              importance_type=None, interaction_constraints=None,
              learning_rate=0.1, max_bin=None, max_cat_threshold=None,
              max_cat_to_onehot=None, max_delta_step=None, max_depth=6,
              max_leaves=None, min_child_weight=None, missing=nan,
              monotone_constraints=None, n_estimators=100, n_jobs=None,
              num_parallel_tree=None, objective='multi:softmax', predictor=None, ...)
```

```
[ ] # Make predictions on the testing set
    y_pred_xg = xg.predict(X_testg_gb)

    # Evaluate the accuracy of the model
    accuracy_xg = accuracy_score(y_testgxg, y_pred_xg)
    print("Accuracy: %.2f%%" % (accuracy_xg * 100.0))

    Accuracy: 93.08%
```

Figure 32

2.5. Recommendation:

LDA:

```
import pandas as pd
```

```
df = pd.read_csv('/content/drive/MyDrive/NTI4M_FP/final_data_yarab_b2a')
df.head(10)
```

	hrefs	headline	briefing	body	category
0	https://gate.ahram.org.eg/News/4272955.aspx	محافظ الغربية يتابع الاستعدادات النهائية لإطلاق...	التقى الدكتور طارق رحمي محافظ الغربية، بالمهند...	Dr. Tariq Rahmi, Governor of Gharbia, met with...	محافظات
1	https://gate.ahram.org.eg/News/4272956.aspx	الشباب والرياضة" بالأقصر تنفذ ندوات عن التوعي...	نفذت إدارة التنمية الشباب بمديرية الشباب والرياض...	The Department of Youth Development at the Dir...	محافظات
2	https://gate.ahram.org.eg/News/4272936.aspx	متابعة عمل الوحدات الصحية والمراكز الطبية في 7...	قامت إدارة المتابعة بمديرية الشئون الصحية بمحا...	Today, Friday, the Follow-up Department of the...	محافظات
3	https://gate.ahram.org.eg/News/4272934.aspx	اعتماد ملعب مركز شباب السلما كأول ملعب لـ«الم...	قام أحمد سمير، رئيس الاتحاد المصرى للمبنى فوقب...	Ahmed Samir, President of the Egyptian Federat...	محافظات
4	https://gate.ahram.org.eg/News/4272941.aspx	إخماد حريق فى أرض مزروعة بالقمح فى مركز الحامو...	تمكنت قوات الحماية المدنية بمركز الحامول فى كف...	The Civil Protection Forces at Al-Hamoul Cente...	محافظات
5	https://gate.ahram.org.eg/News/4272940.aspx	الكشف على 1445 مريضًا فى قافلة طبية بقرية الثا...	نظمت مديرية الصحة فى البحيرة، قافلة طبية بقرية...	The Directorate of Health in Buhaira organized...	محافظات
6	https://gate.ahram.org.eg/News/4272939.aspx	إقامة صلاة الغائب على أرواح الأبرياء ضحايا حاد...	أدى جموع المصلين في الوادى الجديد عقب أداء صلا...	The crowds of worshipers in the New Valley, af...	محافظات
7	https://gate.ahram.org.eg/News/4272933.aspx	تنظيم قافلة سكانية في قرية البياقون في محافظة ...	نظمت وحدة السكان في محافظة البحيرة، قافلة سكان...	The Population Unit in Al-Buhaira Governorate ...	محافظات
8	https://gate.ahram.org.eg/News/4272938.aspx	إزالة حالتي بناء مخالف في مركز طما	سوهاج ... نفذت الوحدة المحلية لمركز ومدينة طما ب	The local unit of the center and city of Tama ...	محافظات
9	https://gate.ahram.org.eg/News/4272937.aspx	محافظ الشرقية يتابع الالتزام بالتعريفة في مواق...	تابع الدكتور ممدوح غراب محافظ الشرقية، سير الع...	Dr. Mamdouh Ghorab, Governor of Sharkia, follo...	محافظات

```
from sklearn.feature_extraction.text import TfidfVectorizer
tf = TfidfVectorizer(max_df=0.95, min_df=2, stop_words='english')
dtm = tf.fit_transform(df['body'])
```

```
from sklearn.feature_extraction.text import CountVectorizer
# cv = CountVectorizer()
cv = CountVectorizer(max_df=0.95, min_df=2, stop_words='english')
dtm1 = cv.fit_transform(df['body'])
```

Figure 33

```
len(cv.get_feature_names_out())
```

```
27131
```

```
import random
for i in range(10):
    random_word_id = random.randint(0,27131)
    print(cv.get_feature_names_out()[random_word_id])
```

```
storming
smashed
315
musafir
maxillofacial
shamelessly
affiliation
baalbek
availability
tuta
```

```
from sklearn.decomposition import LatentDirichletAllocation
LDA = LatentDirichletAllocation(random_state=42)
LDA.fit(dtm1)
```

```
LatentDirichletAllocation(random_state=42)
```

Figure 34

```
[ ] len(LDA.components_)

    10

[ ] LDA.components_

    array([[7.36080413e+01, 1.88808260e+00, 1.00691154e-01, ...,
            2.70999585e+01, 1.00000001e-01, 1.00089853e-01],
           [1.00491409e+03, 6.42346171e+00, 1.00000000e-01, ...,
            1.00000000e-01, 1.00000000e-01, 1.00000000e-01],
           [1.00004907e-01, 1.44823408e+00, 1.00000000e-01, ...,
            1.00028957e-01, 1.00000001e-01, 1.00002494e-01],
           ...,
           [1.00009190e-01, 1.03999136e+02, 1.00000000e-01, ...,
            1.00000000e-01, 2.09917594e+00, 1.00000000e-01],
           [1.01183195e-01, 2.58825148e+02, 1.00000000e-01, ...,
            1.00000029e-01, 1.00000001e-01, 1.00000000e-01],
           [1.25764363e+01, 5.60702767e+02, 1.00000000e-01, ...,
            1.00012506e-01, 1.00000001e-01, 2.09990765e+00]])

[ ] len(LDA.components_) , len(LDA.components_[0])

    (10, 27131)

[ ] for index,topic in enumerate(LDA.components_):
        print(f'THE TOP 15 WORDS FOR TOPIC #{index}')
        print([cv.get_feature_names_out()[i] for i in topic.argsort()[-15:]])
        print('\n')

    THE TOP 15 WORDS FOR TOPIC #0
    ['stadium', 'matches', 'minute', 'goals', 'scored', 'goal', 'round', 'ahly', 'place', 'second', 'team', 'league', 'match', 'points', 'al']


    THE TOP 15 WORDS FOR TOPIC #1
    ['el', 'god', 'air', 'work', 'departure', 'train', 'cairo', 'muhammad', 'ramadan', 'abdel', 'artist', 'time', 'ahmed', 'series', 'al']
```

Figure 35

```
[ ] topic_results = LDA.transform(dtm1)

[ ] topic_results.shape

    (20230, 10)

[ ]

    [2, 3, 4, 5]

[ ] topic_results[0].round(2)

    array([0.02, 0.01, 0.  , 0.62, 0.  , 0.19, 0.03, 0.04, 0.05, 0.04])

[ ] topic_results[0].argmax()

    3

[ ] topic_results.argmax(axis=1)

    array([3, 3, 3, ..., 5, 5, 8])

[ ] ex = df['body'][11]

[ ] ex = cv.transform([ex])
    ex = LDA.transform(ex)

[ ] ex.round(2)

    array([[0.  , 0.  , 0.  , 0.15, 0.  , 0.14, 0.  , 0.06, 0.64, 0.  ]])

[ ] ex.argmax()
```

Figure 36

```
[ ] ex.argmax()
    8

[ ] random.choices(df[ df.Topic == ex.argmax() ].index.tolist(),k = 10)
    [17880, 20088, 945, 19450, 10109, 19299, 19654, 19300, 17826, 85]

