

**Московский авиационный институт
(национальный исследовательский университет)**

**Институт информационные технологии и прикладной
математики**

Кафедра вычислительной математики и программирования

**Лабораторная работа №2 по курсу
«Информационный поиск»**

Студент:	Е.М. Стифеев
Преподаватель:	А.А. Кухтичев
Группа:	М8О-109М-21
Дата:	29.11.21
Оценка:	
Подпись:	

Москва, 2021

Лабораторная работа №2 «Оценка качества поиска»

Необходимо оценить качество своего поиска и сравнить их с двумя альтернативами (для Википедии можно собственный поиск по Википедии, поиск Google или Яндекса с ограничением по сайту Википедии). Как минимум, нужно измерить P, DCG, NDCG и ERR уровней @1, @3 и @5, приветствуется использование дополнительных метрик качества.

Для оценки качества необходимо придумать 30 запросов, отражающих интересы пользователей или, если есть доступ к настоящим запросам пользователей, то выбрать репрезентативную подборку.

В качестве примера посмотрите на 10 запросов к поиску по всей Википедии, подумайте о том, почему именно они были выбраны и какую сложность для поисковой системы они представляют:

1. [из каких книг состоит библия]
2. [что где когда]
3. [игра]
4. [российские авиазаводы]
5. [без меня народ не полный]
6. [как называют жителей набережных челнов]
7. [где короновали николая 2]
8. [товарищ прокурора]
9. [цари газы]
- 10.[административный кодекс]

Проведите анализ результатов оценки качества. Какие у какой поисковой системы сильные и слабые стороны? Как можно бороться с недостатками, что можно сделать, чтобы улучшить качество?

Описание

Корпус

Напомню, что корпус документов имеет следующую структуру, полученную по результатам ЛР1 (доступен по ссылке <https://cloud.mail.ru/public/ZfkX/gccM7hnDR>):

- Корпус документов
 - films1.txt (94 Мб, 15000 документов)
 - films2.txt (96 Мб, 15000 документов)
 - films3.txt (184 Мб, 15000 документов)
 - films4.txt (219 Мб, 15000 документов)
 - films5.txt (322 Мб, 15000 документов)
 - films6.txt (711 Мб, 15000 документов)
 - films7.txt (823 Мб, 15000 документов)
 - films8.txt (226 Мб, 15000 документов)
 - films9.txt (67 Мб, 15000 документов)
 - films10.txt (75 Мб, 15000 документов)
 - films11.txt (99 Мб, 15000 документов)
 - films12.txt (78 Мб, 15000 документов)
 - films13.txt (41 Мб, 6109 документов)

$$\Sigma_{Gb} = 2,899 \text{ Gb}, \Sigma_{docs} = 186109$$

Также, напомню, что получение одного документа могло включать проход по нескольким html-страницам и обработку динамически подгружаемых страниц, поэтому общее количество обработанных страниц было >800'000.

Корпус документов собран с трёх сайтов:

- <https://www.kinopoisk.ru/>
- <https://shikimori.one/>
- <http://filmplace.ru/>

В каждом файле *.txt документы хранятся следующим образом:

- 1 строка 1 документ {....}
- 2 строка 2 документ {....}
- n строка n документ {....}

Каждый документ снабжён прямой ссылкой на источник, откуда был скачен, и хранит только выделенный из html-кода текст в кодировке UTF-8. Например, 234 строка файла films1.txt выглядит так:

```
{
  "page_url": "https://www.kinopoisk.ru/media/article/1773537/",
  "title": "Артур Смольянинов: «Я сомневался, что смогу сыграть ангела»",
  "body": "2 января в российский прокат вышла романтическая комедия Веры Сторожевой „Мой парень — ангел“, главные роли в которой исполнили Артур Смольянинов и Анна Старшенбаум. Мы подготовили небольшой видеосюжет с участием создателей картины...Студентка Саша с большим трудом верит в чудеса. Ангелу Серафиму приходится приложить немало усилий, чтобы доказать ей, что ангелы существуют. Но он не учел одного: если девушка тебе поверит, она, скорее всего, тебя полюбит.",
  "author": "Дарико Цулая",
  "comments": ""
}
```

Индекс

Готовый индекс хранится в четырёх файлах:

- **docs_id.data** (42 Мб)

Файл служит для отображения индекса документа (doc_id) в его текстовое представление в файлах *.txt. Поддерживается переменная длина пути до файлов с документами.

