```
pip install pandas nltk spacy gensim scikit-learn

Requirement already satisfied: pandas in
/usr/local/lib/python3.10/dist-packages (1.5.3)
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-
packages (3.8.1)
Requirement already satisfied: spacy in
/usr/local/lib/python3.10/dist-packages (3.6.1)
Requirement already satisfied: gensim in
/usr/local/lib/python3.10/dist-packages (4.3.2)
Requirement already satisfied: scikit-learn in
/usr/local/lib/python3.10/dist-packages (1.2.2)
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2023.3.post1)
Requirement already satisfied: numpy>=1.21.0 in
/usr/local/lib/python3.10/dist-packages (from pandas) (1.23.5)
Requirement already satisfied: click in
/usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in
/usr/local/lib/python3.10/dist-packages (from nltk) (1.3.2)
Requirement already satisfied: regex>=2021.8.3 in
/usr/local/lib/python3.10/dist-packages (from nltk) (2023.6.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-
packages (from nltk) (4.66.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in
/usr/local/lib/python3.10/dist-packages (from spacy) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/usr/local/lib/python3.10/dist-packages (from spacy) (3.0.9)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in
/usr/local/lib/python3.10/dist-packages (from spacy) (8.1.12)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in
/usr/local/lib/python3.10/dist-packages (from spacy) (1.1.2)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.0.10)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (0.9.0)
Requirement already satisfied: pathy>=0.10.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (0.10.3)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in
/usr/local/lib/python3.10/dist-packages (from spacy) (6.4.0)
```

```
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (2.31.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in
/usr/local/lib/python3.10/dist-packages (from spacy) (1.10.13)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.10/dist-packages (from spacy) (3.1.2)
Requirement already satisfied: setuptools in
/usr/local/lib/python3.10/dist-packages (from spacy) (67.7.2)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (23.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in
/usr/local/lib/python3.10/dist-packages (from spacy) (3.3.0)
Requirement already satisfied: scipy>=1.7.0 in
/usr/local/lib/python3.10/dist-packages (from gensim) (1.11.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn) (3.2.0)
Requirement already satisfied: typing-extensions>=4.2.0 in
/usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!
=1.8.1,<3.0.0,>=1.7.4->spacy) (4.5.0)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1-
>pandas) (1.16.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0-
>spacy) (2023.7.22)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in
/usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8-
>spacy) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in
/usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8-
>spacy) (0.1.3)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->spacy) (2.1.3)
```

```
import pandas as pd
import nltk
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer
import spacy
from gensim.models import Word2Vec
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
# Assuming BBC_DATA.csv is in the same directory
dataset = pd.read_csv('BBC_DATA.csv')

dataset.head()
```

```
   ArticleId                                               Text
Category
0       1833  worldcom ex-boss launches defence lawyers defe...
business
1        154  german business confidence slides german busin...
business
2       1101  bbc poll indicates economic gloom citizens in ...
business
3       1976  lifestyle  governs mobile choice  faster  bett...
tech
4        917  enron bosses in $168m payout eighteen former e...