[ ] df['Topic'] = topic_results.argmax(axis=1)
    df.head(30)
```

	hrefs	headline	briefing	body	category	Topic
0	https://gate.ahram.org.eg/News/4272955.aspx	محافظ الغربية يتابع الاستعدادات النهائية لاحتـ...	التقى الدكتور طارق رحمي محافظ الغربية، بالمهند...	Dr. Tariq Rahmi, Governor of Gharbia, met with...	محافظات	3
1	https://gate.ahram.org.eg/News/4272956.aspx	الشباب والرياضة" بالأقصر تعقد ندوات عن التوعي...	نفذت إدارة التوعية الشباب بمديرية الشباب والرياض...	The Department of Youth Development at the Dir...	محافظات	3
2	https://gate.ahram.org.eg/News/4272936.aspx	متابعة عمل الوحدات الصحية والمراكز الطبية في 7...	قامت إدارة المتابعة بمديرية الشئون الصحية بمحا...	Today, Friday, the Follow-up Department of the...	محافظات	3
3	https://gate.ahram.org.eg/News/4272934.aspx	اعتماد ملعب مركز شباب السلما كأول ملعب لـ«دالمي...	قام أحمد سمير، رئيس الاتحاد المصرى للمبانى فرقاب...	Ahmed Samir, President of the Egyptian Federat...	محافظات	3
4	https://gate.ahram.org.eg/News/4272941.aspx	إخماد حريق في أرض مزروعة بالقمح في مركز الحامو...	تمكنت قوات الحماية المدنية بمركز الحامول في كف...	The Civil Protection Forces at Al-Hamoul Cente...	محافظات	7
5	https://gate.ahram.org.eg/News/4272940.aspx	الكشف على 1445 مريضًا في قافلة طبية بقرية الذا...	نظمت مديرية الصحة في البحيرة، قافلة طبية بقرية...	The Directorate of Health in Buhaira organized...	محافظات	3