- **terms.data** (54 Мб)

Файл служит для хранения словаря с терминами и ссылок (смещений) на файл с словопозициями и координатами. Поддерживается переменная длина термина. Термины упорядочены в лексикографическом порядке.

- **postings_list.data** (2.68 Гб)

Файл служит для хранения словопозиций и координат терминов в документе. Словопозиции упорядочены по возрастанию идентификаторов документов. Координаты упорядочены по возрастанию документа.

- **tf.data** (939 Мб)

Файл служит для быстрого получения компонент документа, как вектора в пространстве терминов. Он нужен для быстрого ранжирования на основе косинуса между вектором запроса и вектором документа. См. подробности в ЛР по ранжированию.

Готовый индекс доступен по ссылке
<https://cloud.mail.ru/public/wynT/adagiBjh9>

Запросы

Учитывая тематику корпуса были придуманы следующие запросы (в форме для обработки поисковиком):

Номер	Запрос	Вид
1	"тихое место 2"	цитатный
2	аватар скачать	фразовый (транслируется в нечёткий)
3	"re zero" / 4	цитатный
4	режиссёр назад в будущее	фразовый
5	фильм для интеллектуалов	фразовый
6	фильм сериал с самым большим рейтингом	фразовый
7	фильмы Макото Синкая	фразовый
8	лучшие фильмы квентина тарантино	фразовый
9	как звали главного героя коносубы	фразовый
10	самый лучший фильм	фразовый
11	сериалы с рейтингом 18+	фразовый
12	высшая школа демонов	фразовый
13	джокер	фразовый
...

Далее, запросы были адресованы поисковой системе Yandex с ограничением по сайтам, дате и с приведением к соответствующему виду.

Например, первый запрос из списка был представлен в форме:

"тихое место 2" (url:www.kinopoisk.ru/film/* | url:www.kinopoisk.ru/media/* | site:filmlplace.ru | site:shikimori.one) date:<20210930

**Тихое место 2 (2020) — смотреть онлайн — Кинопоиск**[kinopoisk.ru > film/1129900/](https://www.kinopoisk.ru/film/1129900/) ...

Были вчера · **Тихое место 2** (2020). A Quiet Place Part II 16+. Семья Эббот ищет новое укрытие для ... Поймав радиосигнал, Реган вычисляет **место** предполагаемой колонии выживших и решает во что бы то ни стало её разыскать. Рейтинг фильма. Читать ещё

[Сборы](#) · [Рецензии зрителей](#) · [Связи](#) · [Трейлеры](#) · [Саундтреки](#)

**Тихое место 2 (2020) — актеры и съемочная группа**[kinopoisk.ru > film/1129900/cast/](https://www.kinopoisk.ru/film/1129900/cast/) ...

Тихое место 2 (2020). Актеры, режиссер, продюсеры и другие участники съемочной группы.

**На высоте мечты (2021) — смотреть онлайн — Кинопоиск**[kinopoisk.ru > film/464302/](https://www.kinopoisk.ru/film/464302/) ...

Были 27 ноя · Жанр: мюзикл, мелодрама, драма. Режиссёр: Джон М. Чу. В ролях: Энтони Рамос, Мелисса Баррера, Лесли Грейс. Вашингтон-Хайтс. Запах горячего кофе витает в воздухе у станции метро «181-я улица», а калейдоскоп желаний собирает там шумных и дружных...

