|  |  |  |  |
| --- | --- | --- | --- |
| Министерство образования Республики Беларусь  Учреждение образования  Белорусский Государственный Университет Информатики и Радиоэлектроники | | | |
| Факультет | Компьютерных сетей и систем | | |
| Кафедра | Информатики  Дисциплина: Конструирование те технологии электронных вычислительных средств | | |
|  |  | | |
| Интеллектуальные интернет-технологии  ЛАБОРАТОРНАЯ РАБОТА №2  «Изучение информационно-поисковых систем»  БГУИР 1-40 81 04 | | | |
| Магистрант:  гр. 858641  Кукареко А.В. | |  | Проверил:  Захаров В. В. |
| Минск, 2020 | | | |

# 

ХОД РАБОТЫ

**Задание.**

Разработать и реализовать документальную гипертекстовую информационно-поисковую систему по документам реализованной в предыдущей лабораторной модели гипертекста.

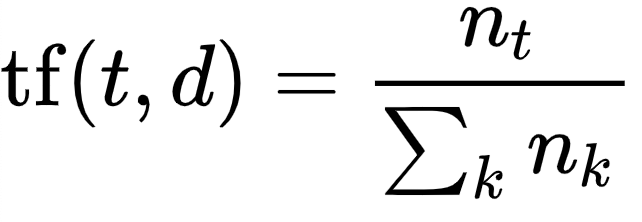
**Результат выполнения:**

Для реализации функции поиска был выбран алгоритм «TF-IDF». Данный алгоритм предназначен для расчета важности слова для какого-либо документа относительно других документов.

Если термин часто используется в определенном тексте, но редко в других, то он имеет большую значимость для данного текста.

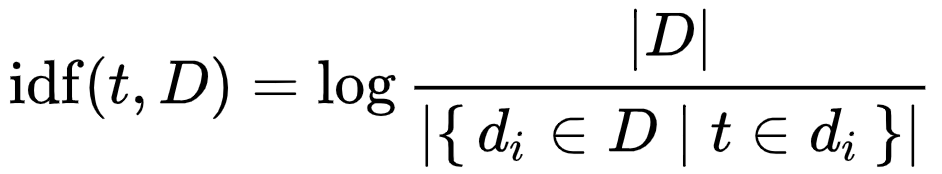
TF (Term Frequency — частота слова) - показывает насколько часто термин встречается в документе. Показывает отношение количества упоминаний слова к сумме всех слов на странице, т.е. частотность слова формула (1). Числитель - вхождение слова в документ, знаменатель - общее число слов в данном документе.

(1)



IDF (Inverse Document Frequency — обратная частота документа) - отношение всего числа документов к тем, которые имеют заданное слово. Уменьшает вес слова в зависимости от его частоты и показывает релевантность текста ключевому запросу6 формула (2).

(2)

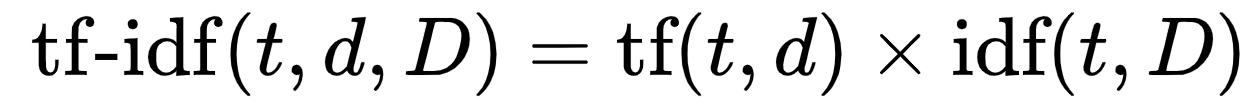


где:

* |D| - число документов в коллекции;
* - число документов из коллекции D, в которых встречается t (когда nt != 0).

В итоге получим значимость конкретного слова в пределах одного текста, формула (3).

(3)



Большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

В дополнение к «TF-IDF», алгоритм поиска дополнительно ощущает текст во время индексирования:

* убираются знаки пунктуации;
* убираются цифры;
* все слова приводятся к нижнему регистру;
* убираются двойные пробелы.

Затем для каждого документа считается TF по каждому слову. В свою очередь «IDF» считается во время запроса.

Пример работы алгоритма можно посмотреть на рисунках 1, 2 и 3.

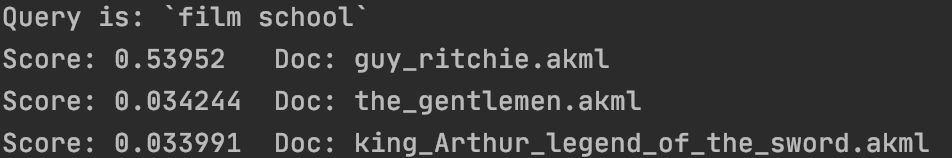


Рисунок 1 – результат поиска запроса «film school».

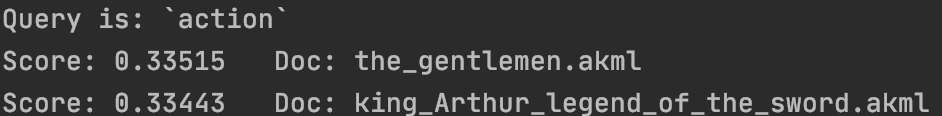


Рисунок 2 – результат поиска запроса «action».

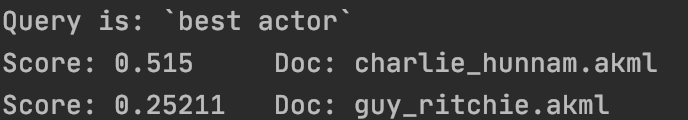


Рисунок 1 – результат поиска запроса «best actor».

**Вывод.**

В ходе выполнения лабораторной, я изучил информационно-поисковые системы, реализовал свою информационно-поисковую система на основе моедли разметки, составленной в первой лабораторной работе. В качестве метрики была использована мера tf-idf.