6	https://gate.ahram.org.eg/News/4272939.aspx	إقامة صلاة الغائب على أرواح الأبرياء ضحايا حاد...	أدى جموع المصلين في الوادى الجديد عقب أداء صلا...	The crowds of worshipers in the New Valley, af...	محافظات	1
7	https://gate.ahram.org.eg/News/4272933.aspx	تنظيم قافلة سكانية في قرية البساتين في محافظة ...	نظمت وحدة السكان في محافظة البحيرة، قافلة سكان...	The Population Unit in Al-Buhaira Governorate ...	محافظات	3
8	https://gate.ahram.org.eg/News/4272938.aspx	إزالة حالتي بناء مخالف في مركز طما	صور ... نفذت الوحدة المحلية لمركز ومدينة طما بسوهاج ...	The local unit of the center and city of Tama ...	محافظات	7
9	https://gate.ahram.org.eg/News/4272937.aspx	محافظ الشرقية يتابع الالتزام بالتعريفة في مواق...	تابع الدكتور ممدوح غراب محافظ الشرقية، سير الع...	Dr. Mamdouh Ghorab, Governor of Sharkia, follo...	محافظات	7
10	https://gate.ahram.org.eg/News/4272914.aspx	جولات تموينية على محطات الوقود والصوامع في مرك...	تفقد السيد دايرة المشرف العام على مديرية التمو...	Today, Friday, Mr. Dayra, the general supervis...	محافظات	3
11	https://gate.ahram.org.eg/News/4272924.aspx	توريد 11088 طنا و454 كليو قمح إلى شون وصوامع ا...	قال اللواء عصام سعد محافظ أسيوط، إنه تم توريد ...	Major General Essam Saad, Governor of Assiut, ...	محافظات	8
