# Московский авиационный институт (национальный исследовательский университет)

## Факультет информационных технологий и прикладной математики

## Кафедра вычислительной математики и программирования


**Лабораторная работа №5 по курсу «Информационный поиск»**


Студент: В. М. Филиппов
Преподаватель: А. А. Кухтичев
Группа: М8О-410Б

Дата:
Оценка:
Подпись:

**Москва, 2025**

# Лабораторная работа №5 «Булев поиск»

Необходимо реализовать алгоритм булева поиска по корпусу документов при условии, что индекс сдамплен на диск.

# 1 Описание

При выполнении данной лабораторной работы, я реализовал алгоритм булева поиска по документам.

## 1 Булев поиск

Булев поиск — это модель информационного поиска, основанная на формальной логике. В этой модели запрос пользователя представляет собой логическое выражение, а результатом является множество документов, которые либо строго соответствуют условию, либо нет.

Булев поиск строится на трех базовых операторах алгебры логики:

- AND (И): Находит документы, в которых присутствуют оба слова.
- OR (ИЛИ): Находит документы, в которых есть хотя бы одно из слов (или оба сразу).
- NOT (НЕ): Исключает документы, содержащие определенное слово.

Алгоритм поиска работает следующим образом: для каждого токена из запроса мы получаем список документов. Далее преобразуем наш запрос в обратную польскую нотацию при помощи алгоритма сортировочной станции Дейкстры. Затем при помощи стэка операндов и операторов вычисляем предикат. Берем два операнда и один оператор с вершины стэка и выполняем операцию над множествами, кладём результат в стэк. Операции над множествами реализованы через два указателя, что позволяет выполнять пересечение и объединение списков за линейное время относительно их длины.

Если между двумя токенами нет никакого оператора, считается, что между ними стоит оператор AND

Из недостатков можно выделить то, что мы никак не ранжируем результаты поиска.

# 2 Исходный код

Центральным узлом системы является абстрактный класс ISearcher, реализующий паттерн "шаблонный метод". Этот паттерн позволяет зафиксировать высокоуровневый алгоритм поиска, делегируя детали реализации конкретных шагов классам-наследникам.

Основные этапы алгоритма поиска документа:

1. Лексический анализ (parseQuery). Строка запроса разбивается на токены (слова и операторы). Это позволяет абстрагировать логику обработки текста (например, удаление стоп-слов или стемминг) от логики поиска.

2. Трансформация в ОПН (sortingStation). Запрос преобразуется в обратную польскую нотацию. Этот этап необходим для корректной обработки приоритетов логических операций и управления вложенными скобками.

3. Вычисление (evaluate). Происходит непосредственное выполнение булевых операций над списками вхождений, полученными из источника данных. Результатом этого этапа является «сырой» список идентификаторов документов, которые соответствуют логическому условию запроса.

4. Пост-процессинг и ранжирование (processResults). Это ключевая точка расширения. Метод принимает список найденных документов и исходные токены запроса, возвращая финальную выдачу с весами релевантности.