[В главных ролях](#) · [Рецензии](#) · [Трейлеры](#) · [Студии](#) · [Слова](#) · [Связи](#)

★ ★ ★ ★ ★ 6,7 из 10 — 614 оценок · 2 ч 23 мин · 2021 · США

**На старт! (2021) — трейлеры, даты премьер — Кинопоиск**[kinopoisk.ru > film/1323863/](https://www.kinopoisk.ru/film/1323863/) ...

Жанр: драма, комедия, спорт. Режиссёр: Чу Кэн Гуань. В ролях: Ван Яньхуэй, Чжан Юохао, Гун Бэйби. Отец-одиночка с сыном принимают участие в марафонском забеге.

1 ч 44 мин · 2021 · Китай

**Тихое место (2018) — смотреть онлайн — Кинопоиск**[kinopoisk.ru > film/1044906/](https://www.kinopoisk.ru/film/1044906/) ...

Жанр: ужасы, фантастика, драма. Режиссёр: Джон Красински. В ролях: Эмили Блант, Джон Красински, Миллисент Симмондс. Семья с двумя детьми живёт на отдалённой ферме. Казалось бы, жизнь этих людей совершенно не отличается от жизни других таких семей, но они...

[Все награды](#) · [Рецензии](#) · [Трейлеры](#) · [Связи](#) · [Сборы](#) · [В России](#)

★ ★ ★ ★ ★ 6,8 из 10 — 238 тыс. оценок · 1 ч 30 мин · 2018 · США

**Что смотреть в 2020 году: 50 самых ожидаемых фильмов...**[kinopoisk.ru > media/article/4000204/](https://www.kinopoisk.ru/media/article/4000204/) ...

Были 27 ноя · В команде Квинн нашлось **место** Черной Канарейке (Джерни Смоллетт), Охотнице (Мэри Элизабет Уинстед) и Рене Монтойе (Рози Перес). Читать ещё

**Что смотреть в 2021 году: 50 самых ожидаемых фильмов...**[kinopoisk.ru > media/article/4003767/](https://www.kinopoisk.ru/media/article/4003767/) ...

Тогда же выпустят «**Тихое место 2**» (22 апреля), «Черную Вдову» (6 мая), «Неизведанное: Удача Дрейка» (15 июля) и хоррор Эдгара Райта «Прошлой ночью...» Читать ещё

Нашлось 46 результатов

[Разместить рекламу](#)

Далее, для каждого запроса из списка вручную был отобран список релевантных ссылок, ввиду того, что поисковая система Яндекс'а даёт сбои.

```

1 =====
2 "тихое место 2"
3 =====
4 https://www.kinopoisk.ru/film/1129900/
5 http://filmlplace.ru/film/tihoe-mesto-2.html
6 https://www.kinopoisk.ru/film/1044906/
7 https://www.kinopoisk.ru/media/news/4000234/
8 https://www.kinopoisk.ru/media/news/4000472/
9 https://www.kinopoisk.ru/media/news/4004687/
10 https://www.kinopoisk.ru/media/news/3414753/
11 https://www.kinopoisk.ru/media/news/3390248/
12 https://www.kinopoisk.ru/media/news/3344771/
13 https://www.kinopoisk.ru/media/news/4000865/
14 =====
15 аватар скачать
16 =====
17 http://filmlplace.ru/film/avatar-1.html
18 http://filmlplace.ru/film/avatar.html
19 http://filmlplace.ru/film/legenda-o-korre.html
20 http://filmlplace.ru/film/povelitel-stihij.html
21 http://filmlplace.ru/film/avatar-poslednij-mag-vozduha.html
22 https://www.kinopoisk.ru/media/news/1434058/
23 https://www.kinopoisk.ru/media/news/1766571/
24 https://www.kinopoisk.ru/media/news/1124915/
25 http://filmlplace.ru/film/avatar-sozdanie-mira-pandoryi.html
26 http://filmlplace.ru/film/kenau.html
27 =====
28 "re zero" / 4
29 =====

```

С полным списком запросов и релевантных ссылок читатель может ознакомиться по ссылке: <https://cloud.mail.ru/public/Hp9R/KMn7Vo1HJ> (эта директория также подаётся на вход программе-оценщику).

