Lab Assignment 3

Name: **J Viswaksena**

Roll.No: **AM.EN.U4AIE21035**

```
import requests
from bs4 import BeautifulSoup
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import matplotlib.pyplot as plt
import seaborn as sns

url = "https://en.wikipedia.org/wiki/COVID-19_pandemic"
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')
```

```
paragraphs = [p.get_text() for p in soup.find_all('p')]
text_data = ' '.join(paragraphs)
```

```
text_data
```

```
'\n The COVID-19 pandemic, also known as the coronavirus pandemic, is a global pandemic of coronavirus disease 2019 (CO
VID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The novel virus was first identified in
an outbreak in Wuhan, the capital of Hubei, China, in December 2019, before it spread to other areas of Asia, and then
worldwide in early 2020. The World Health Organization (WHO) declared the outbreak a public health emergency of interna
tional concern (PHEIC) on 30 January 2020, and assessed the outbreak had become a pandemic on 11 March 2020.[3] The WHO
ended the PHEIC on 5 May 2023.[4] As of 29 March 2024, the pandemic has caused 7,037,007[5] confirmed deaths, making it
the fifth deadliest pandemic or epidemic in history.\n COVID-19 symptoms range from asymptomatic to deadly, but most co
mmonly include fever, sore throat, nocturnal cough, and fatigue. Transmission of the virus is often through airborne pa
```

```
text_data_cleaned = ' '.join(word for word in text_data.split() if word.isalnum() or ' ' in word)
```

```
text_data_cleaned
```

```
'The also known as the coronavirus is a global pandemic of coronavirus diseas
e 2019 caused by severe acute respiratory syndrome coronavirus 2 The novel vi
rus was first identified in an outbreak in the capital of in December before
it spread to other areas of and then worldwide in early The World Health Orga
nization declared the outbreak a public health emergency of international con
cern on 30 January and assessed the outbreak had become a pandemic on 11 Marc
h The WHO ended the PHEIC on 5 May As of 29 March the pandemic has caused con
firmed making it the fifth deadliest pandemic or epidemic in symptoms range f
```

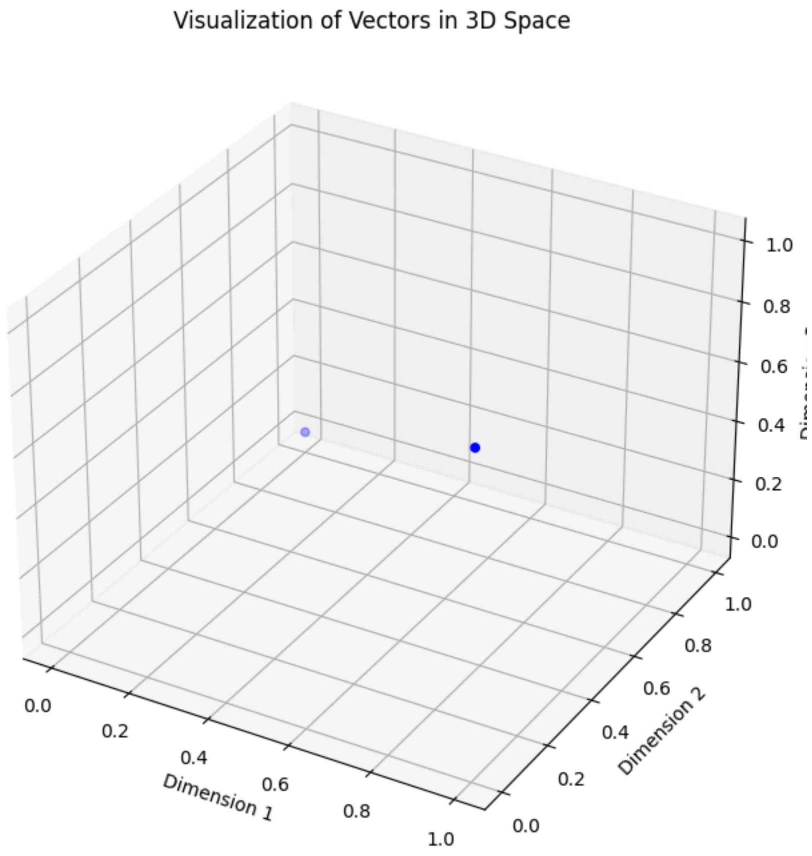
```
paragraph1 = text_data_cleaned[:300]
paragraph2 = text_data_cleaned[300:600]
```

```
vectorizer = CountVectorizer().fit_transform([paragraph1, paragraph2])
cosine_sim_count = cosine_similarity(vectorizer)

tfidf_vectorizer = TfidfVectorizer().fit_transform([paragraph1, paragraph2])
cosine_sim_tfidf = cosine_similarity(tfidf_vectorizer)
```