12	https://gate.ahram.org.eg/News/4272927.aspx	صيدلة 49 كشاف إدارة في مركز الشلجات بمحافظة ا...	أعلن وائل زغلول رئيس الوحدة المحلية لمركز ومدي...	Wael Zaghloul, head of the local unit of the c...	محافظات	7
13	https://gate.ahram.org.eg/News/4272932.aspx	ضبط 410 أطنان قمح خلال حملة في المأثر من رمضان	أكدت فايزة عبد الرحمن، وكيلة وزارة التموين بال...	Fayza Abdel Rahman, Undersecretary of the Mini...	محافظات	7

Figure 37

Tf-idf similarity:

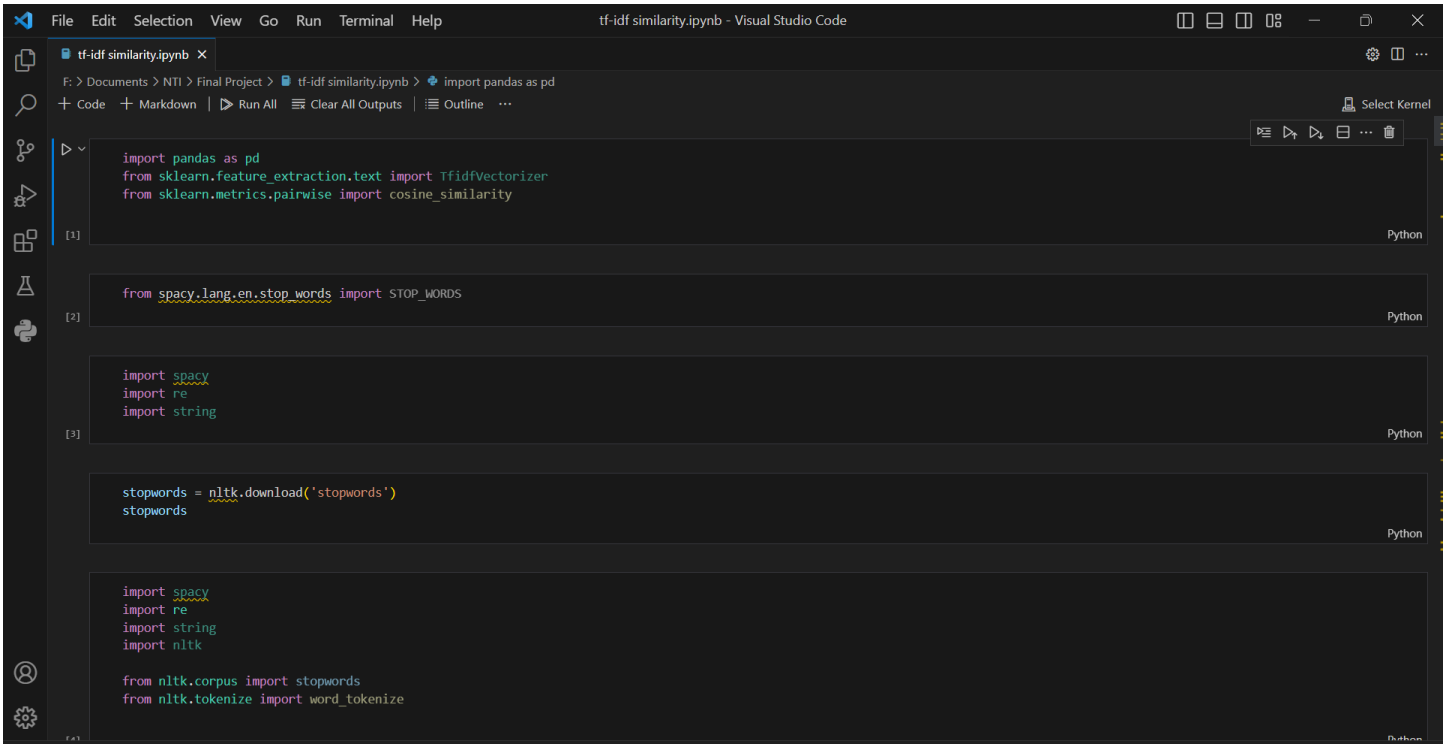


Figure 38

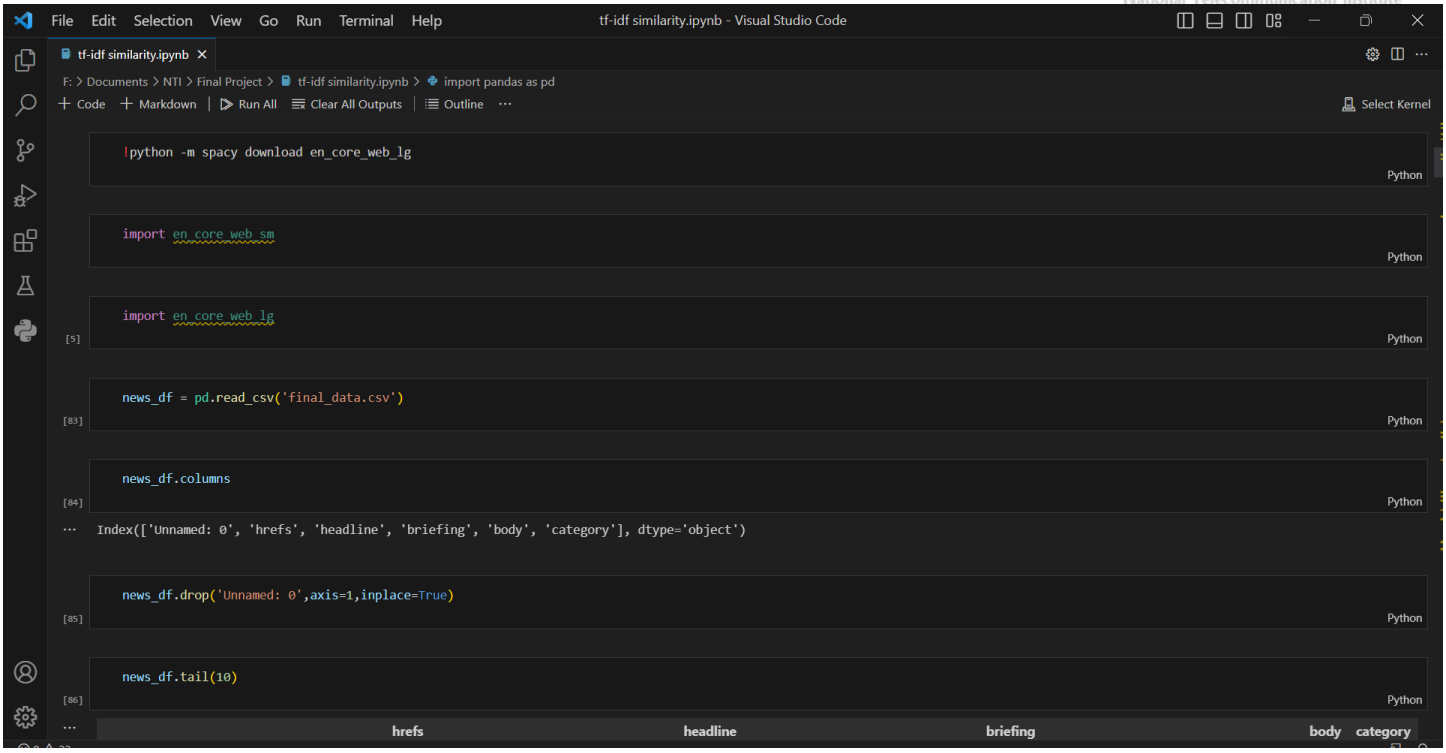


Figure 39

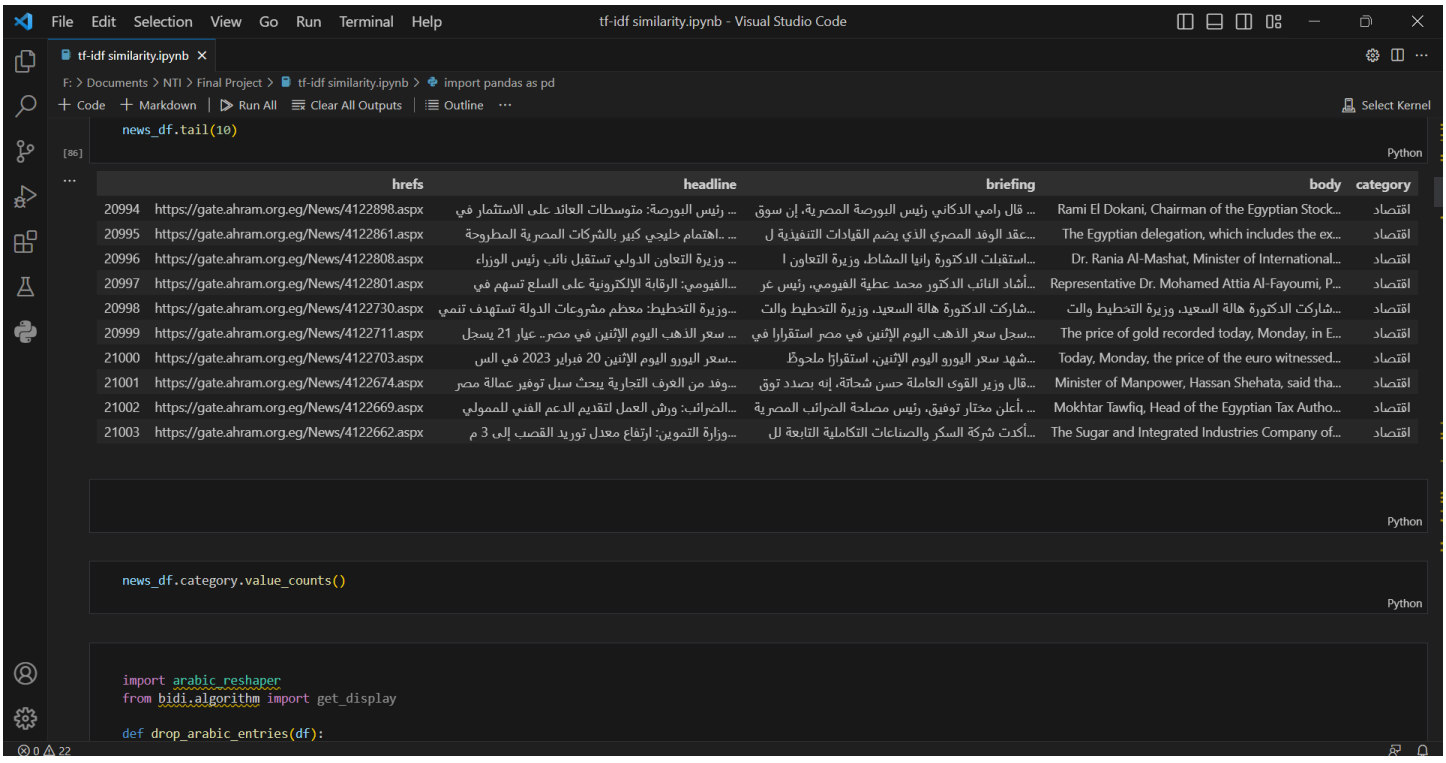


Figure 40

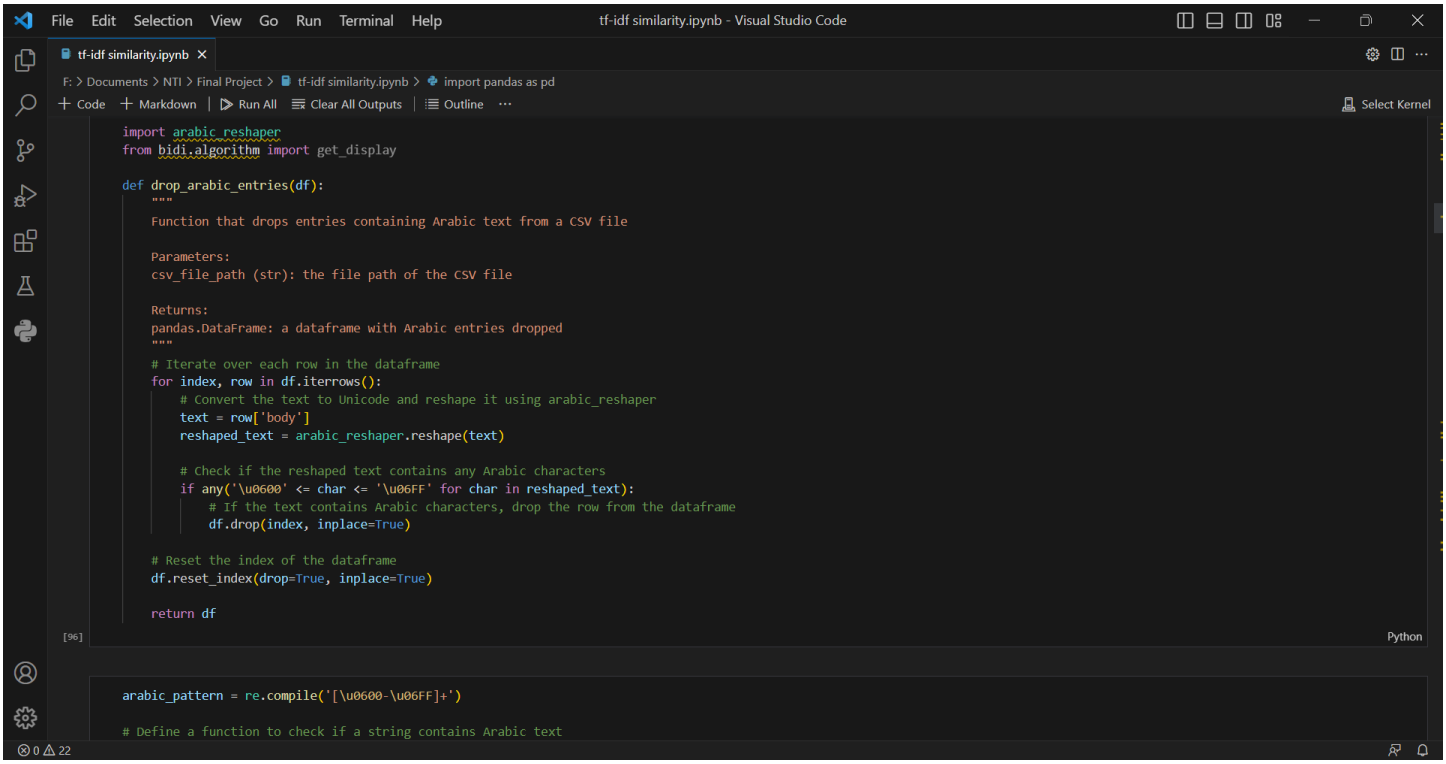


Figure 41

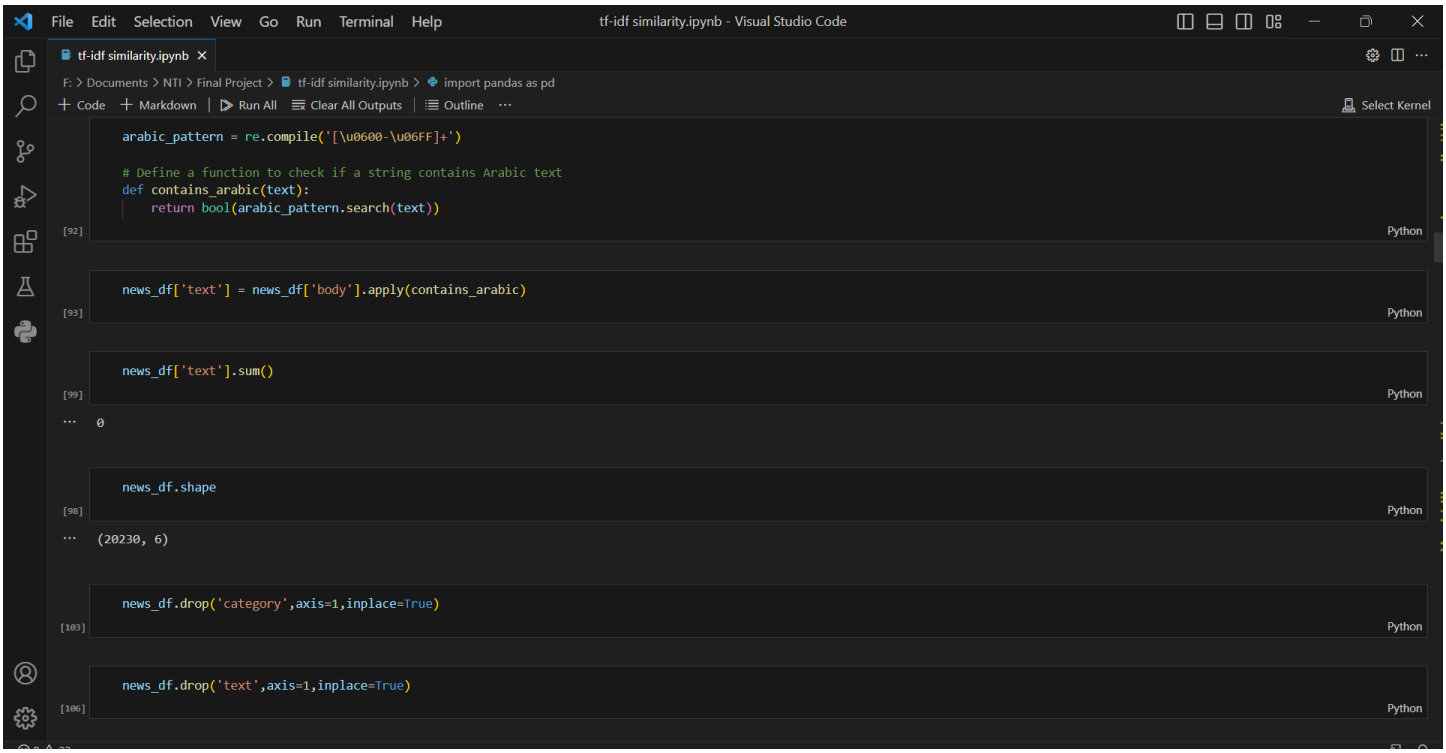


