Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет прикладної математики
Кафедра програмного забезпечення комп’ютерних систем

Методи організації пошуку інформації

Лабораторна робота № 1

«Реалізувати теоретико-множинну (стандартну булеву) та алгебраїчну (векторно-просторову) модель подання документів, таким чином ознайомившись з найбільш поширеними моделями та набувши практичні навички з їх реалізації.»

Виконав:
студент
групи КП-93
Долгов Олексій
Варіант 7

Київ 2023

Мета

Реалізувати теоретико-множинну (стандартну булеву) та алгебраїчну (векторно-просторову) модель подання документів, таким чином ознайомившись з найбільш поширеними моделями та набувши практичні навички з їх реалізації.

Постановка завдання

Програмне забезпечення повинно мати два режими роботи: перший – з використанням теоретико-множинної моделі, другий – з використанням алгебраїчної моделі. Це може бути реалізовано будь-яким способом на вибір студента, наприклад параметром запуску для консольного застосунку, перемикачем у графічному інтерфейсі тощо.

Код Розробленого програмного забезпечення

https://github.com/dolho/search-labs/tree/master/lab1

Результати роботи програми

1) Конструювання індексних термів та індексація текстового корпусу

```
$ python main.py init-index-terms initial_index_terms.json
$ python main.py index-text-corpus
```

```
{
    "index_terms": {
        "index_terms": {
            "quick": [
                "foxes.txt"
            ],
            "brown": [
                "brown eagle.txt"
            ],
            "fox": [
                "foxes.txt"
            ],
            "eagle": [
                "eagle.txt",
                "brown eagle.txt"
            ],
            "wolf": [
                "wolfs.txt"
            ],
            "prey": [
                "eagle.txt",
                "wolfs.txt",
                "foxes.txt"
            ],
            "is": [
                "eagle.txt",
```


Вміст файлу index_terms.json, після ініціалізації термів та індексації текстового корпусу

```
(lab1-py3.10) axel@axel-ThinkPad-E14:~/Desktop/uni/4_курс/sem2/search-labs/lab1$ python main.py search-boolean '"eagle" | "prey" & "fox" | "wolf"'
Found documents:
foxes.txt
wolfs.txt
```

Пошук документів за запитом у нормальній кон'юктивній формі.

```
(lab1-py3.10) axel@axel-ThinkPad-E14:~/Desktop/uni/4_курс/sem2/search-labs/lab1$ python main.py search-boolean '"eagle" | "prey" & "fox" & "wolf"'
Documents not found :(
```

Виконання запиту, який не знаходить ніяких документів

```
1$ python main.py replace-vector-index-terms
1$ python main.py index-text-corpus-vector
```

Встановлення термів для векторного пошуку та індексація текстового корпусу


```
    "indexed_documents": {
        "eagle.txt": {
            "eagle": {
                "occurrences": 2,
                "tf": 0.4,
                "idf": 0.6931471805599453,
                "tf_idf": 0.2772588722239781
            },
            "is": {
                "occurrences": 2,
                "tf": 0.4,
                "idf": 0.28768207245178085,
                "tf_idf": 0.11507282898071235
            },
            "prey": {
                "occurrences": 1,
                "tf": 0.2,
                "idf": 0.28768207245178085,
                "tf_idf": 0.05753641449035617
            }
        },
        "brown eagle.txt": {
            "brown": {
                "occurrences": 1,
                "tf": 0.5,
                "idf": 1.3862943611198906,
                "tf_idf": 0.6931471805599453
            },
```


Приклад проіндексованих документів

```
(lab1-py3.10) axel@axel-ThinkPad-E14:~/Desktop/uni/4_курс/sem2/search-labs/lab1$ python main.py search-vector '"brown" "eagle" "fox"'
Found documents:
brown eagle.txt; Score: 0.7745966692414834
eagle.txt; Score: 0.5773502691896257
foxes.txt; Score: 0.5773502691896257
```

Виконня пошукового запиту. Можна побачити, що документ `brown eagle.txt` отримав вищу оцінку, ніж eagle.txt та foxes.txt, оскільки він є більш релевантним до запиту.

```
(lab1-py3.10) axel@axel-ThinkPad-E14:~/Desktop/uni/4_курс/sem2/search-labs/lab1$ python main.py search-vector '"fox" "nonexistent term"'
Found documents:
foxes.txt; Score: 1.0
```

Терм fox зустрічається тільки у документі foxes.txt, тому він отримує найбільш возможну оцінку. Не існуючі терми ігноруються.

```
(lab1-py3.10) axel@axel-ThinkPad-E14:~/Desktop/uni/4_курс/sem2/search-labs/lab1$ python main.py search-vector '"fox" "prey"'
Found documents:
foxes.txt; Score: 0.8360329614711213
eagle.txt; Score: 0.7071067811865475
wolfs.txt; Score: 0.7071067811865475
(lab1-py3.10) axel@axel-ThinkPad-E14:~/Desktop/uni/4_курс/sem2/search-labs/lab1$ python main.py search-vector '"quick" "fox" "prey"'
Found documents:
foxes.txt; Score: 0.8916671860126297
eagle.txt; Score: 0.5773502691896258
wolfs.txt; Score: 0.5773502691896258
```

Покращення релевантності видачі із додаванням до пошукового запиту терму quick.

Висновки

В ході даної лабораторної роботи було побудовано консольну пошукову систему, яка дозволяє шукати документи за стандартною булевою та векторною моделлю.