business
```

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

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

```
True
```

```
sample_article = dataset['Text'].iloc[0]

# Tokenization
words = word_tokenize(sample_article)
sentences = sent_tokenize(sample_article)

print(words)
print(sentences)
```

```
['worldcom', 'ex-boss', 'launches', 'defence', 'lawyers', 'defending',
'former', 'worldcom', 'chief', 'bernie', 'ebbers', 'against', 'a',
'battery', 'of', 'fraud', 'charges', 'have', 'called', 'a', 'company',
'whistleblower', 'as', 'their', 'first', 'witness', '.', 'cynthia',
'cooper', 'worldcom', 's', 'ex-head', 'of', 'internal', 'accounting',
'alerted', 'directors', 'to', 'irregular', 'accounting', 'practices',
'at', 'the', 'us', 'telecoms', 'giant', 'in', '2002.', 'her',
'warnings', 'led', 'to', 'the', 'collapse', 'of', 'the', 'firm',
'following', 'the', 'discovery', 'of', 'an', '$', '11bn', '(',
'£5.7bn', ')', 'accounting', 'fraud', '.', 'mr', 'ebbers', 'has',
'pleaded', 'not', 'guilty', 'to', 'charges', 'of', 'fraud', 'and',
'conspiracy', '.', 'prosecution', 'lawyers', 'have', 'argued', 'that',
'mr', 'ebbers', 'orchestrated', 'a', 'series', 'of', 'accounting',
'tricks', 'at', 'worldcom', 'ordering', 'employees', 'to', 'hide',
'expenses', 'and', 'inflate', 'revenues', 'to', 'meet', 'wall',
'street', 'earnings', 'estimates', '.', 'but', 'ms', 'cooper', 'who',
```

```
'now', 'runs', 'her', 'own', 'consulting', 'business', 'told', 'a',
'jury', 'in', 'new', 'york', 'on', 'wednesday', 'that', 'external',
'auditors', 'arthur', 'andersen', 'had', 'approved', 'worldcom', 's',
'accounting', 'in', 'early', '2001', 'and', '2002.', 'she', 'said',
'andersen', 'had', 'given', 'a', 'green', 'light', 'to', 'the',
'procedures', 'and', 'practices', 'used', 'by', 'worldcom', '.', 'mr',
'ebber', 's', 'lawyers', 'have', 'said', 'he', 'was', 'unaware', 'of',
'the', 'fraud', 'arguing', 'that', 'auditors', 'did', 'not', 'alert',
'him', 'to', 'any', 'problems', '.', 'ms', 'cooper', 'also', 'said',
'that', 'during', 'shareholder', 'meetings', 'mr', 'ebbers', 'often',
'passed', 'over', 'technical', 'questions', 'to', 'the', 'company',
's', 'finance', 'chief', 'giving', 'only', 'brief', 'answers',
'himself', '.', 'the', 'prosecution', 's', 'star', 'witness',
'former', 'worldcom', 'financial', 'chief', 'scott', 'sullivan',
'has', 'said', 'that', 'mr', 'ebbers', 'ordered', 'accounting',
'adjustments', 'at', 'the', 'firm', 'telling', 'him', 'to', 'hit',
'our', 'books', '.', 'however', 'ms', 'cooper', 'said', 'mr',
'sullivan', 'had', 'not', 'mentioned', 'anything', 'uncomfortable',
'about', 'worldcom', 's', 'accounting', 'during', 'a', '2001',
'audit', 'committee', 'meeting', '.', 'mr', 'ebbers', 'could', 'face',
'a', 'jail', 'sentence', 'of', '85', 'years', 'if', 'convicted', 'of',
'all', 'the', 'charges', 'he', 'is', 'facing', '.', 'worldcom',
'emerged', 'from', 'bankruptcy', 'protection', 'in', '2004', 'and',
'is', 'now', 'known', 'as', 'mci', '.', 'last', 'week', 'mci',
'agreed', 'to', 'a', 'buyout', 'by', 'verizon', 'communications',
'in', 'a', 'deal', 'valued', 'at', '$', '6.75bn', '.']
['worldcom ex-boss launches defence lawyers defending former worldcom
chief bernie ebbers against a battery of fraud charges have called a
company whistleblower as their first witness.', 'cynthia cooper
worldcom s ex-head of internal accounting  alerted directors to
irregular accounting practices at the us telecoms giant in 2002. her
warnings led to the collapse of the firm following the discovery of an
$11bn (£5.7bn) accounting fraud.', 'mr ebbers has pleaded not guilty
to charges of fraud and conspiracy.', 'prosecution lawyers have argued
that mr ebbers orchestrated a series of accounting tricks at worldcom
ordering employees to hide expenses and inflate revenues to meet wall
street earnings estimates.', 'but ms cooper  who now runs her own
consulting business  told a jury in new york on wednesday that
external auditors arthur andersen had approved worldcom s accounting
in early 2001 and 2002. she said andersen had given a  green light  to
the procedures and practices used by worldcom.', 'mr ebber s lawyers
have said he was unaware of the fraud  arguing that auditors did not
alert him to any problems.', 'ms cooper also said that during
shareholder meetings mr ebbers often passed over technical questions
to the company s finance chief  giving only  brief  answers himself.',
'the prosecution s star witness  former worldcom financial chief scott
sullivan  has said that mr ebbers ordered accounting adjustments at
the firm  telling him to  hit our books .', 'however  ms cooper said
mr sullivan had not mentioned  anything uncomfortable  about worldcom
```

```
s accounting during a 2001 audit committee meeting.', 'mr ebbers could face a jail sentence of 85 years if convicted of all the charges he is facing.', 'worldcom emerged from bankruptcy protection in 2004  and is now known as mci.', 'last week  mci agreed to a buyout by verizon communications in a deal valued at $6.75bn.']
