

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**Український державний університет науки і технологій**

Кафедра «Комп’ютерні інформаційні технології»

**Лабораторна робота №1**

**з дисципліни «Інтернет-технології»**

**на тему:**

**«**Порівняння веб-сайтів за допомогою косинусної подібності.»

Виконав:

студент гр. ПЗ2421

Кулик С. В.

Прийняв:

Андрющенко В. О.

Дніпро, 2025

**Тема:** Порівняння веб-сайтів за допомогою косинусної подібності.

**Постановка задачі:** Розробити програмний засіб для визначення подібності веб-сайтів за змістом за допомогою їх косинусної подібності.

**Хід Роботи**

import math  
from collections import Counter  
import requests  
from bs4 import BeautifulSoup  
  
# Функція для отримання тексту з веб-сайту  
def fetch\_website\_content(url):  
 try:  
 response = requests.get(url)  
 soup = BeautifulSoup(response.text, 'html.parser')  
 # Видалення скриптів і стилів  
 for script\_or\_style in soup(['script', 'style']):  
 script\_or\_style.decompose()  
 return ' '.join(soup.stripped\_strings)  
 except Exception as e:  
 print(f"Помилка при отриманні даних з {url}: {e}")  
 return ""  
  
# Функція для обчислення TF  
def compute\_tf(text):  
 words = text.split()  
 total\_words = len(words)  
 word\_counts = Counter(words)  
 return {word: count / total\_words for word, count in word\_counts.items()}  
  
# Функція для обчислення IDF  
def compute\_idf(documents):  
 total\_documents = len(documents)  
 word\_document\_counts = Counter(  
 word for doc in documents for word in set(doc.split())  
 )  
 return {word: math.log(total\_documents / (1 + count)) for word, count in word\_document\_counts.items()}  
  
# Функція для обчислення TF-IDF  
def compute\_tfidf(tf, idf):  
 return {word: tf.get(word, 0) \* idf.get(word, 0) for word in idf.keys()}  
  
# Функція для обчислення косинусної подібності  
def cosine\_similarity(vector1, vector2):  
 dot\_product = sum(vector1.get(word, 0) \* vector2.get(word, 0) for word in set(vector1.keys()).union(vector2.keys()))  
 magnitude1 = math.sqrt(sum(value \*\* 2 for value in vector1.values()))  
 magnitude2 = math.sqrt(sum(value \*\* 2 for value in vector2.values()))  
 if magnitude1 == 0 or magnitude2 == 0:  
 return 0.0  
 return dot\_product / (magnitude1 \* magnitude2)  
  
# Основний процес  
urls = [  
 "https://www.geeksforgeeks.org",  
 "https://realpython.com",  
 "https://www.w3schools.com",  
 "https://www.freecodecamp.org",  
 "https://developer.mozilla.org",  
]  
  
# Завантаження контенту з веб-сайтів  
documents = [fetch\_website\_content(url) for url in urls]  
  
# Розрахунок TF, IDF та TF-IDF для кожного документу  
tfs = [compute\_tf(doc) for doc in documents]  
idf = compute\_idf(documents)  
tfidfs = [compute\_tfidf(tf, idf) for tf in tfs]  
  
# Розрахунок косинусної подібності між веб-сайтами  
for i, url in enumerate(urls):  
 print(f"Подібність для {url}:")  
 for j, other\_url in enumerate(urls):  
 if i != j:  
 similarity = cosine\_similarity(tfidfs[i], tfidfs[j])  
 print(f" З {other\_url}: {similarity:.4f}")  
 print("\n")  
  
  
  
# urls = [  
# "https://www.geeksforgeeks.org",  
# "https://realpython.com",  
# "https://www.w3schools.com",  
# "https://www.freecodecamp.org",  
# "https://developer.mozilla.org",  
# ]

**Результати**

Подібність для https://www.geeksforgeeks.org:  
 З https://realpython.com: 0.0289  
 З https://www.w3schools.com: 0.0816  
 З https://www.freecodecamp.org: 0.0000  
 З https://developer.mozilla.org: 0.0021  
  
  
Подібність для https://realpython.com:  
 З https://www.geeksforgeeks.org: 0.0289  
 З https://www.w3schools.com: 0.0372  
 З https://www.freecodecamp.org: 0.0000  
 З https://developer.mozilla.org: 0.0073  
  
  
Подібність для https://www.w3schools.com:  
 З https://www.geeksforgeeks.org: 0.0816  
 З https://realpython.com: 0.0372  
 З https://www.freecodecamp.org: 0.0000  
 З https://developer.mozilla.org: 0.0184  
  
  
Подібність для https://www.freecodecamp.org:  
 З https://www.geeksforgeeks.org: 0.0000  
 З https://realpython.com: 0.0000  
 З https://www.w3schools.com: 0.0000  
 З https://developer.mozilla.org: 0.0000  
  
  
Подібність для https://developer.mozilla.org:  
 З https://www.geeksforgeeks.org: 0.0021  
 З https://realpython.com: 0.0073  
 З https://www.w3schools.com: 0.0184  
 З https://www.freecodecamp.org: 0.0000

**Висновок:** Розроблена програма дозволяє порівнювати веб-сайти за змістом, використовуючи метод косинусної подібності. Вона виконує збір тексту з сайтів і розрахунок TF-IDF та порівнює їх.