Учреждение образования

«Белорусский государственный университет информатики и радиоэлектроники»

Кафедра интеллектуальных информационных технологий

ОТЧЁТ

по лабораторной работе №1 по дисциплине «ЕЯзИИС»

на тему: «Разработка информационно-поисковой системы и методы оценки качества ее работы»

Выполнили студенты группы 821701: Поживилко П.С.
Витушко Л. Д.

Проверил: Крапивин Ю.Б.

Минск, 2021

Цель работы: освоить на практике основные принципы реализации информационно-поисковых систем и методы оценки качества их работы.

Задание

Вариант 1.

Сфера применения: локальная вычислительная сеть.

Стратегия поиска: логическая.

Язык: русский.

Ход работы.

Структура разработанной системы

Система представляет собой консольное приложение. Для простоты системы, данные берутся из указанного каталога.

Структура базы данных системы

При реализации системы было принято решение хранить необходимую информацию в оперативной памяти приложения, а данные вычитывать на старте из файловой системы. В данной работе использование БД нет необходимости. Также это упрощает разработку системы

Основные алгоритмы реализации компонентов системы

Основной алгоритм заключается в разбиении логического выражения, принимаемого на вход, на подстроки, которое выполняется рекурсивно и заменяет во входящем поисковом выражении слова на булевы значения индивидуально для каждого файла.

Запрос формулируется как произвольное булевы выражения, связывающие термины с помощью стандартных логических операций: AND, OR или NOT.

Если в файле присутствует слово, то оно заменяется на 1, если отсутствует – 0. Мерой соответствия запроса документу служит значение статуса выборки (RSV, retrieval status value). В логической модели статус выборки равен либо 1, если для данного документа вычисление выражения запроса дает значение «истина», либо 0 в противном случае. Все документы с RSV = 1 считаются релевантными запросу.

```
def parens(token_lst):
    left_lst = find(token_lst, '(')

    if not left_lst:
        return False, -1, -1

    left = left_lst[-1]

    if token_lst[left + 1] != 0 and token_lst[left + 1] != 1:
        right = find(token_lst, ')', left + 3)[0]
    else:
        right = find(token_lst, ')', left + 4)[0]

    return True, left, right


def bool_eval(token_lst):
    if len(token_lst) == 2:
        return token_lst[0](token_lst[1])
    else:
        return token_lst[1](token_lst[0], token_lst[2])


def formatted_bool_eval(token_lst, empty_res=empty_res):
    if not token_lst:
        return empty_res

    if len(token_lst) == 1:
        return token_lst[0]

    has_parens, l_paren, r_paren = parens(token_lst)

    if not has_parens:
        return bool_eval(token_lst)

    token_lst[l_paren:r_paren + 1] = [bool_eval(token_lst[l_paren+1:r_paren])]

    return formatted_bool_eval(token_lst, bool_eval)
```

Рисунок 1. Основной алгоритм расчета значения логического выражения

Результаты тестирования системы

Результаты тестирование системы приведены на следующих изображениях.

В данном случае запрос составлен для нахождения файлов, в которых одновременно присутствуют слова «Рогожин» и «Мышкин» либо слово «Каренина».

```
--------------Меню---------------
1. Логический поиск
2. Информация о метриках
3. Справка
0. Выход
1
Введите логическую формулу для поиска:
(('Рогожин' AND 'Мышкин') OR 'Каренина')
Файл: D:\BSUIR\ЕЯИИС\LR1\avidreaders.ru__anna-karenina.txt
Список присутствующих слов: ['Каренина']
Файл: D:\BSUIR\ЕЯИИС\LR1\avidreaders.ru__idiot.txt
Список присутствующих слов: ['Рогожин', 'Мышкин']
```

Рисунок 2. Результат тестирования системы

Как видим, условию удовлетворяют два файла: «Идиот» Ф.М. Достоевского, содержащий фамилии «Рогожин» и «Мышкин», и «Анна Каренина» Л.Н. Толстого, в котором присутствует слово «Каренина».

```
--------------Меню---------------
1. Логический поиск
2. Информация о метриках
3. Справка
0. Выход
3
Логическая модель трактует термины в запросе как булевы переменные.
При наличии термина в документе соответствующая переменная принимает значение «true» (истина).
Присваивание терминам весовых коэффициентов не допускается.
 Запросы формулируются как произвольные булевы выражения, связывающие термины с помощью стандартных логических операций: AND, OR или NOT.
Мерой соответствия запроса документу служит значение статуса выборки (RSV, retrieval status value).
В булевой модели статус выборки равен либо 1, если для данного документа вычисление выражения запроса дает значение «истина», либо 0 в противном случае.
Все документы с RSV = 1 считаются релевантными запросу.
```

Рисунок 3. Результат вызова справки

Информация о тестовой коллекции документов

Тестовая коллекция документов представляет из классические произведения в текстовом формате:

1. «Анна Каренина» Л.Н. Толстого,
2. «Братья Карамазовы» Ф.М. Достоевского,
3. «Идиот» Ф.М. Достоевского,
4. «Преступление и наказание» Ф.М. Достоевского.

Результаты оценки по каждой из метрик

Результаты оценки по каждой из метрик приведены на следующем изображении.

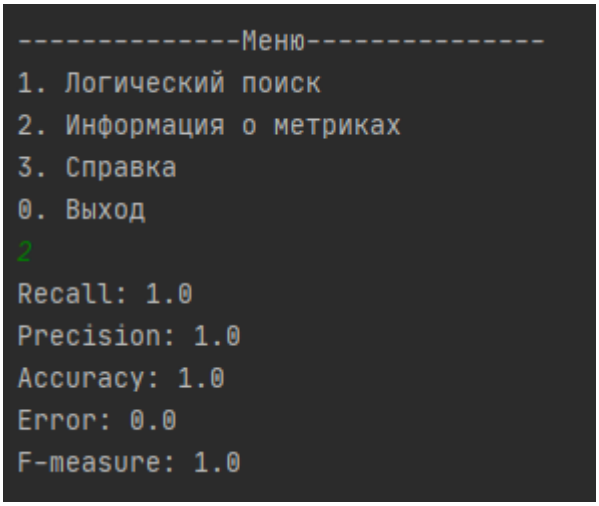

Рисунок 4. Результат вызова справки

Результаты анализа полученных данных

Полученные данные свидетельствует о том, что алгоритм является точным.

Описание готовых использованных компонент

При реализации системы были использованы следующие компоненты:

- Написан алгоритм на языке Python.

Вывод:

В данной лабораторной работе был реализована информационно-поисковая система с нуля помощью языка программирования Python.

Реализованная система была протестирована и проанализирована с помощью метрик.

Были приведены алгоритмы реализации системы.

В результате лабораторной работы были закреплены принципы реализации информационно-поисковых систем.