```

```
import nltk
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
```

```
True
```

```
# Stemming
porter_stemmer = PorterStemmer()
stemmed_words = [porter_stemmer.stem(word) for word in words]

# Lemmatization
lemmatizer = WordNetLemmatizer()
lemmatized_words = [lemmatizer.lemmatize(word) for word in words]

print(stemmed_words)
print(lemmatized_words)
```

```
['worldcom', 'ex-boss', 'launch', 'defenc', 'lawyer', 'defend',
'former', 'worldcom', 'chief', 'berni', 'ebber', 'against', 'a',
'batteri', 'of', 'fraud', 'charg', 'have', 'call', 'a', 'compani',
'whistleblow', 'as', 'their', 'first', 'wit', '.', 'cynthia',
'cooper', 'worldcom', 's', 'ex-head', 'of', 'intern', 'account',
'alert', 'director', 'to', 'irregular', 'account', 'practic', 'at',
'the', 'us', 'telecom', 'giant', 'in', '2002.', 'her', 'warn', 'led',
'to', 'the', 'collaps', 'of', 'the', 'firm', 'follow', 'the',
'discoveri', 'of', 'an', '$', '11bn', '(', '£5.7bn', ')', 'account',
'fraud', '.', 'mr', 'ebber', 'ha', 'plead', 'not', 'guilti', 'to',
'charg', 'of', 'fraud', 'and', 'conspiraci', '.', 'prosecut',
'lawyer', 'have', 'argu', 'that', 'mr', 'ebber', 'orchestr', 'a',
'seri', 'of', 'account', 'trick', 'at', 'worldcom', 'order',
'employe', 'to', 'hide', 'expens', 'and', 'inflat', 'revenu', 'to',
'meet', 'wall', 'street', 'earn', 'estim', '.', 'but', 'ms', 'cooper',
'who', 'now', 'run', 'her', 'own', 'consult', 'busi', 'told', 'a',
'juri', 'in', 'new', 'york', 'on', 'wednesday', 'that', 'extern',
'auditor', 'arthur', 'andersen', 'had', 'approv', 'worldcom', 's',
'account', 'in', 'earli', '2001', 'and', '2002.', 'she', 'said',
'andersen', 'had', 'given', 'a', 'green', 'light', 'to', 'the',
'procedur', 'and', 'practic', 'use', 'by', 'worldcom', '.', 'mr',
'ebber', 's', 'lawyer', 'have', 'said', 'he', 'wa', 'unawar', 'of',
'the', 'fraud', 'argu', 'that', 'auditor', 'did', 'not', 'alert',
'him', 'to', 'ani', 'problem', '.', 'ms', 'cooper', 'also', 'said',
'that', 'dure', 'sharehold', 'meet', 'mr', 'ebber', 'often', 'pass',
'over', 'technic', 'question', 'to', 'the', 'compani', 's', 'financ',
'chief', 'give', 'onli', 'brief', 'answer', 'himself', '.', 'the',
```

```
'prosecut', 's', 'star', 'wit', 'former', 'worldcom', 'financi',
'chief', 'scott', 'sullivan', 'ha', 'said', 'that', 'mr', 'ebber',
'order', 'account', 'adjust', 'at', 'the', 'firm', 'tell', 'him',
'to', 'hit', 'our', 'book', '.', 'howev', 'ms', 'cooper', 'said',
'mr', 'sullivan', 'had', 'not', 'mention', 'anyth', 'uncomfort',
'about', 'worldcom', 's', 'account', 'dure', 'a', '2001', 'audit',
'committe', 'meet', '.', 'mr', 'ebber', 'could', 'face', 'a', 'jail',
'sentenc', 'of', '85', 'year', 'if', 'convict', 'of', 'all', 'the',
'charg', 'he', 'is', 'face', '.', 'worldcom', 'emerg', 'from',
'bankruptci', 'protect', 'in', '2004', 'and', 'is', 'now', 'known',
'as', 'mci', '.', 'last', 'week', 'mci', 'agre', 'to', 'a', 'buyout',
'by', 'verizon', 'commun', 'in', 'a', 'deal', 'valu', 'at', '$',
'6.75bn', '.']
