NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

Data Mining (LAB)

Lab Report - 05

Submitted to
Dr. Moiz Ullah Ghouri

Submitted By
Junaid Asif
(BSAI-144)

Submission Date: November 12, 2024

- **Calculate cosine similarity of four paragraphs and also print out the count for each word in the paragraphs. For Example: is appeared how many times in all three paragraphs?**

```python
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd

def cosine_similarity(x, y):
    if len(x) != len(y):
        return None

    dot_product = np.dot(x, y)

    IIxII = np.sqrt(np.sum(x**2))
    IIyII = np.sqrt(np.sum(y**2))

    cos_similarity_formula = dot_product / (IIxII * IIyII)
    return cos_similarity_formula

twitter = """
Twitter is an online social media and social networking service owned
and operated by American company X Corp.,
the legal successor of Twitter, Inc. Twitter users outside the United
States are legally served by the Ireland-based
Twitter International Unlimited Company, which makes these users
subject to Irish and European Union data protection laws.
On Twitter users post texts, photos and videos known as 'tweets'.
Registered users can tweet, like, 'retweet' tweets,
and direct message (DM) other registered users, while unregistered
users only have the ability to view public tweets.
Users interact with Twitter through browser or mobile frontend
software, or programmatically via its APIs.
"""
```

```python
facebook = """
Facebook is an online social media and social networking service owned
by American technology giant Meta Platforms.
Created in 2004 by Mark Zuckerberg with fellow Harvard College
students and roommates Eduardo Saverin, Andrew McCollum,
Dustin Moskovitz, and Chris Hughes, its name derives from the face
book directories often given to American university students.
Membership was initially limited to only Harvard students, gradually
expanding to other North American universities and,
since 2006, anyone over 13 years old. As of December 2022, Facebook
claimed 2.96 billion monthly active users, and ranked third
worldwide among the most visited websites. It was the most downloaded
mobile app of the 2010s. Facebook can be accessed from devices
with Internet connectivity, such as personal computers, tablets and
smartphones. After registering, users can create a profile
revealing information about themselves. They can post text, photos and
multimedia which are shared with any other users who have
agreed to be their friend' or, with different privacy settings,
publicly. Users can also communicate directly with each other with
Messenger, join common-interest groups, and receive notifications on
the activities of their Facebook friends and the pages they follow.
"""
tiktok = """
TikTok, and its Chinese counterpart Douyin (Chinese: 抖音; pinyin:
Dǒuyīn), is a short-form video hosting service owned by ByteDance.
It hosts user-submitted videos, which can range in duration from 3
seconds to 10 minutes. Since their launches, TikTok and Douyin have
gained global popularity.[6][7] In October 2020, TikTok surpassed 2
billion mobile downloads worldwide. Morning Consult named TikTok the
third-fastest growing brand of 2020, after Zoom and Peacock.
Cloudflare ranked TikTok the most popular website of 2021,
surpassing google.com.
"""
```

```
instagram = """
Instagram is a photo and video sharing social networking service owned
by American company Meta Platforms. The app allows users to
upload media that can be edited with filters and organized by hashtags
and geographical tagging. Posts can be shared publicly or
with preapproved followers. Users can browse other users' content by
tag and location, view trending content, like photos, and follow
other users to add their content to a personal feed. Instagram was
originally distinguished by allowing content to be framed only in a
square (1:1) aspect ratio of 640 pixels to match the display width of
the iPhone at the time. In 2015, this restriction was eased with
an increase to 1080 pixels. It also added messaging features, the
ability to include multiple images or videos in a single post, and a
Stories feature—similar to its main competitor Snapchat—which allowed
users to post their content to a sequential feed, with each post
accessible to others for 24 hours. As of January 2019, Stories is used
by 500 million people daily.
"""
```

```
document = [twitter, facebook, tiktok, instagram]
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(document).toarray()

cos_sim_1_2 = cosine_similarity(X[0, :], X[1, :])
cos_sim_1_3 = cosine_similarity(X[0, :], X[2, :])
cos_sim_1_4 = cosine_similarity(X[0, :], X[3, :])
cos_sim_2_3 = cosine_similarity(X[1, :], X[2, :])
cos_sim_2_4 = cosine_similarity(X[1, :], X[3, :])
cos_sim_3_4 = cosine_similarity(X[2, :], X[3, :])

print('\nCosine Similarity between:')
print(f'\tDocument 1 (Twitter) and Document 2 (Facebook): {cos_sim_1_2}')
print(f'\tDocument 1 (Twitter) and Document 3 (TikTok): {cos_sim_1_3}')
print(f'\tDocument 1 (Twitter) and Document 4 (Instagram): {cos_sim_1_4}')
print(f'\tDocument 2 (Facebook) and Document 3 (TikTok): {cos_sim_2_3}')
print(f'\tDocument 2 (Facebook) and Document 4 (Instagram): {cos_sim_2_4}')
print(f'\tDocument 3 (TikTok) and Document 4 (Instagram): {cos_sim_3_4}')


word_counts = pd.DataFrame(X, columns=vectorizer.get_feature_names_out())
print(word_counts)
```

Output:

```
Cosine Similarity between:
        Document 1 (Twitter) and Document 2 (Facebook): 0.49812777753930826
        Document 1 (Twitter) and Document 3 (TikTok): 0.2326957082444114
        Document 1 (Twitter) and Document 4 (Instagram): 0.4939598741083312
        Document 2 (Facebook) and Document 3 (TikTok): 0.34922161264379814
        Document 2 (Facebook) and Document 4 (Instagram): 0.6035279596937492
        Document 3 (TikTok) and Document 4 (Instagram): 0.30479179723505573
   10  1080  13  2004  2006  2010s  2015  2019  2020  2021  ...  which  while  \
0   0     0   0     0     0      0     0     0     0     0  ...      1      1
1   0     0   1     1     1      1     0     0     0     0  ...      1      0
2   1     0   0     0     0      0     0     0     2     1  ...      1      0
3   0     1   0     0     0      0     1     1     0     0  ...      1      0

   who  width  with  worldwide  years  zoom  zuckerberg  抖音
0    0      0     1          0      0     0           0   0
1    1      0     6          1      1     0           1   0
2    0      0     0          1      0     1           0   1
3    0      1     4          0      0     0           0   0

[4 rows x 302 columns]
```