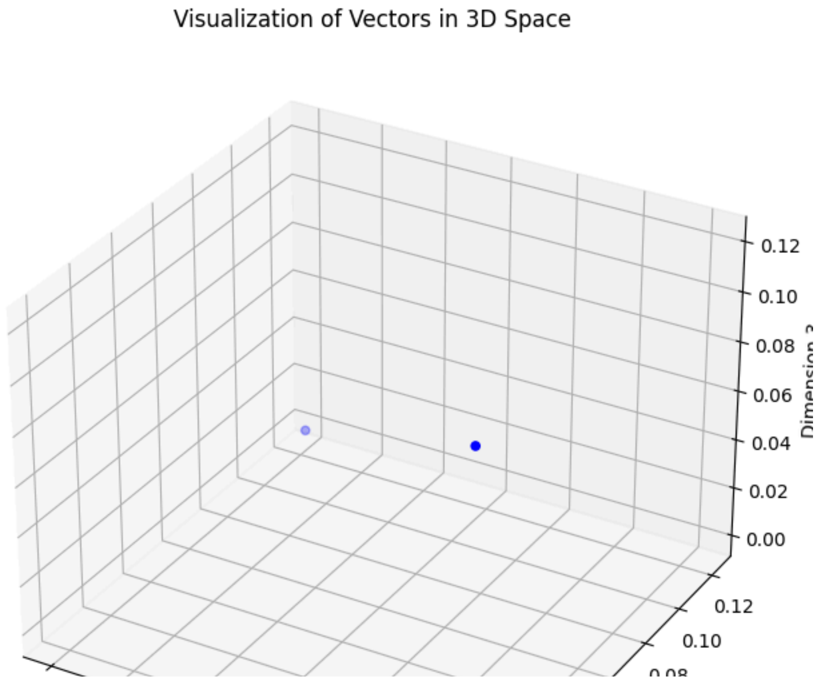
```
print(vectorizer.toarray())
```

```
[[0 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 0 0 3 0 1 0 1 1 0 0 0 0 1 1 0 0 1 0 1 4
  0 1 1 0 1 0 0 0 1 3 0 0 0 1 1 1 0 0 1 1 1 1 0 5 1 1 1 1 0 1 1]
 [1 0 1 1 0 0 0 1 0 1 1 1 0 0 0 1 1 1 0 1 0 1 0 0 1 1 1 1 0 0 1 1 0 1 0 1
  1 0 1 1 0 1 2 1 0 2 3 1 1 0 2 3 1 1 0 0 0 0 1 6 0 0 0 0 1 0 0]]
```

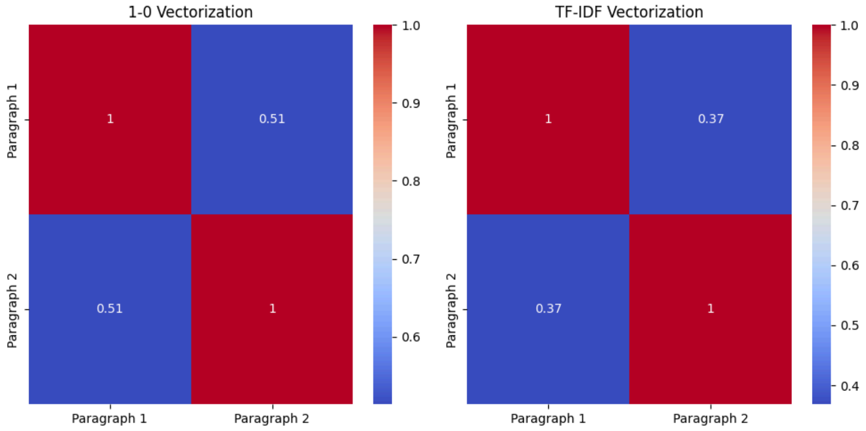
```
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

vectors = np.array(vectorizer.toarray())

fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')

ax.scatter(vectors[:, 0], vectors[:, 1], vectors[:, 2], color='b', marker='o')

ax.set_xlabel('Dimension 1')
ax.set_ylabel('Dimension 2')
ax.set_zlabel('Dimension 3')
ax.set_title('Visualization of Vectors in 3D Space')

plt.show()
```



```
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

vectors = np.array(tfidf_vectorizer.toarray())

fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')

ax.scatter(vectors[:, 0], vectors[:, 1], vectors[:, 2], color='b', marker='o')

ax.set_xlabel('Dimension 1')
ax.set_ylabel('Dimension 2')
ax.set_zlabel('Dimension 3')
ax.set_title('Visualization of Vectors in 3D Space')

plt.show()
```



```
plt.figure(figsize=(10, 5))

plt.subplot(1, 2, 1)
sns.heatmap(cosine_sim_count, annot=True, cmap='coolwarm', xticklabels=['Paragraph 1', 'Paragraph 2'], yticklabels=['Paragrap
plt.title('1-0 Vectorization')

plt.subplot(1, 2, 2)
sns.heatmap(cosine_sim_tfidf, annot=True, cmap='coolwarm', xticklabels=['Paragraph 1', 'Paragraph 2'], yticklabels=['Paragrap
plt.title('TF-IDF Vectorization')

plt.tight_layout()
plt.show()
```


1-0 Vectorization
Paragraph 1
Paragraph 2
Paragraph 1
Paragraph 2
TF-IDF Vectorization
Paragraph 1
Paragraph 2
Paragraph 1
Paragraph 2