```
import numpy as np
import pandas as pd
import re
import string

import nltk
from nltk.corpus import stopwords, wordnet
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

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

```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
True
```

```
documents = [
    # Sports
    "The football team won the championship",
    "Cricket is the most popular sport in India",
    "The player scored a hat trick",
    "Olympic athletes train very hard",
    "The match was postponed due to rain",

    # Politics
    "The government passed a new law",
    "Elections will be held next year",
    "The president addressed the nation",
    "Political parties campaigned aggressively",
    "The parliament debated the bill",

    # Health
    "Doctors recommend regular exercise",
    "A balanced diet improves health",
    "The patient recovered after treatment",
    "Hospitals provide medical care",
```

```
    "Mental health awareness is important",

    # Technology
    "Artificial intelligence is transforming industries",
    "Machine learning improves predictions",
    "Cybersecurity is crucial in the digital age",
    "Smartphones use advanced processors",
    "Technology evolves rapidly"
]

df = pd.DataFrame({"Text": documents})
df
```

	Text
0	The football team won the championship
1	Cricket is the most popular sport in India
2	The player scored a hat trick
3	Olympic athletes train very hard
4	The match was postponed due to rain
5	The government passed a new law
6	Elections will be held next year
7	The president addressed the nation
8	Political parties campaigned aggressively
9	The parliament debated the bill
10	Doctors recommend regular exercise
11	A balanced diet improves health
12	The patient recovered after treatment
13	Hospitals provide medical care
14	Mental health awareness is important
15	Artificial intelligence is transforming indust...
16	Machine learning improves predictions
17	Cybersecurity is crucial in the digital age
18	Smartphones use advanced processors
19	Technology evolves rapidly

Next steps: Generate code with df | New interactive sheet

```
nltk.download('punkt_tab')
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()

def preprocess(text):
    text = text.lower()
    text = re.sub(r'\d+', '', text)
    text = text.translate(str.maketrans('', '', string.punctuation))
    tokens = word_tokenize(text)
    tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in stop
    return " ".join(tokens)

df["Clean_Text"] = df["Text"].apply(preprocess)
df
```

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

	Text	Clean_Text
0	The football team won the championship	football team championship
1	Cricket is the most popular sport in India	cricket popular sport india
2	The player scored a hat trick	player scored hat trick
3	Olympic athletes train very hard	olympic athlete train hard
4	The match was postponed due to rain	match postponed due rain
5	The government passed a new law	government passed new law
6	Elections will be held next year	election held next year
7	The president addressed the nation	president addressed nation
8	Political parties campaigned aggressively	political party campaigned aggressively
9	The parliament debated the bill	parliament debated bill
10	Doctors recommend regular exercise	doctor recommend regular exercise
11	A balanced diet improves health	balanced diet improves health
12	The patient recovered after treatment	patient recovered treatment
13	Hospitals provide medical care	hospital provide medical care
14	Mental health awareness is important	mental health awareness important
15	Artificial intelligence is transforming indust...	artificial intelligence transforming industry
16	Machine learning improves predictions	machine learning improves prediction
17	Cybersecurity is crucial in the digital age	cybersecurity crucial digital age
18	Smartphones use advanced processors	smartphones use advanced processor
19	Technology evolves rapidly	technology evolves rapidly

Next steps: Generate code with df | New interactive sheet

```
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()

def preprocess(text):
    text = text.lower()
    text = re.sub(r'\d+', '', text)
    text = text.translate(str.maketrans('', '', string.punctuation))
    tokens = word_tokenize(text)
    tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in stop
    return " ".join(tokens)
```

```
df["Clean_Text"] = df["Text"].apply(preprocess)
df
```

	Text	Clean_Text
0	The football team won the championship	football team championship
1	Cricket is the most popular sport in India	cricket popular sport india
2	The player scored a hat trick	player scored hat trick
3	Olympic athletes train very hard	olympic athlete train hard
4	The match was postponed due to rain	match postponed due rain
5	The government passed a new law	government passed new law
6	Elections will be held next year	election held next year
7	The president addressed the nation	president addressed nation
8	Political parties campaigned aggressively	political party campaigned aggressively
9	The parliament debated the bill	parliament debated bill
10	Doctors recommend regular exercise	doctor recommend regular exercise
11	A balanced diet improves health	balanced diet improves health
12	The patient recovered after treatment	patient recovered treatment
13	Hospitals provide medical care	hospital provide medical care
14	Mental health awareness is important	mental health awareness important
15	Artificial intelligence is transforming indust...	artificial intelligence transforming industry
16	Machine learning improves predictions	machine learning improves prediction
17	Cybersecurity is crucial in the digital age	cybersecurity crucial digital age
18	Smartphones use advanced processors	smartphones use advanced processor
19	Technology evolves rapidly	technology evolves rapidly

Next steps: Generate code with df | New interactive sheet

```
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(df["Clean_Text"])

cosine_sim = cosine_similarity(tfidf_matrix)
cosine_df = pd.DataFrame(cosine_sim)
```

```
cosine_df.iloc[:5, :5]
```

	0	1	2	3	4
0	1.0	0.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0
2	0.0	0.0	1.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0
4	0.0	0.0	0.0	0.0	1.0

```
def jaccard_similarity(text1, text2):
    set1 = set(text1.split())
    set2 = set(text2.split())
    return len(set1 & set2) / len(set1 | set2)

for i in range(5):
    print("Sentence 1:", df["Text"][0])
    print("Sentence 2:", df["Text"][i])
    print("Jaccard Similarity:", jaccard_similarity(df["Clean_Text"][0], df["Cl
    print()
```

```
Sentence 1: The football team won the championship
Sentence 2: The football team won the championship
Jaccard Similarity: 1.0

Sentence 1: The football team won the championship
Sentence 2: Cricket is the most popular sport in India
Jaccard Similarity: 0.0

Sentence 1: The football team won the championship
Sentence 2: The player scored a hat trick
Jaccard Similarity: 0.0

Sentence 1: The football team won the championship
Sentence 2: Olympic athletes train very hard
Jaccard Similarity: 0.0

Sentence 1: The football team won the championship
Sentence 2: The match was postponed due to rain
Jaccard Similarity: 0.0
```

```
def wordnet_similarity(sent1, sent2):
    words1 = sent1.split()
    words2 = sent2.split()
    scores = []
```

```
        for w1 in words1:
            for w2 in words2:
                syn1 = wordnet.synsets(w1)
                syn2 = wordnet.synsets(w2)
                if syn1 and syn2:
                    sim = syn1[0].wup_similarity(syn2[0])
                    if sim:
                        scores.append(sim)

    return sum(scores)/len(scores) if scores else 0

for i in range(10):
    print("Sentence A:", df["Text"][i])
    print("Sentence B:", df["Text"][i+10])
    print("WordNet Similarity:", wordnet_similarity(df["Clean_Text"][i], df["Cle
    print()
```

```
Sentence A: The football team won the championship
Sentence B: Doctors recommend regular exercise
WordNet Similarity: 0.18203965796845054

Sentence A: Cricket is the most popular sport in India
Sentence B: A balanced diet improves health
WordNet Similarity: 0.17783359893460512

Sentence A: The player scored a hat trick
Sentence B: The patient recovered after treatment
WordNet Similarity: 0.2560432622932623

Sentence A: Olympic athletes train very hard
Sentence B: Hospitals provide medical care
WordNet Similarity: 0.21697624002674312

Sentence A: The match was postponed due to rain
Sentence B: Mental health awareness is important
WordNet Similarity: 0.1791543397048428

Sentence A: The government passed a new law
Sentence B: Artificial intelligence is transforming industries
```