['worldcom', 'ex-boss', 'launch', 'defence', 'lawyer', 'defending',
'former', 'worldcom', 'chief', 'bernie', 'ebbers', 'against', 'a',
'battery', 'of', 'fraud', 'charge', 'have', 'called', 'a', 'company',
'whistleblower', 'a', 'their', 'first', 'witness', '.', 'cynthia',
'cooper', 'worldcom', 's', 'ex-head', 'of', 'internal', 'accounting',
'alerted', 'director', 'to', 'irregular', 'accounting', 'practice',
'at', 'the', 'u', 'telecom', 'giant', 'in', '2002.', 'her', 'warning',
'led', 'to', 'the', 'collapse', 'of', 'the', 'firm', 'following',
'the', 'discovery', 'of', 'an', '$', '11bn', '(', '£5.7bn', ')',
'accounting', 'fraud', '.', 'mr', 'ebbers', 'ha', 'pleaded', 'not',
'guilty', 'to', 'charge', 'of', 'fraud', 'and', 'conspiracy', '.',
'prosecution', 'lawyer', 'have', 'argued', 'that', 'mr', 'ebbers',
'orchestrated', 'a', 'series', 'of', 'accounting', 'trick', 'at',
'worldcom', 'ordering', 'employee', 'to', 'hide', 'expense', 'and',
'inflate', 'revenue', 'to', 'meet', 'wall', 'street', 'earnings',
'estimate', '.', 'but', 'm', 'cooper', 'who', 'now', 'run', 'her',
'own', 'consulting', 'business', 'told', 'a', 'jury', 'in', 'new',
'york', 'on', 'wednesday', 'that', 'external', 'auditor', 'arthur',
'andersen', 'had', 'approved', 'worldcom', 's', 'accounting', 'in',
'early', '2001', 'and', '2002.', 'she', 'said', 'andersen', 'had',
'given', 'a', 'green', 'light', 'to', 'the', 'procedure', 'and',
'practice', 'used', 'by', 'worldcom', '.', 'mr', 'ebber', 's',
'lawyer', 'have', 'said', 'he', 'wa', 'unaware', 'of', 'the', 'fraud',
'arguing', 'that', 'auditor', 'did', 'not', 'alert', 'him', 'to',
'any', 'problem', '.', 'm', 'cooper', 'also', 'said', 'that',
'during', 'shareholder', 'meeting', 'mr', 'ebbers', 'often', 'passed',
'over', 'technical', 'question', 'to', 'the', 'company', 's',
'finance', 'chief', 'giving', 'only', 'brief', 'answer', 'himself',
'.', 'the', 'prosecution', 's', 'star', 'witness', 'former',
'worldcom', 'financial', 'chief', 'scott', 'sullivan', 'ha', 'said',
'that', 'mr', 'ebbers', 'ordered', 'accounting', 'adjustment', 'at',
'the', 'firm', 'telling', 'him', 'to', 'hit', 'our', 'book', '.',
'however', 'm', 'cooper', 'said', 'mr', 'sullivan', 'had', 'not',
'mentioned', 'anything', 'uncomfortable', 'about', 'worldcom', 's',
'accounting', 'during', 'a', '2001', 'audit', 'committee', 'meeting',
'.', 'mr', 'ebbers', 'could', 'face', 'a', 'jail', 'sentence', 'of',
```

```
'85', 'year', 'if', 'convicted', 'of', 'all', 'the', 'charge', 'he',
'is', 'facing', '.', 'worldcom', 'emerged', 'from', 'bankruptcy',
'protection', 'in', '2004', 'and', 'is', 'now', 'known', 'a', 'mci',
'.', 'last', 'week', 'mci', 'agreed', 'to', 'a', 'buyout', 'by',
'verizon', 'communication', 'in', 'a', 'deal', 'valued', 'at', '$',
'6.75bn', '.']