Figure 42

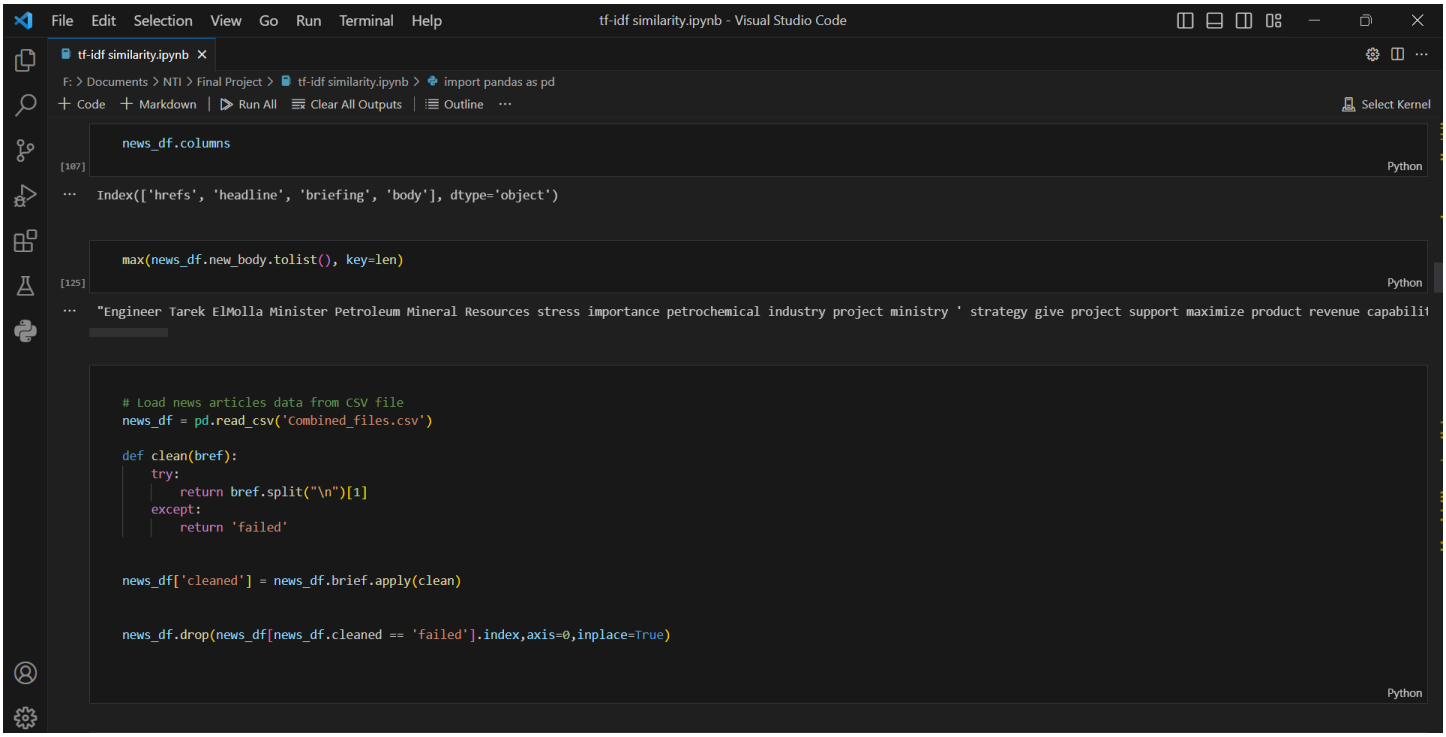


Figure 43

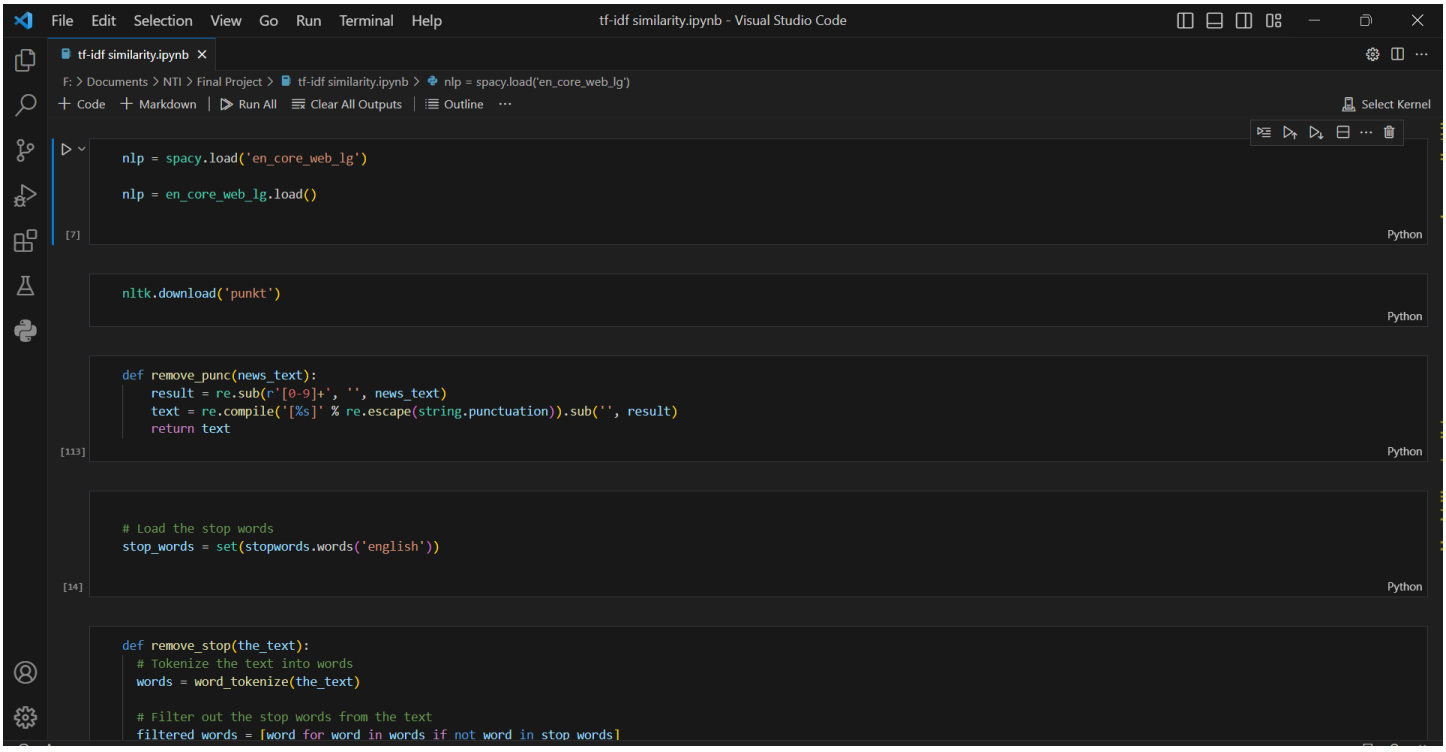


Figure 44

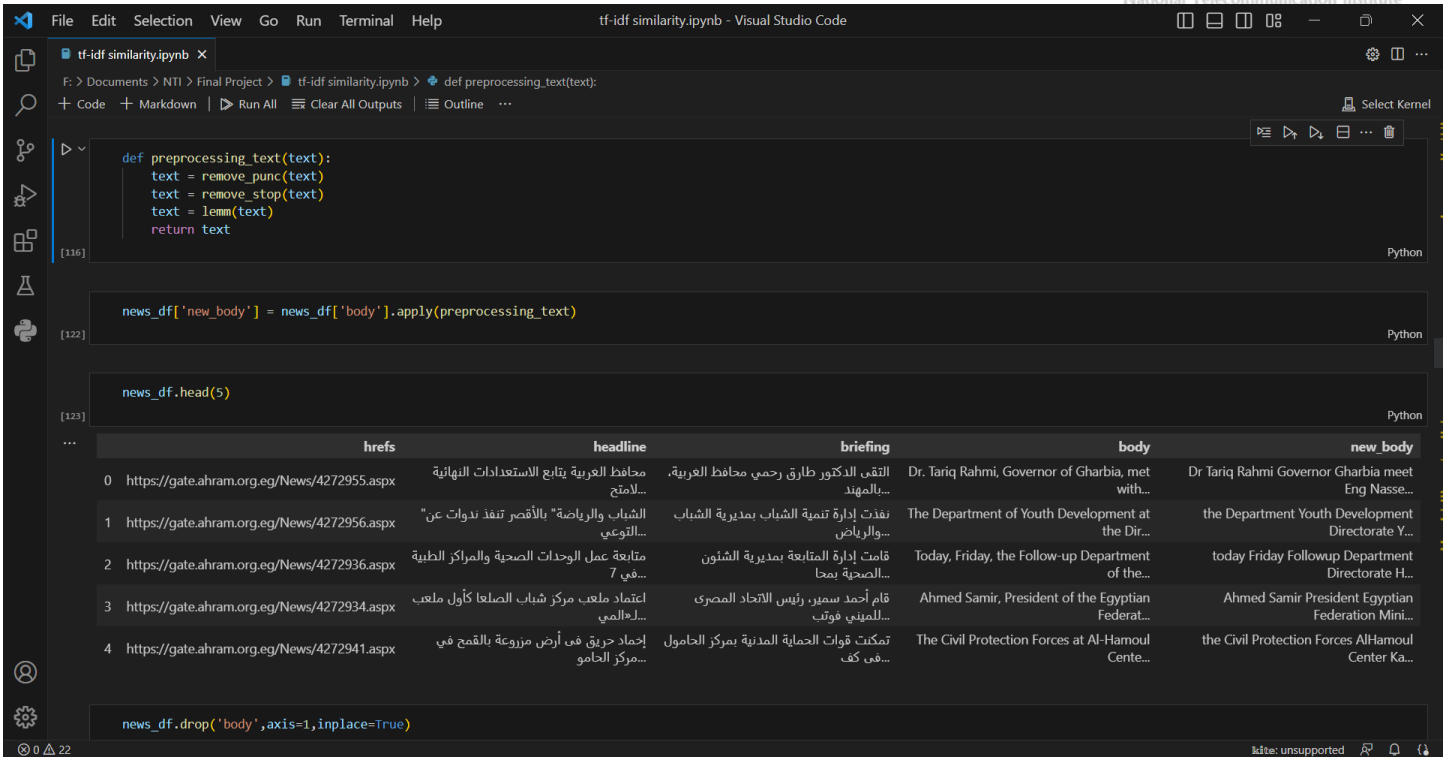


Figure 45

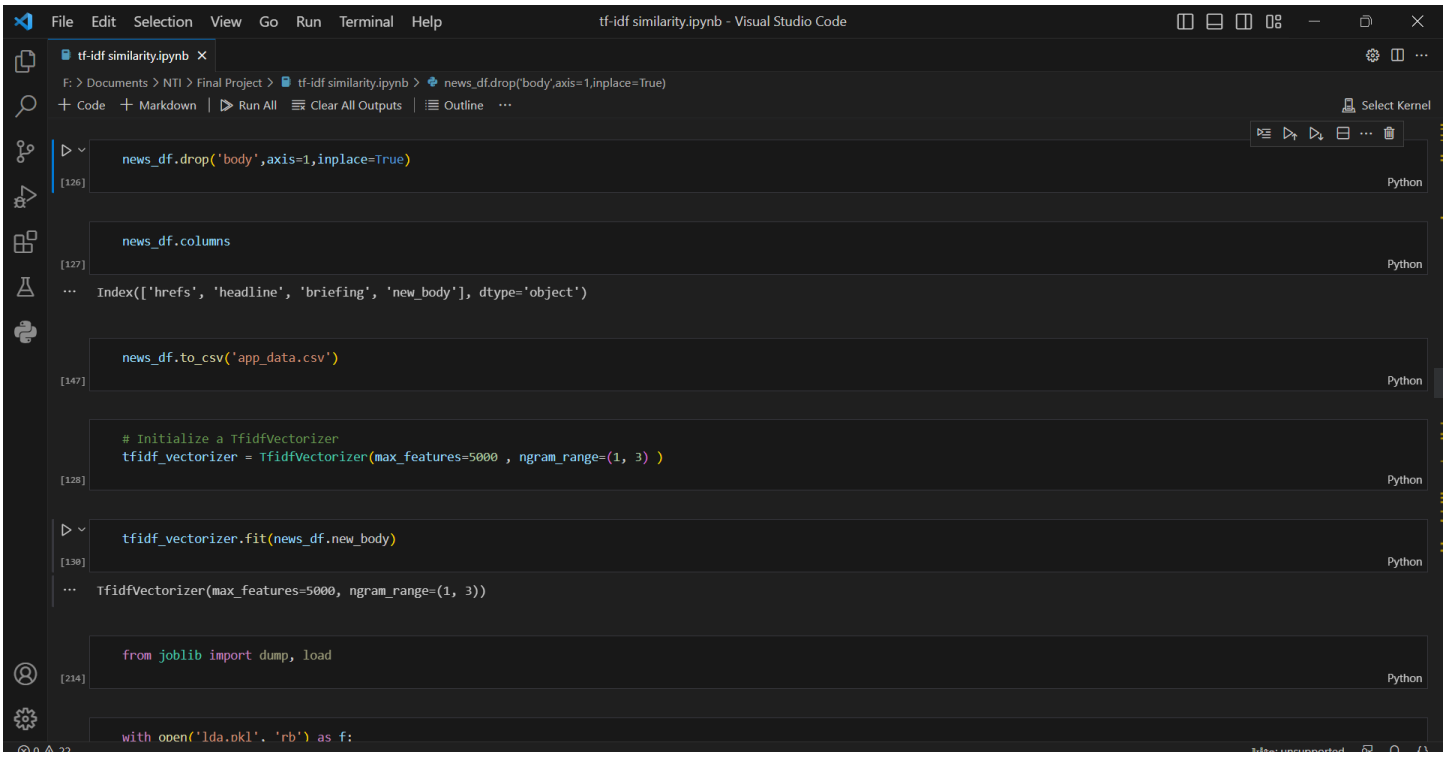


Figure 46

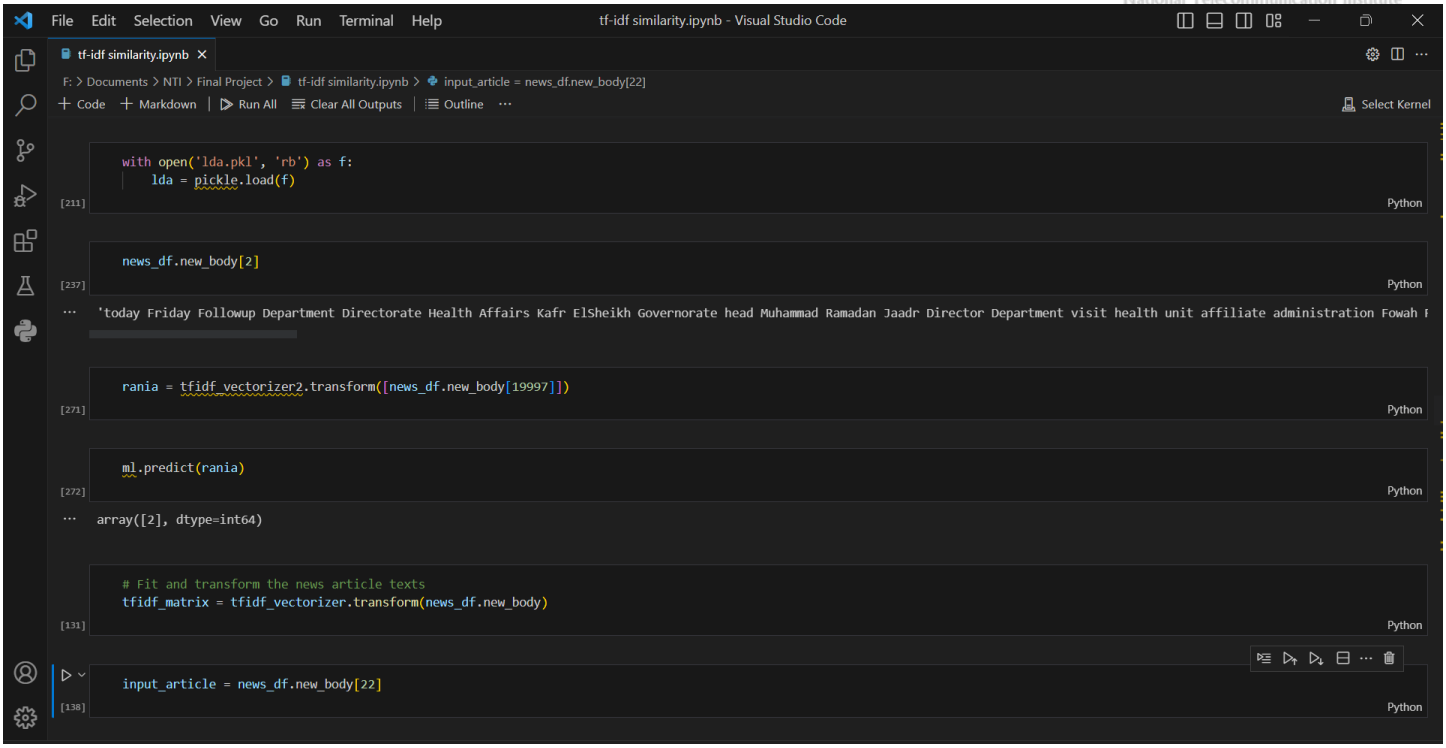


Figure 47

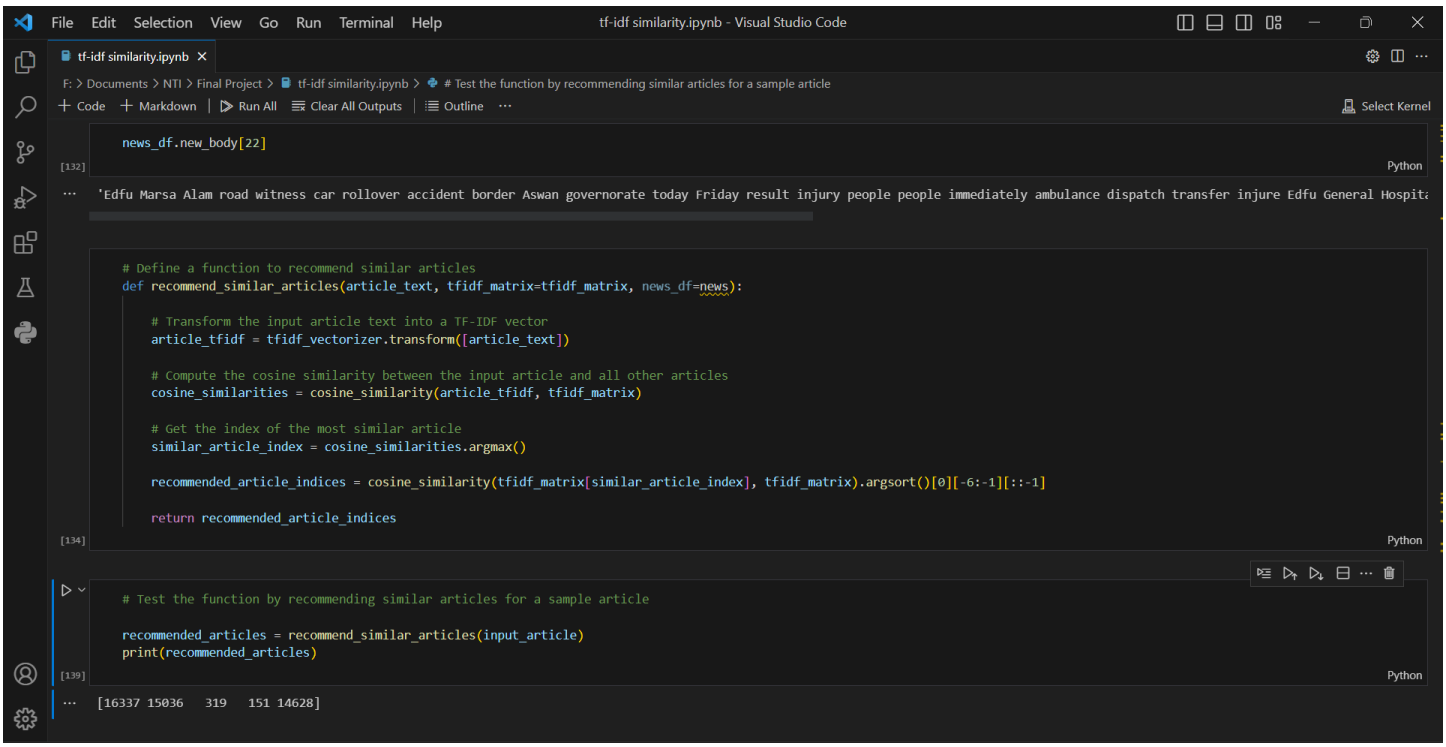


Figure 48

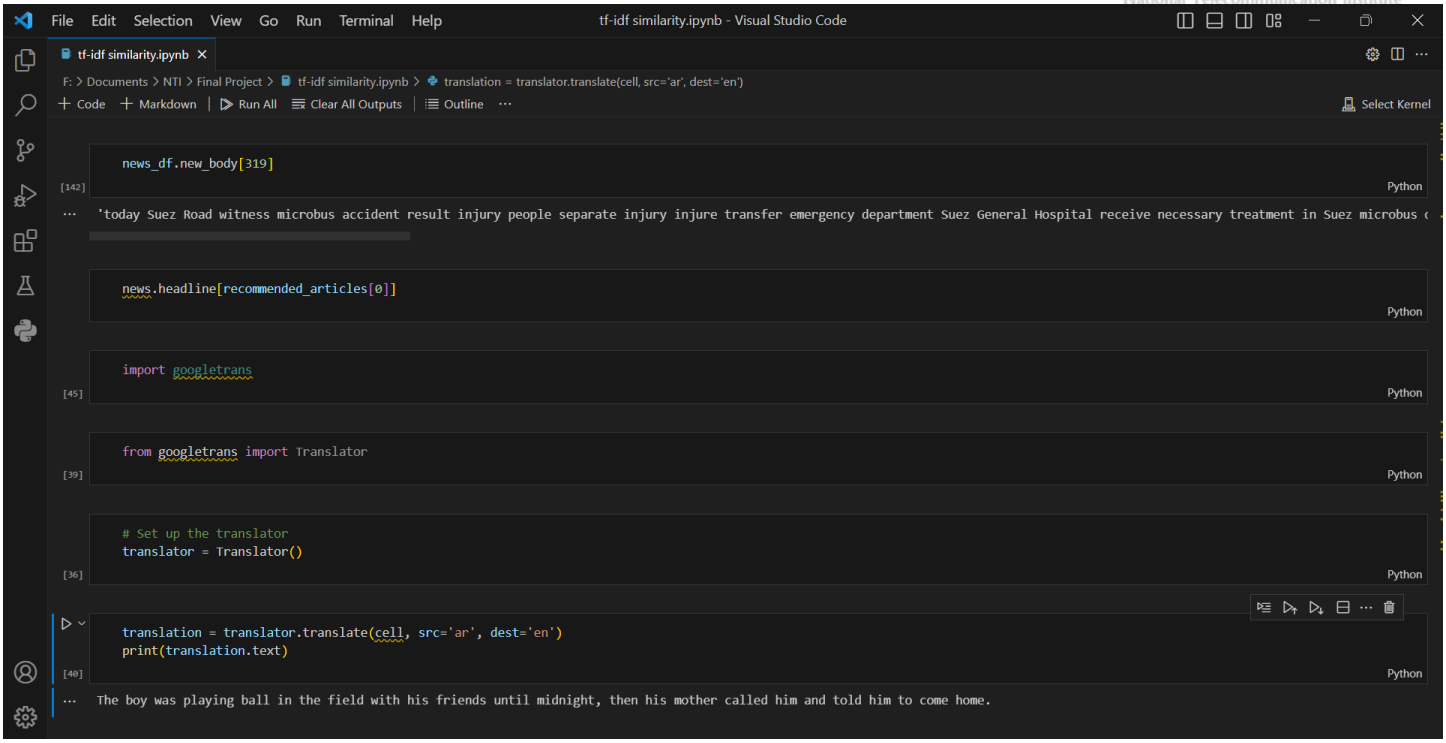


Figure 49

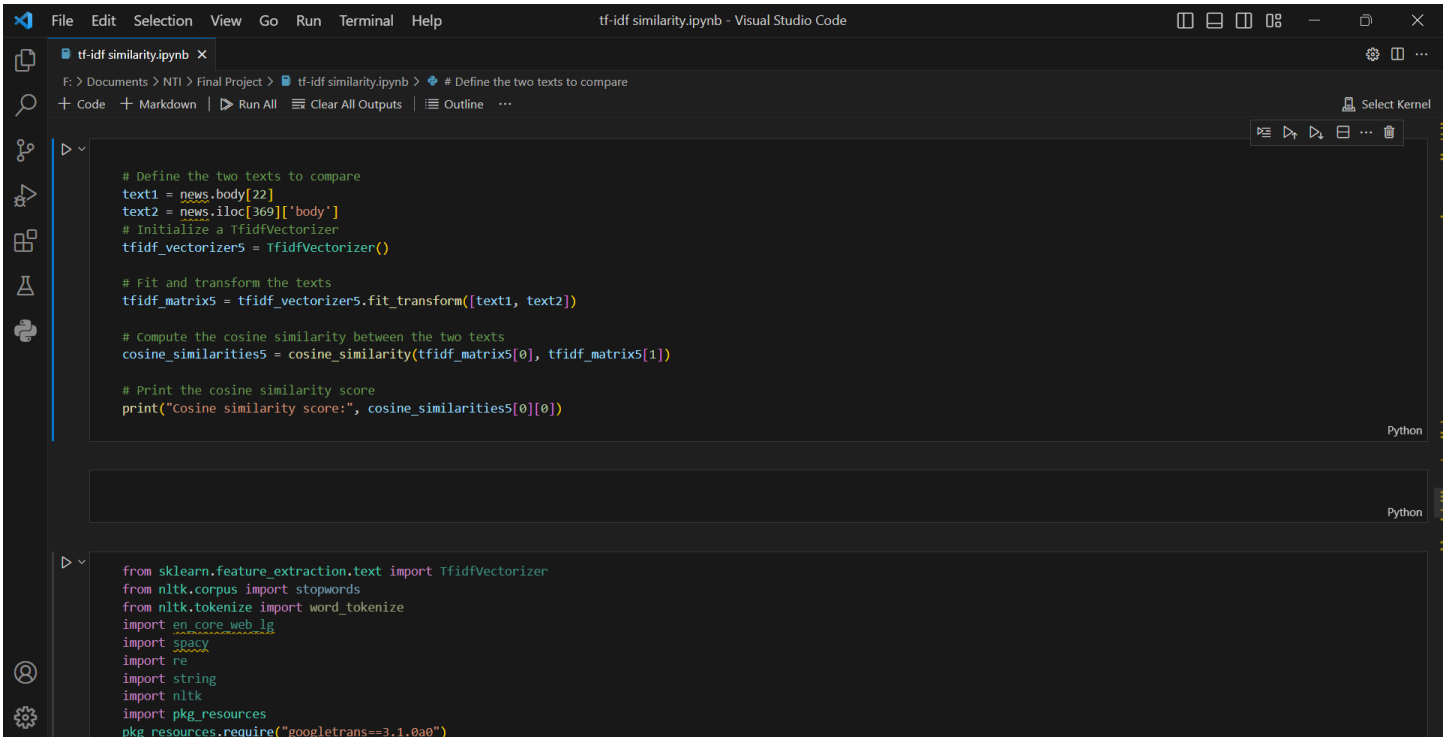


Figure 50

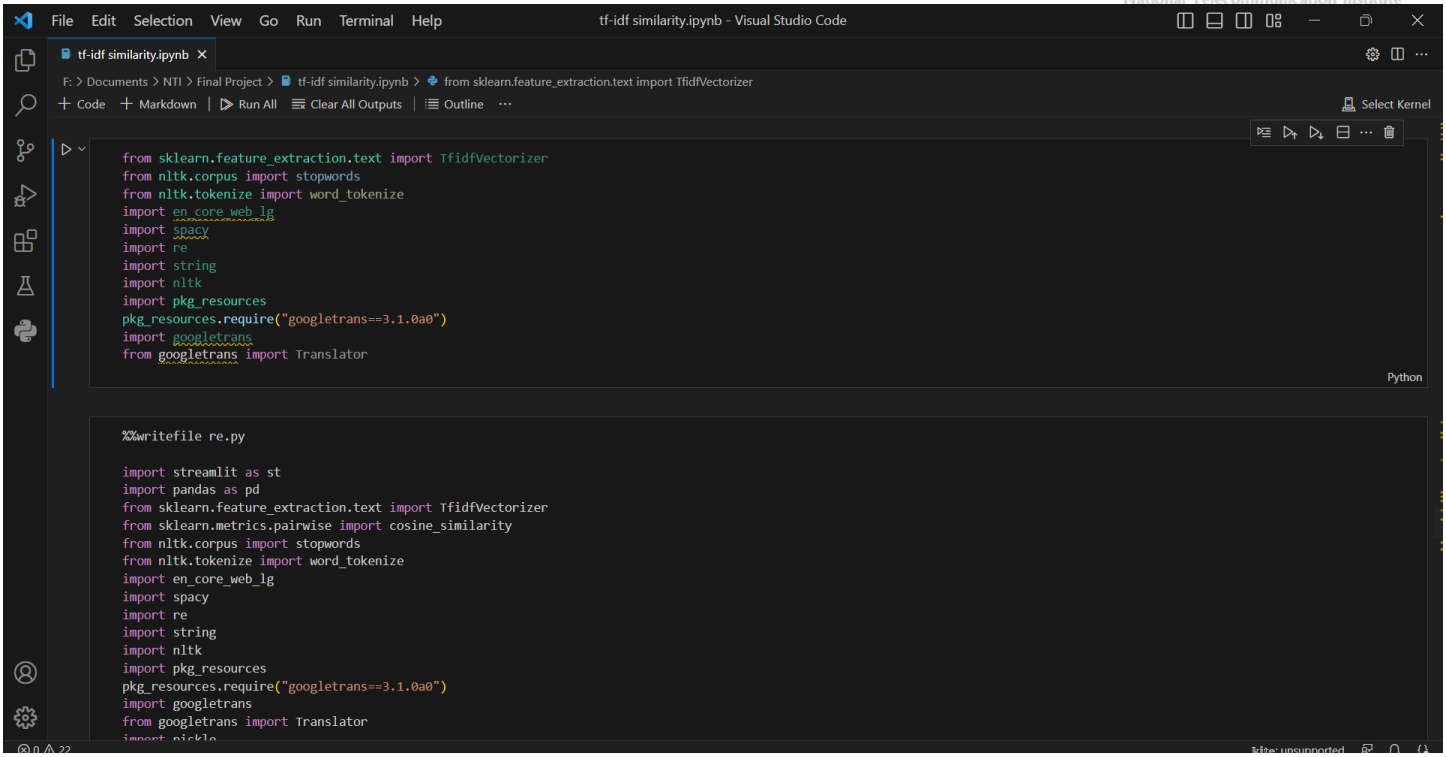


Figure 51

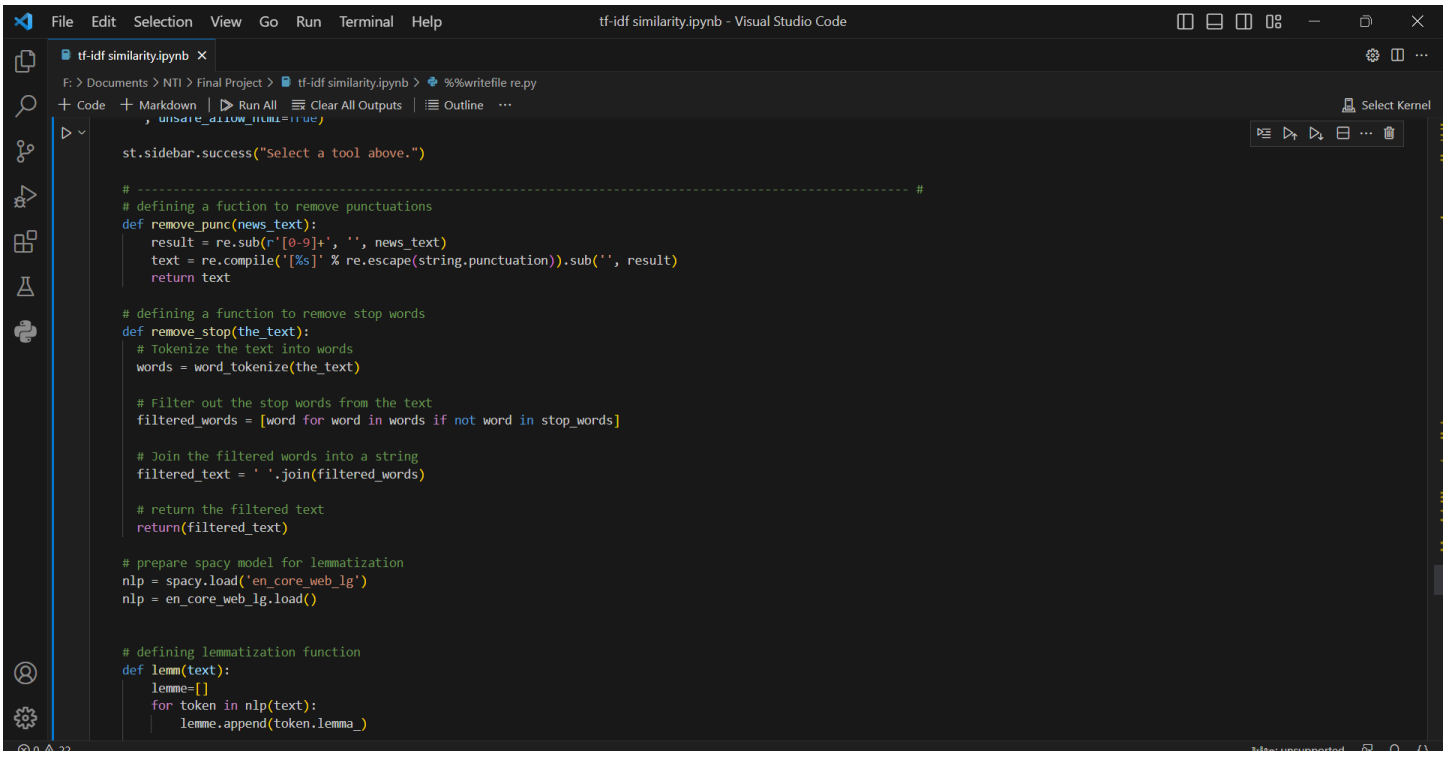


Figure 52

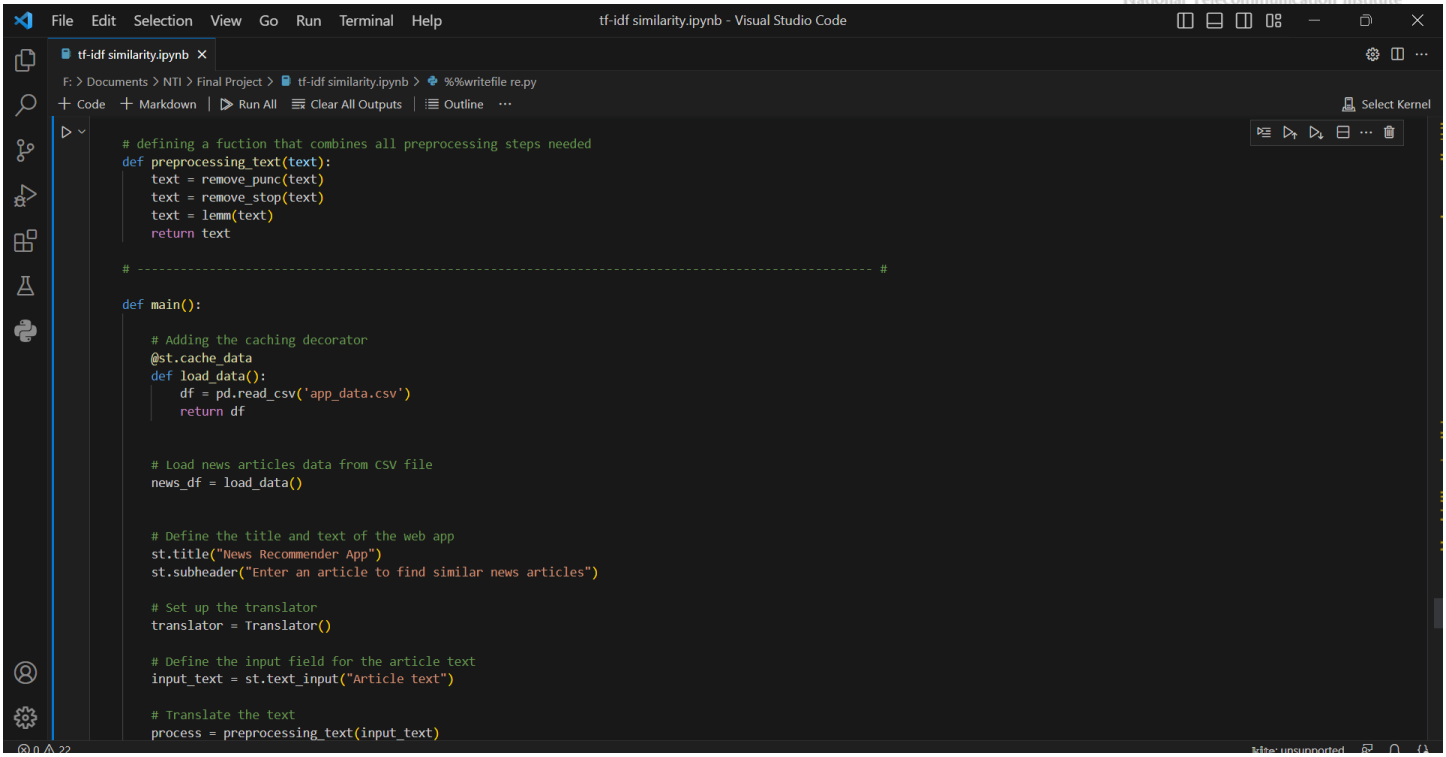


Figure 53

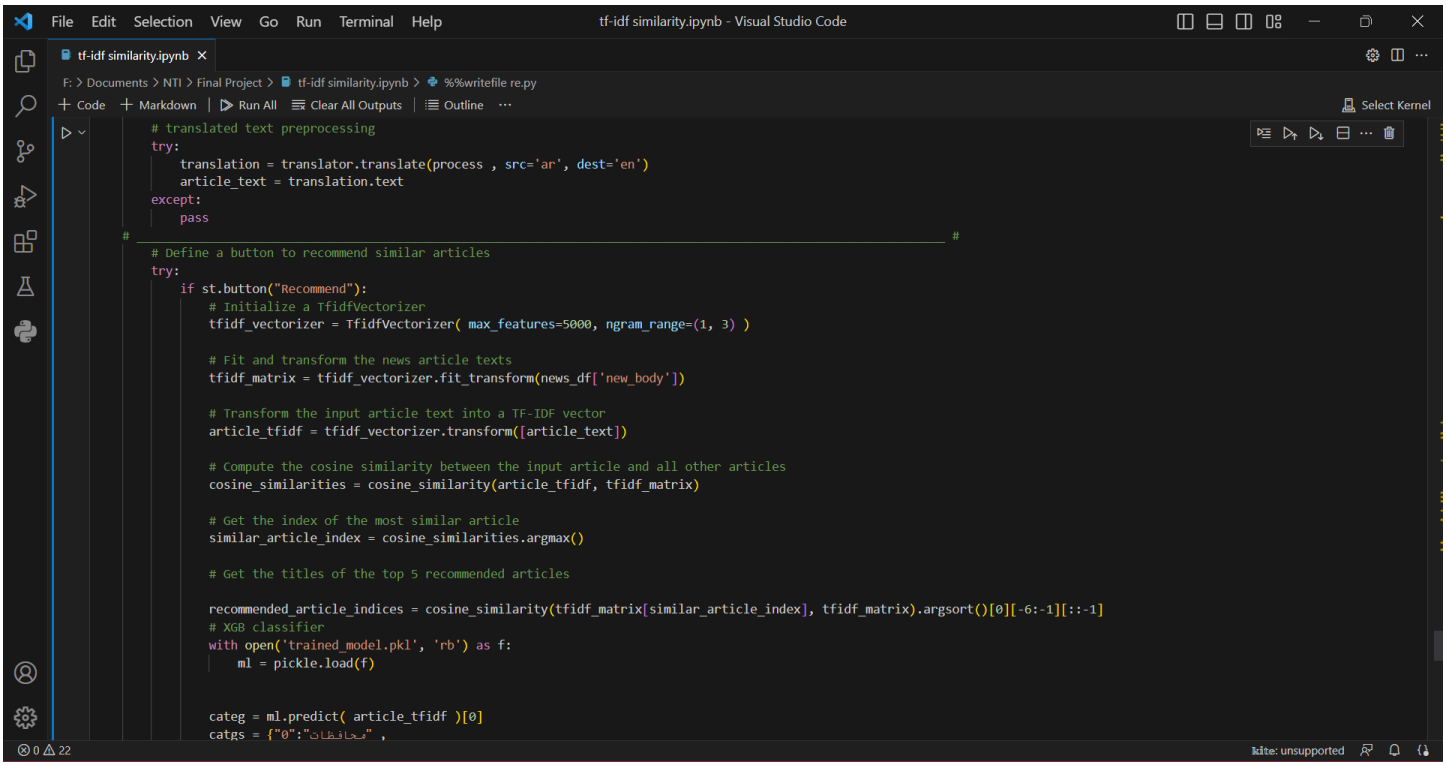


Figure 54

