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 № 3

«Доповнити інформаційно-пошукову систему, реалізовану в рамках другої лабораторної роботи, можливістю повнотекстового пошуку.»

Виконав:
студент
групи КП-93
Долгов Олексій

Київ 2022

Мета

Доповнити інформаційно-пошукову систему, реалізовану в рамках другої лабораторної роботи, можливістю повнотекстового пошуку.

Код Розробленого програмного забезпечення

https://github.com/dolho/search-labs/tree/master/lab2

Постановка завдання

Предметна галузь - музика.

Результати роботи програми

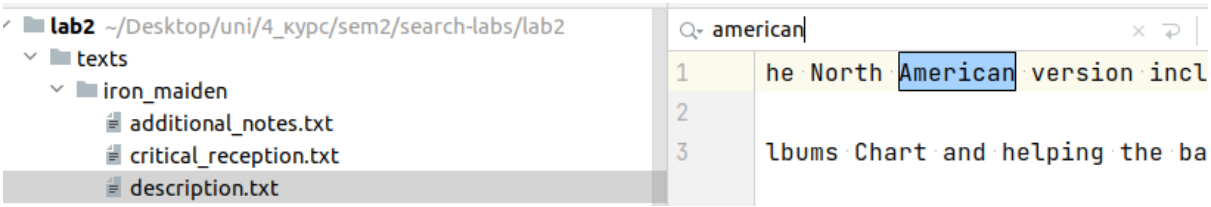
Хід роботи програми аналогічний до ходу роботи програми у другій лабораторній: конструювання запиту та виконная пошуку.
Приклад повнотекстового пошуку за запитом “american duo”

```
python main.py add-match must description 'american duo'
python main.py search
```

```
Album(album_name='Soulvaki', re
tion='', critical_reception='',
Album(album_name='Iron Maiden',
ion='', critical_reception='',
Neil Halstead', 'Rachel Goswell', 'Christian Savill', 'Nic
584-4ed3-80f8-d0ba7359a1f0', _score=2.3715408)
:["Paul Di'Anno", 'Dave Murray', 'Dennis Stratton', 'Steve
aa-4a88-8525-c67d5ccd3d51', _score=0.732833)
```

За заданим запитом було знайдено два документи - Soulvaki та Iron Maiden. Різниця між оцінками для них зумовлена різною релевантністю документів:

Текст для Soulvaki має два терми з запиту - American duo



Текст для Iron Maiden відповідає лише терму American.

Було побудовано кастомний аналайзер на основі стандартного аналайзеру для англійської мови. Для нього було додано кастомний pattern_replacement чар фільтр, який прибирає підстроки виду [12345…], що часто зустрічаються у необробленому тексту з Вікіпедії.

Кастомний аналайзер

```
{
   "analysis": {
       "filter": {
           "english_stop": {
               "type": "stop",
               "stopwords": "_english_"
           },
           "english_keywords": {
               "type": "keyword_marker",
               "keywords": ["example"]
           },
           "english_stemmer": {
               "type": "stemmer",
               "language": "english"
           },
           "english_possessive_stemmer": {
               "type": "stemmer",
               "language": "possessive_english"
           }
```

```
        },
        "analyzer": {
            "custom_wikipedia_analyzer": {
                "tokenizer": "standard",
                "char_filter": [
                    "wikipedia_symbols"
                ],
                "filter": [
                    "lowercase",
                    "english_possessive_stemmer",
                    "english_stop",
                    "english_stemmer"
                ]
            }
        },
        "char_filter": {
            "wikipedia_symbols": {
                "type": "pattern_replace",
                "pattern": "\[\d+\]",
                "replacement": " "
            },
        }
    }
}
```

Маппінг для текстових полів

```
"description": {
   "type": "text",
   "analyzer": "standard"
},
"critical_reception": {
   "type": "text",
   "analyzer": "custom_wikipedia_analyzer"
},
"additional_notes": {
   "type": "text",
   "analyzer": "english"
}
```

Висновки

В ході даної лабораторної роботи було добудовано дії для повнотекстового пошуку на основі системи з другої лабораторної роботи. Було розроблено кастомний аналайзер, для видалення підстрок, які вказують на номер посилання у вікіпедії.