# Load SpaCy's pre-trained model
nlp = spacy.load('en_core_web_sm')

# Process the sample article
doc = nlp(sample_article)

# Extract named entities
named_entities = [(ent.text, ent.label_) for ent in doc.ents]
print(named_entities)

[('worldcom ex-boss', 'PERSON'), ('worldcom', 'ORG'), ('bernie',
'PERSON'), ('first', 'ORDINAL'), ('cynthia cooper  worldcom s ex-
head', 'PERSON'), ('us', 'GPE'), ('2002', 'DATE'), ('5.7bn', 'MONEY'),
('worldcom', 'ORG'), ('new york', 'GPE'), ('wednesday', 'DATE'),
('arthur andersen', 'PERSON'), ('worldcom', 'ORG'), ('early 2001 and
2002', 'DATE'), ('worldcom', 'ORG'), ('cooper', 'PERSON'), ('worldcom
financial', 'ORG'), ('scott sullivan', 'PERSON'), ('sullivan',
'PERSON'), ('2001', 'DATE'), ('85 years', 'DATE'), ('worldcom',
'ORG'), ('2004', 'DATE'), ('mci', 'ORG'), ('last week', 'DATE'),
('mci', 'ORG'), ('6.75bn', 'MONEY')]

# Tokenize the entire dataset
tokenized_data = [word_tokenize(article) for article in
dataset['Text']]

# Train Word2Vec model
word2vec_model = Word2Vec(sentences=tokenized_data, vector_size=100,
window=5, min_count=1, workers=4)


# Get a word from the vocabulary of the Word2Vec model
sample_word = word2vec_model.wv.index_to_key[0]  # Choose index 0 or
any other valid index

# Get vector representation of the chosen word
sample_word_vector = word2vec_model.wv[sample_word]
print(sample_word_vector)
print(sample_word)
print(word2vec_model)

[ 2.7879721e-02  1.3422313e-03  1.6377891e+00  1.0293640e+00
 -1.0290486e+00 -1.3601092e+00  1.2974035e+00  9.9968672e-01
 -8.5291559e-01 -1.7252359e+00  5.7322627e-01 -6.5874749e-01
 -1.5534254e+00  1.3109279e-01 -7.6012686e-03 -1.0684609e-01
```

```
  5.5359185e-01  4.6730842e-02 -4.6097383e-01 -1.0784883e+00
  1.2382365e+00 -3.4004995e-01  1.8526311e+00 -8.9855832e-01
  8.0788380e-01 -2.3810017e-01 -7.3202866e-01  8.5141951e-01
 -6.0484940e-01 -1.3663299e-01  1.7441951e-01  4.1195318e-01
  2.2754297e-01 -1.4277290e+00  2.6205432e-01 -2.3459113e-01
  3.2779139e-01 -1.0027425e+00 -2.6524103e-01 -9.7996330e-01
 -5.3939018e-02 -7.3855877e-01 -3.2385218e-01 -4.5562556e-01
  3.3828276e-01 -9.2608529e-01 -7.6590401e-01  5.1656097e-01
  5.9468585e-01  8.1698108e-01 -2.2878036e-01 -5.8444786e-01
 -5.9562641e-01  2.0109849e-01 -4.9469033e-01 -3.5529356e-02
 -7.3464908e-02 -3.1536564e-01 -2.6733142e-01  7.5271177e-01
  2.9261222e-01  5.9244394e-01  3.9192730e-01 -1.2188308e-01
 -4.8659769e-01  9.6173590e-01 -7.2267249e-02  1.5975096e+00
 -4.6635282e-01  1.2832092e-01 -5.1829058e-01  1.7379866e+00
  6.9617790e-01 -5.4129392e-02  1.5675797e+00  7.4697667e-01
  4.9156135e-01 -2.3404866e-01 -5.3845477e-01 -1.3944783e+00
  1.6136480e-02 -8.8061994e-01  1.6445722e-01  1.2899598e+00
 -8.1304860e-01 -1.2082011e-01  6.8087047e-01 -4.7792643e-01
 -2.0605896e-01  5.2059090e-01 -8.6493358e-02 -5.8890021e-01
