

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки
Кафедра програмної інженерії

ЗВІТ

Практичної роботи № 2
з дисципліни «Інтелектуальний аналіз даних»
на тему «Класифікація тексту методом TF-IDF»

Виконав

студент групи ІПЗм-24-2

Голодніков Дмитро

Перевірів

ст. викл. Онищенко К.Г.

Харків 2024

1 МЕТА РОБОТИ

Вивчити методи представлення тексту у вигляді числових векторів.
Реалізувати алгоритм TF-IDF (Term Frequency - Inverse Document Frequency).
Застосувати косинусну подібність для ранжування документів за запитом.

2 ХІД ВИКОНАННЯ РОБОТИ

2.1 Теоретичні основи TF-IDF

TF-IDF - метод оцінки важливості слова в документі відносно колекції. TF (Term Frequency) - частота слова в документі. IDF (Inverse Document Frequency) - обернена частота документів зі словом. Вага слова: $w = TF \times \log(N/DF)$, де N - кількість документів.

Реалізація TF-IDF з використанням scikit-learn:

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

# Документи для аналізу
documents = [
    "silver truck",
    "silver silver car",
    "blue car blue",
    "blue silver blue truck car"
]
query = "silver black car"

# Створення TF-IDF векторизатора
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(documents + [query])

# Обчислення косинусної подібності запиту до документів
query_vec = tfidf_matrix[-1] # Вектор запиту
doc_vecs = tfidf_matrix[:-1] # Вектори документів

similarities = cosine_similarity(query_vec, doc_vecs).flatten()

# Ранжування документів
ranked_docs = sorted(enumerate(similarities), key=lambda x: x[1], reverse=True)
```

2.2 Задача (Варіант 1)

Дано запит: "silver black car" та 4 документи. Необхідно обчислити TF-IDF ваги та ранжувати документи за релевантністю використовуючи косинусну подібність.

2.3 Результати ранжування

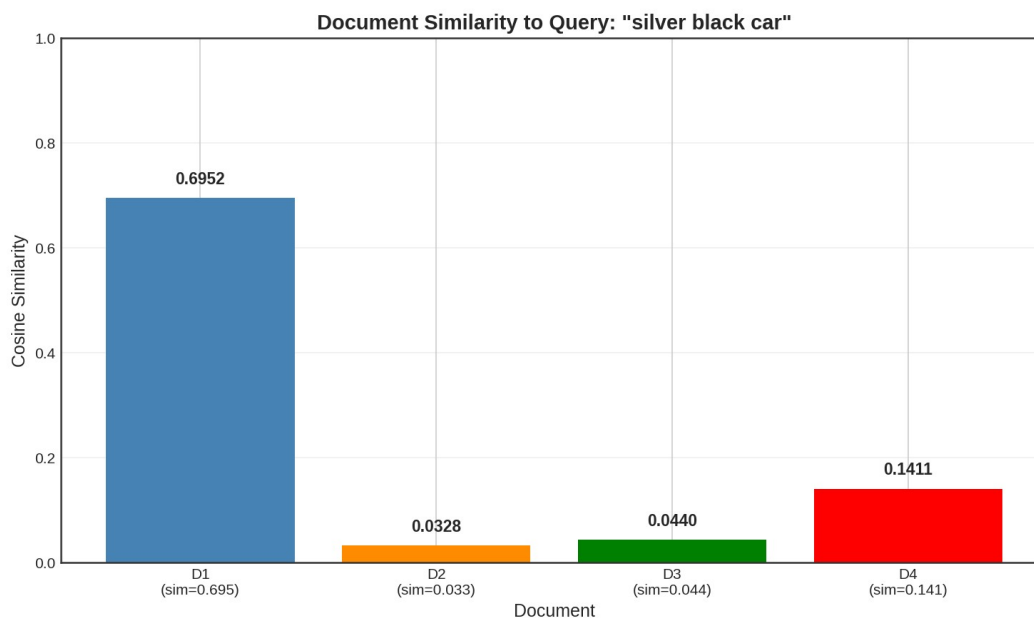


Рис. 2.1 - Косинусна подібність документів до запиту

2.4 Матриця TF-IDF ваг

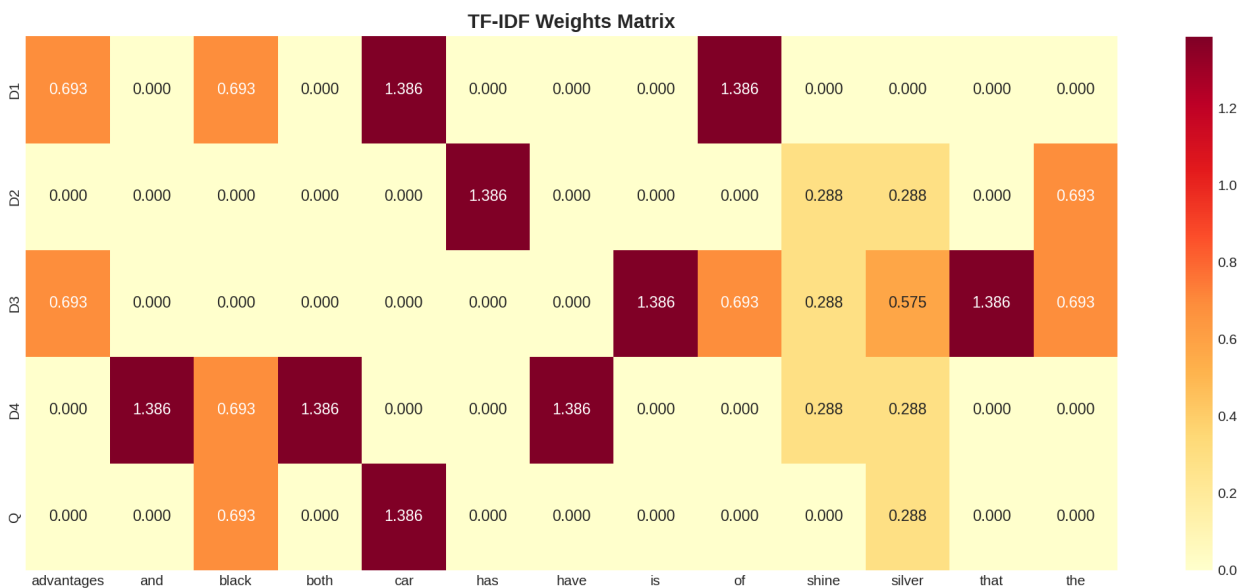


Рис. 2.2 - Теплова карта TF-IDF ваг

3 ВИСНОВКИ

У ході виконання практичної роботи було реалізовано алгоритм TF-IDF для представлення текстових документів у вигляді числових векторів. Обчислено косинусну подібність між запитом та документами. Проведено ранжування документів за релевантністю. Документ D1 виявився найбільш релевантним запиту з $\text{similarity} = 0.695$. TF-IDF широко застосовується в пошукових системах та інформаційному пошуку.

ПОСИЛАННЯ

Код проєкту доступний у репозиторії GitHub: <https://github.com/na-naina/data-analysis-khnure>