```
# Task 1: Import necessary libraries

import pandas as pd
import nltk
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer
import spacy
from spacy import displacy
from gensim.models import Word2Vec
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
nltk.download('punkt')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
```

```
True
```

```
# Task 2: Load the dataset
file_path = "/content/drive/MyDrive/Colab Notebooks/BBC_DATA.csv"
df = pd.read_csv(file_path)
display(df.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
```

```
# Task 3: Tokenization with NLTK

sample_article = df.iloc[0, 1]  # Assuming the 'Text' column contains
the news articles
tokens_words = word_tokenize(sample_article)
tokens_sentences = sent_tokenize(sample_article)

print("\nTokenization with NLTK:")
print("Tokenized Words:", tokens_words)
print("Tokenized Sentences:", tokens_sentences)
```

```

Tokenization with NLTK:
```

```
Tokenized Words: ['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',
'.']
Tokenized 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.']
```

```
# Task 4: Stemming and Lemmatization with NLTK
porter_stemmer = PorterStemmer()
wordnet_lemmatizer = WordNetLemmatizer()

stemmed_words = [porter_stemmer.stem(word) for word in tokens_words]
lemmatized_words = [wordnet_lemmatizer.lemmatize(word) for word in
tokens_words]

print("\nStemming with NLTK:", stemmed_words)
print("Lemmatization with NLTK:", lemmatized_words)
```

```
Stemming with NLTK: ['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', '$',
'6.75bn', '.']
Lemmatization with NLTK: ['worldcom', 'ex-boss', 'launch', 'defence',
'lawyer', 'defending', 'former', 'worldcom', 'chief', 'bernie',
'ebbers', 'against', 'a', 'battery', 'of', 'fraud', 'charge', 'have',
'called', 'a', 'company', 'whistleblower', 'a', 'their', 'first',
'witness', '.', 'cynthia', 'cooper', 'worldcom', 's', 'ex-head', 'of',
'internal', 'accounting', 'alerted', 'director', 'to', 'irregular',
'accounting', 'practice', 'at', 'the', 'u', 'telecom', 'giant', 'in',
'2002.', 'her', 'warning', 'led', 'to', 'the', 'collapse', 'of',
'the', 'firm', 'following', 'the', 'discovery', 'of', 'an', '$',
'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', '.']
```

```
# Task 5: Named Entity Recognition with SpaCy
nlp = spacy.load("en_core_web_sm")
doc = nlp(sample_article)
displacy.render(doc, style="ent", jupyter=True)
```

```
<IPython.core.display.HTML object>
```

```
# Task 6: Word2Vec with gensim
sentences = [word_tokenize(article) for article in df['Text']]
word2vec_model = Word2Vec(sentences, vector_size=100, window=5,
min_count=1, workers=4)

# Get vector representation of a sample word
sample_word_vector = word2vec_model.wv['sample']
print("\nWord2Vec with gensim - Vector representation of 'sample':",
sample_word_vector)
```

```
Word2Vec with gensim - Vector representation of 'sample': [-0.01151732
0.0864388  -0.00820681  0.01198654  0.04814653 -0.10581327
 -0.02857972  0.13516779 -0.06370393 -0.0513782  -0.02205362 -
0.06476305
  0.00464555  0.00840096  0.06711    -0.06885602  0.07714604 -
0.06072763
  0.03082088 -0.10884771  0.05934846  0.00984306  0.06359379 -
0.0338015
  0.01568069 -0.046829   -0.11158412  0.01679148 -0.05996404 -
0.00053319
  0.09854335 -0.0388101   0.02445992 -0.12494642  0.00191479
0.05541306
  0.01439658 -0.01624904 -0.02484631 -0.09023548 -0.01364749 -
```

```
0.05923644
 -0.04967679  0.04008827  0.0537195   0.02267633 -0.07054133
0.00340478
  0.00569881  0.02455191  0.05622242 -0.08629406 -0.0229742  -
0.02066166
  0.01637956  0.01121017  0.06532785  0.03676561 -0.07041118 -
0.01215285
  0.01180393  0.01714104 -0.01940402  0.03000361 -0.01001607
0.06273773
  0.0072099   0.03676554 -0.05372756  0.04453144 -0.04235607 -
0.01348015
  0.09369236  0.02946641  0.0319985  -0.05690531  0.01878851 -
0.03534034
 -0.01182422 -0.01577108 -0.06313758 -0.00524855 -0.07180045
0.06245581
 -0.00546193 -0.01636546  0.01432091  0.08605643  0.0658258
0.02146672
  0.11522219 -0.00460271  0.00081966 -0.01413905  0.08193518
0.03804189
 -0.01032828 -0.04182971 -0.00175747 -0.02085764]

# Task 7: TF-IDF with scikit-learn
tfidf_vectorizer = TfidfVectorizer()
tfidf_matrix = tfidf_vectorizer.fit_transform(df['Text'])

# Calculate cosine similarity between two news articles
article1_index = 0
article2_index = 1
cosine_sim = cosine_similarity(tfidf_matrix[article1_index],
tfidf_matrix[article2_index])

print("\nTF-IDF with scikit-learn - Cosine Similarity between Article
1 and Article 2:", cosine_sim[0][0])


TF-IDF with scikit-learn - Cosine Similarity between Article 1 and
Article 2: 0.07875931547482325
```