  2.4558719e-01 -3.5645044e-01  1.6963493e+00  1.1045080e+00
  3.1815717e-01 -1.4082348e+00 -2.2831902e-02  4.1127384e-01]
the
Word2Vec<vocab=28178, vector_size=100, alpha=0.025>

# Initialize TfidfVectorizer
tfidf_vectorizer = TfidfVectorizer()

# Fit and transform the dataset
tfidf_matrix = tfidf_vectorizer.fit_transform(dataset['Text'])

# Calculate cosine similarity between two news articles
cosine_similarity = (tfidf_matrix * tfidf_matrix.T).toarray()
print(tfidf_matrix)
print(cosine_similarity)

  (0, 1009)     0.05643509689668293
  (0, 23572)    0.05171683198420305
  (0, 6472)     0.02902507640882694
  (0, 5342)     0.044434208471766566
  (0, 23661)    0.05875637443991937
  (0, 4150)     0.0601801944482067
  (0, 1652)     0.034102087269467486
  (0, 24104)    0.02383356638673293
  (0, 13075)    0.017706106467856114
  (0, 14251)    0.10916318514966579
  (0, 12865)    0.03184197913561059
  (0, 407) 0.025019777226527817
  (0, 17579)    0.039549514785291666
  (0, 2875)     0.04289120555750058
  (0, 9615)     0.012344388705610583
```

```
  (0, 8036)      0.043630523793597796
  (0, 8748)      0.038172940645020706
  (0, 12224)     0.021277420879685587
  (0, 1792)      0.017252502375521908
  (0, 5810)      0.049059642984184255
  (0, 11448)     0.018347791308608738
  (0, 24617)     0.02055116391848126
  (0, 1076)      0.048608990336853036
  (0, 19913)     0.049533415911568075
  (0, 12316)     0.04415847017280601
  :   :
  (1489, 11033)  0.08625548429839593
  (1489, 15792)  0.020993536539211335
  (1489, 16046)  0.017799215951455238
  (1489, 17416)  0.03244209969602519
  (1489, 4156)   0.07275128184524933
  (1489, 23499)  0.08217323575020663
  (1489, 17420)  0.057758102195817854
  (1489, 19389)  0.022384922212533638
  (1489, 10506)  0.030646580533812535
  (1489, 15777)  0.03189506635943616
  (1489, 15313)  0.03465549786303456
  (1489, 15546)  0.03931849073930841
  (1489, 22278)  0.03424608962685311
  (1489, 1998)   0.0291486251801776
  (1489, 15497)  0.0430417823480323
  (1489, 10702)  0.011509131115391957
  (1489, 9167)   0.02691202781223377
  (1489, 11619)  0.03878671010933062
  (1489, 22285)  0.24047370574433408
  (1489, 22511)  0.08691960927507664
  (1489, 9171)   0.038599233268436596
  (1489, 22294)  0.016037927102274776
  (1489, 2378)   0.024793810877961175
  (1489, 10734)  0.012414143702070527
  (1489, 15683)  0.04845088020646398
[[1.         0.07875932 0.0791499  ... 0.09308276 0.09546118
0.06290533]
 [0.07875932 1.         0.2000389  ... 0.16967046 0.16940074
0.10128641]
 [0.0791499  0.2000389  1.         ... 0.1568732  0.21010698
0.13737273]
 ...
 [0.09308276 0.16967046 0.1568732  ... 1.         0.17307143
0.09767576]
 [0.09546118 0.16940074 0.21010698 ... 0.17307143 1.
0.13288278]
 [0.06290533 0.10128641 0.13737273 ... 0.09767576 0.13288278 1.
]]
```