## 1 Примеры запросов и результаты выдачи

В среднем запрос занимает меньше чем 0.01 секунды. Поиск работает корректно, если полазить по ссылкам, то при поиске по странице мы будем находить токены из нашего Tokens searching.

```
Enter query: manchester united
Tokens searching: [manchest, unit, ]
Top 10 results
[1] https://theguardian.com/football/2025/dec/17/premier-league-teams-africa-cup-of-nations-sunderland-morocco-chelsea-arsenal-aston-villa TF-IDF: 0
[2] https://theguardian.com/football/2025/dec/17/the-football-daily-christmas-awards-2025 TF-IDF: 0
[3] https://theguardian.com/football/2025/dec/17/the-knowledge-football-match-wham-watching-wrote-last-christmas TF-IDF: 0
[4] https://theguardian.com/football/2025/dec/14/ruben-amorim-kobbie-mainoo-loan-manchester-united-bournemouth-premier-league TF-IDF: 0
[5] https://theguardian.com/football/2025/dec/14/brentford-leeds-premier-league-match-report TF-IDF: 0
```

```
[6] https://theguardian.com/football/2025/dec/16/wsl-at-halfway-best-of-the-season-so-far-moving-the-goalposts TF-IDF: 0
[7] https://theguardian.com/football/2025/dec/16/nottingham-forest-sean-dyche-revival TF-IDF: 0
[8] https://theguardian.com/football/2025/dec/15/ruben-amorim-defends-manchester-united-defenders-despite-conceding-four TF-IDF: 0
[9] https://theguardian.com/football/live/2025/dec/15/manchester-united-v-bournemouth-premier-league-updates-live TF-IDF: 0
[10] https://theguardian.com/football/2025/dec/15/manchester-united-bournemouth-premier-league-match-report TF-IDF: 0
Query time: 0.0131275 sec
Number of results: 6065 items
```

```
Enter query: epstein
Tokens searching: [epstein, ]
Top 10 results
[1] https://theguardian.com/football/2025/dec/12/schmaltz-theatre-and-sharp-teeth-wrexham-reveal-the-hard-truth-about-football TF-IDF: 0
[2] https://theguardian.com/football/2025/dec/08/joey-barton-gets-suspended-prison-sentence-for-offensive-social-media-posts TF-IDF: 0
[3] https://en.wikipedia.org/wiki/Gatorade_shower TF-IDF: 0
[4] https://en.wikipedia.org/wiki/Sports_analytics TF-IDF: 0
[5] https://en.wikipedia.org/wiki/Mechanics_of_Oscar_Pistorius%27s_running_blades TF-IDF: 0
[6] https://en.wikipedia.org/wiki/Invention TF-IDF: 0
[7] https://en.wikipedia.org/wiki/Culture TF-IDF: 0
[8] https://en.wikipedia.org/wiki/Cultural_issues TF-IDF: 0
[9] https://theguardian.com/us-news/2021/mar/22/leon-black-quits-apollo-jeffrey-epstein-ties-inquiry TF-IDF: 0
[10] https://theguardian.com/us-news/live/2025/dec/17/donald-trump-venezuela-oil-jack-smith-fcc-jimmy-kimmel-us-politics-live-news-updates TF-IDF: 0
Query time: 0.00518029 sec
Number of results: 1479 items
```

```
Enter query: best football player
Tokens searching: [best, footbal, play, ]
Top 10 results
[1] https://theguardian.com/football/2025/dec/17/the-football-daily-christmas-awards-2025 TF-IDF: 0
[2] https://theguardian.com/football/2025/dec/17/the-knowledge-football-match-wham-watching-wrote-last-christmas TF-IDF: 0
[3] https://theguardian.com/sport/2025/dec/15/philip-rivers-indianapolis-colts-nfl-return TF-IDF: 0
[4] https://theguardian.com/football/2025/dec/15/how-the-guardian-ranked-the-100-best-male-footballers-in-the-world-2025 TF-IDF: 0
[5] https://theguardian.com/sport/2025/dec/14/all TF-IDF: 0
[6] https://theguardian.com/football/live/2025/dec/14/brentford-v-leeds-united-premier-league-live TF-IDF: 0
[7] https://theguardian.com/football/2025/dec/14/brentford-leeds-premier-league-match-report TF-IDF: 0
```

```
[8] https://theguardian.com/football/2025/dec/16/wsl-at-halfway-best-of-the-season-so-far-moving-the-goalposts TF-IDF: 0
[9] https://theguardian.com/sport/2025/dec/16/all TF-IDF: 0
[10] https://theguardian.com/sport/2025/dec/11/shedeur-sanders-cleveland-browns-quarterback-nfl TF-IDF: 0
Query time: 0.0226403 sec
Number of results: 6711 items
```

```
Enter query: (manchester & united) | messi
Tokens searching: [manchest, unit, messi, ]
Top 10 results
[1] https://theguardian.com/football/2025/dec/17/premier-league-teams-africa-cup-of-nations-sunderland-morocco-chelsea-arsenal-aston-villa TF-IDF: 0
[2] https://theguardian.com/football/2025/dec/17/the-football-daily-christmas-awards-2025 TF-IDF: 0
[3] https://theguardian.com/football/2025/dec/17/the-knowledge-football-match-wham-watching-wrote-last-christmas TF-IDF: 0
[4] https://theguardian.com/sport/2025/dec/17/the-anti-sports-personality-of-the-year-awards-2025 TF-IDF: 0
[5] https://theguardian.com/football/2025/dec/14/ruben-amorim-kobbie-mainoo-loan-manchester-united-bournemouth-premier-league TF-IDF: 0
[6] https://theguardian.com/football/2025/dec/14/brentford-leeds-premier-league-match-report TF-IDF: 0
[7] https://theguardian.com/football/2025/dec/16/wsl-at-halfway-best-of-the-season-so-far-moving-the-goalposts TF-IDF: 0
[8] https://theguardian.com/sport/2025/dec/16/anthony-joshua-jake-paul-miami-joe-louis-al-mccoy-boxing TF-IDF: 0
[9] https://theguardian.com/football/2025/dec/16/nottingham-forest-sean-dyche-revival TF-IDF: 0
[10] https://theguardian.com/football/2025/dec/15/ruben-amorim-defends-manchester-united-defenders-despite-conceding-four TF-IDF: 0
Query time: 0.0156707 sec
Number of results: 7561 items
```