```
        recommended_article_indices = cosine_similarity(tfidf_matrix[similar_article_index], tfidf_matrix).argsort()[0][-6:-1][::-1]
        # XGB classifier
        with open('trained_model.pkl', 'rb') as f:
            ml = pickle.load(f)


        categ = ml.predict( article_tfidf )[0]
        catgs = {"0":"محافظات" ,
                 "1" : "رياضة",
                 "2" : "عرب و عالم",
                 "3": "أخبار" ,
                 "4" : "ثقافة وفنون",
                 "5" : "حوادث",
                 "6" : "إقتصاد"}


        # Print the recommended articles
        st.header("Recommended articles:")


        for order,indx in list(enumerate(recommended_article_indices)):
            st.write(order+1 , "- " + news_df.iloc[indx]['headline'])
            st.write("Brief:     ", news_df.iloc[indx]['briefing'])
            st.write("Go to: ", news_df.iloc[indx]['hrefs'])
            st.write(" -----/----/-----/-----/-----/-----/------/-----/-----/-----/------ ")

        st.write("Classification of News: ", catgs[str(categ)])
    except:
        st.subheader("Please enter the news text")


if __name__ == '__main__':
    main()
```

Figure 55

2.6. Summarization:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[ ] import pandas as pd
    import numpy as np
```

```
[ ] df_combined=pd.read_csv("/content/drive/MyDrive/final project_NTI/Data/final data/completeDS1.csv")
```

```
[ ] np.array(df_combined['body'][0])
```

```
array('في دورته الـ 71.ويضع المهرجان جائزة المركز الخاصة للفنان مدحت صالح، وتذهب جائزة الإبداع الفني لكل من، الكاتب مدحت العدل، الموسيقار هشام نزيه، الناقد \xa0ماجد مهرجان المركز الكاثوليكي للفنان الكبير يحيى الفخراني، جائزة الريادة السينمائية عن مجمل أعماله الفنية، وذلك القني نادر عدلي، مدير التصوير أيمن أبو المكارم، وتذهب جائزة التميز الإعلامي لـ محمود مدد.أما بالنسبة للأعمال الدرامية التي قُدمت عام 2022 فقد قررت اللجنة الثنائية للمهرجان منح جائزة أحسن ممثلة في عمل درامي، للنجمة نيللي كريم عن دورها في مسلسل فاتن أمل حربي، جائزة أفضل من: الفنانة ماجدة زكي، الفنان مجدي كامل، الفنانة التونسية علاشة\xa0ممثل عن دوره في مسلسل جزيرة غمام، وتذهب جائزة المركز التشجيعية للفنانة نارا عماد.وتشكّلت لجنة تحكيم مهرجان المركز الكاثوليكي المصري للسينما في دورته الحادية والسبعين برئاسة المخرج هاني لاشين، وعضوية كل بن أحمد، رئيس أكاديمية الفنون المونتيرة غادة جُبارة، مدير التصوير سامح سليم، الكاتبة والناقدة الفنية داهد صلاح، والموسيقار خالد الكمّار.وكتفت اللجنة الثنائية لمهرجان المركز الكاثوليكي المصري للسينما برئاسة الأب بطرس دانيال وعضوية كلاً من الأستاذ مجدي سامي والأستاذ ميشيل ماهر عن الخمسة أعلام المشارك بالمسابقة الرسمية لهذا العام بين خمسة وعشرين فيلماً عُرضت عرضاً تجارياً بدور السينما إلى جانب فيلمين عُرضا عبر المنصات، ولأول مرة تختار اللجنة الثنائية للمهرجان فيلماً تم عرضه بمهرجان قبل أن يُعرض تجارياً. وجميعهم من إنتاج عام 2022.والأفلام الخمسة التي تم اختيارها هي: "ليلة قمر 14" للمخرج هادي الباجوري. فيلم "11:11" للمخرج كريم أبو زيد. فيلم "حدث في 2 طلعت حرب" للمخرج مجدي أحمد علي. فيلم "كاملة" للمخرج جون إكرام. فيلم "الباب الأخضر" للمخرج رؤوف عبد العزيز.',
      dtype='<U1555')
```

```
[ ] df_combined['category'].value_counts()
```

```
10000    ثقافة وفنون
10000          حوادث
10000       اقتصاد
10000       محافظات
10000          رياضة
```

Figure 56

```
[ ] from transformers import AutoTokenizer, AutoModelWithLMHead
    import torch
    import nltk
    nltk.download('punkt')
    from nltk.tokenize import sent_tokenize

    [nltk_data] Downloading package punkt to /root/nltk_data...
    [nltk_data]   Unzipping tokenizers/punkt.zip.

[ ] device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
    m_name = "marefa-nlp/summarization-arabic-english-news"

[ ] tokenizer = AutoTokenizer.from_pretrained(m_name)
    model = AutoModelWithLMHead.from_pretrained(m_name).to(device)

    Downloading (…)okenizer_config.json: 100% 432/432 [00:00<00:00, 22.0kB/s]
    Downloading spiece.model: 100% 847k/847k [00:00<00:00, 2.78MB/s]
    Downloading (…)/main/tokenizer.json: 100% 1.40M/1.40M [00:00<00:00, 5.17MB/s]
    Downloading (…)cial_tokens_map.json: 100% 65.0/65.0 [00:00<00:00, 3.94kB/s]
    /usr/local/lib/python3.10/dist-packages/transformers/models/auto/modeling_auto.py:1322: FutureWarning: The class `AutoModelWithLMHead` is deprecated and will be removed in a future version. Pl
      warnings.warn(
    Downloading (…)lve/main/config.json: 100% 765/765 [00:00<00:00, 35.9kB/s]
    Downloading pytorch_model.bin: 100% 3.16G/3.16G [00:53<00:00, 61.2MB/s]
```

Figure 57

```
[ ] def get_summary(text, tokenizer, model, device="cpu", num_beams=2):
        if len(text.strip()) < 50:
            return ["Please provide more longer text"]

        text = "summarize: <paragraph> " + " <paragraph> ".join([ s.strip() for s in sent_tokenize(text) if s.strip() != ""]) + " </s>"
        text = text.strip().replace("\n","")

        tokenized_text = tokenizer.encode(text, return_tensors="pt").to(device)

        summary_ids = model.generate(
                tokenized_text,
                max_length=512,
                num_beams=num_beams,
                repetition_penalty=1.5,
                length_penalty=1.0,
                early_stopping=True
        )

        output = tokenizer.decode(summary_ids[0], skip_special_tokens=True)
        return [ s.strip() for s in output.split("<hl>") if s.strip() != "" ]
```

Figure 58

```
[ ] ## Get summaries if the are alot of lines in string

    print("Original Article:", text)
    print("\n==========\nSummary: \n")
    hls = get_summary(text, tokenizer, model, device)
    for hl in hls:
        print("\t-", hl)

    Original Article: قال محيط، في مقابلة مع سي إن بي سي عربية، إن مصر لن تطرح سندات أو صكوك دولية حتى نهاية يوليو المقبل، وفي حالة انخفاض الفائدة عالمياً من الممكن أن تطرح سندات أو صكوك دولية.

    -أكد أن مصر من الدول القليلة التي حققت معدلات نمو بلغت 3.6% خلال أزمة كورونا، مشيرا إلى أن نسبة الدين إلى الناتج المحلي من المفترض أن تكون 78%. فالاقتصاد المصري تحمل الصدمات واستمر في تحقيق معدلات نمو رغم الأزمات -على حد تعبيره.

    وأكد محيط أن ارتفاع أسعار الغذاء عالميا زاد كلفة استيراد السلع من الخارج، مشيرا إلى أن الاقتصاد المصري أصبح تحت ضغط بسبب تداعيات الحرب الروسية الأوكرانية ، موضحا أن 80% من واردات مصر من القمح تأتي من دولتي روسيا وأوكرانيا.

    ==========
    Summary:

            - وقد قال وزير الاقتصاد المصري أن مصر لن تطرح سندات أو صكوك دولية حتى نهاية يوليو المقبل.
```

Figure 59

CONCLUSION

In conclusion, the News Recommendation System project has demonstrated the feasibility and effectiveness of using machine learning techniques to recommend personalized news articles to users. By analyzing user behavior patterns and preferences, we were able to create a system that delivers relevant news articles in real-time, improving the user experience and increasing engagement. The project also highlighted the importance of incorporating ethical considerations into the development of such systems. We ensured that our system does not perpetuate filter bubbles or bias, and that user data is kept secure and private. Moving forward, we recommend further exploration of potential applications and improvements to the system, as discussed in the previous section. We also encourage the adoption of similar ethical considerations in the development of news recommendation systems. Overall, the News Recommendation System project has provided valuable insights into the possibilities of using machine learning to personalize the news and improve the way we stay informed. We hope that our work can inspire further innovation in this field and contribute to a more informed and diverse society.

REFERENCES

1- What is a recommendation system? NVIDIA Data Science Glossary. Available at: https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/ (Accessed: 09 May 2023).

2- *What is scraping: About price & web scraping tools: Imperva* (2019) *Learning Center*. Available at: https://www.imperva.com/learn/application-security/web-scraping-attack/ (Accessed: 10 May 2023).

3- *Text classification: What it is and why it matters* (no date) *MonkeyLearn*. Available at: https://monkeylearn.com/text-classification/ (Accessed: 09 May 2023).

4- *What is summarization? - hugging face* (no date) *What is Summarization? - Hugging Face*. Available at: https://huggingface.co/tasks/summarization (Accessed: 09 May 2023).

5- *Streamlit docs* (no date) *Streamlit documentation*. Available at: https://docs.streamlit.io/ (Accessed: 09 May 2023).

6- https://www.edlitera.com/en/blog/posts/text-summarization-nlp-how-to