```
pip install pandas nltk 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: 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) (2023.4)
Requirement already satisfied: numpy>=1.21.0 in /usr/local/lib/python3.10/dist-packages (from pandas) (1.25.2)
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) (2023.12.25)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.2)
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.11.4)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (3.3.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1->pandas) (1.16.0)
```

```
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

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

df = pd.read_csv('documents.csv')

stop_words = list(stopwords.words('english'))
tfidf_vectorizer = TfidfVectorizer(stop_words='english')
tfidf_matrix = tfidf_vectorizer.fit_transform(df['text'])

def search(query, df, tfidf_vectorizer, tfidf_matrix):

    query_tokens = word_tokenize(query.lower())
    query_vector = tfidf_vectorizer.transform([" ".join(query_tokens)])

    cosine_similarities = cosine_similarity(query_vector, tfidf_matrix).flatten()

    similar_documents = sorted(range(len(cosine_similarities)), key=lambda i: cosine_similarities[i], reverse=True)

    for idx in similar_documents[:3]:
        print(f"Document {idx+1}: {df.iloc[idx]['text']}")
query = "information retrieval"
search(query, df, tfidf_vectorizer, tfidf_matrix)
```

```
Document 62: Amount moment notice day offer. Agree enter interesting information cover read. Commercial operation people sound power
Document 4: Guess listen relationship nor. Join south television really several generation. Really professional factor information.
Document 57: After much couple. Sport nothing table well especially line.
Read how statement. Rise write senior hold area information within. Worker decide factor he east material or.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
query = input("Enter your search query: ")
search(query, df, tfidf_vectorizer, tfidf_matrix)
```

```
Enter your search query: couple
Document 59: Board site group our.
Pay deal father yes. Eye just prove suffer peace from. Suffer education work couple.
Document 38: Test would couple. Two smile join notice. Loss drive rule indeed central.
Which store hair. Popular peace agency budget Mrs social need.
Document 57: After much couple. Sport nothing table well especially line.
Read how statement. Rise write senior hold area information within. Worker decide factor he east material or.
```