Вот здесь видим проблему со стэммингом. Гарэт Бейл из футболиста превратился в мяч, поэтому в выдаче мы видим статью про баскетбол.

```
Enter query: !messi | ronaldo & Bale
Tokens searching: [messi, ronaldo, bal, ]
Top 10 results
[1] https://theguardian.com/sport/2025/dec/17/nba-cup-takeaways-spurs-knicks-victor-wembanyama TF-IDF: 0
[2] https://theguardian.com/sport/2025/dec/17/all TF-IDF: 0
[3] https://theguardian.com/football/2025/dec/17/premier-league-teams-africa-cup-of-nations-sunderland-morocco-chelsea-arsenal-aston-villa TF-IDF: 0
[4] https://theguardian.com/football/2025/dec/17/celtic-chair-peter-lawwell-to-stand-down-after-intolerable-abuse-from-fans TF-IDF: 0
[5] https://theguardian.com/football/2025/dec/17/fifa-50-per-cent-increase-2026-world-cup-prize-money-50m-dollars-winners TF-IDF: 0
[6] https://theguardian.com/football/2025/dec/17/the-football-daily-christmas-awards
```

```
    -2025 TF-IDF: 0
[7] https://theguardian.com/football/2025/dec/17/brendan-rodgers-saudi-arabia-pro-league-al-qadsiah-aramco TF-IDF: 0
[8] https://theguardian.com/football/2025/dec/17/macclesfield-footballer-ethan-mcleod-21-dies-in-car-accident TF-IDF: 0
[9] https://theguardian.com/sport/2025/dec/17/khawaja-carey-rise-up-as-england-squander-australia-ashes-gifts TF-IDF: 0
[10] https://theguardian.com/sport/2025/dec/17/jofra-archer-england-australia-third-ashes-test-cricket TF-IDF: 0
Query time: 0.0422179 sec
Number of results: 303572 items
```

# 3 Выводы

За время выполнения этой лабораторной работы, я узнал что такое булев поиск, реализовал алгоритмы пересечения и объединения множеств методом двух указателей. Это обеспечило линейную сложность обработки списков вхождений, что критически важно для производительности системы. Вспомнил как работает алгоритм сортировочной станции Дейкстры.

# Список литературы

[1] Маннинг К. Д., Рагхаван П., Шютце Х. *Введение в информационный поиск* : пер. с англ. — М. : ООО «И.Д. Вильямс», 2011. — 528 с.

[2] Польская нотация или как легко распарсить алгебраическое выражение [Электронный ресурс] // Хабр. — 2021. — URL: https://habr.com/ru/articles/596925/ (дата обращения: 17.12.2025).

[3] Boolean search для чайников и кофейников [Электронный ресурс] // Хабр. — 2023. — URL: https://habr.com/ru/articles/716968/ (дата обращения: 17.12.2025).