Далее, запросы из списка были адресованы поисковой системе, созданной по результатам ЛР по курсу.

Поиск по корпусу документов

Параметры

“Тихое место 2”

Поиск Загрузить индекс Построить индекс Выполнить сканирование

Ссылка	Заголовок	Подробности
https://www.kin...	Видео о съемках «Тихого места 2»: Бесшумный ап...	.. "Видео о съемках «ТИХОГО МЕСТА 2»: Бесшумный апокали...
https://www.kin...	Режиссер «Мада» Джефф Николс напишет и пост...	..фильм во вселенной «ТИХОГО МЕСТА», "body": "Джефф Н...жении «Тихого места 2», а о новой истории...
https://www.kin...	«В мире есть люди, которых надо спасти»: Трейле...	..о спасении: Трейлер «ТИХОГО МЕСТА 2», "body": "Оригина...
https://www.kin...	Завершились съемки «Тихого места 2»	..Завершились съемки «ТИХОГО МЕСТА 2», "body": "Режиссе...
https://www.kin...	Стартовали съемки «Тихого места 2»	.. "Стартовали съемки «ТИХОГО МЕСТА 2», "body": "Джон Кр...
https://www.kin...	Киллиан Мерфи может сыграть в сиквеле «Тихог...	.. сыграть в сиквеле «ТИХОГО МЕСТА», "body": "Звезда ...емки «Тихого места 2» начнутся этим лето...
https://www.kin...	Тизер «Тихого места 2»: Эмили Блант идет к неизв...	..", "title": "Тизер «ТИХОГО МЕСТА 2»: Эмили Блант идет ..
https://www.kin...	«Топ Ган: Мэверик» перенесли на декабрь 2020-го...	..а декабрь 2020-го. «ТИХОЕ МЕСТО 2» выйдет в сентябре" ..
https://www.kin...	Релиз триллера «Тихое место 2» перенесли из-за "Релиз триллера «ТИХОЕ МЕСТО 2» перенесли из-за ко...
https://www.kin...	«Большинство людей уже потеряли надежду»: Фи...	.. Финальный трейлер «ТИХОГО МЕСТА 2» с Эмили Блант и Ки...
https://www.kin...	Paramount отложила «Тихое место 2» и «Топ Ган:Paramount отложила «ТИХОЕ МЕСТО 2» и «Топ Ган: Мэверик...
https://www.kin...	Премьеру «Форсажа 9» перенесли на год из-за ко...	.. Везде затишье, но «ТИХОЕ МЕСТО 2» продолжает собират...
https://www.kin...	Что происходит в прокате: «Тихое место 2» — лид...	..исходит в прокате: «ТИХОЕ МЕСТО 2» — лидер в мире и в...
https://www.kin...	Paramount на год отложила релиз сиквела «Лучш...	..ама «Лучший стрелок 2» на год. Сиквел уви...релиза для сиквела «ТИХОГО МЕСТА». Продолжение хорро...
https://www.kin...	Опрос: Американские зрители предпочтут посмо...	..ема «умирать» — 18% «ТИХОЕ МЕСТО 2» — 12% «Форсаж 9» — ...
https://www.kin...	Слух дня: Джон Красински встречался с Marvel истановщик триллера «ТИХОЕ МЕСТО» Джон Красински мож...у над «Тихим местом 2». Лента ожидалась в...
https://www.kin...	«Оскар-2019»: Лонг-лист фильмов с лучшими виз...	..Мир Юрского периода 2"/n"Миссия невыполнима...Суперсемейка 2"/n"ТИХОЕ МЕСТО"/n"Удивительный ми...
https://www.kin...	Новые части «Миссия: невыполнима» отложилие студия перенесла «ТИХОЕ МЕСТО 2», «Топ Ган: Мэверик...
https://www.kin...	Universal перенесла сиквел «Миньонов» на 2021 г...	..2021 год. «Зверопой 2» отложен на полгода...1984», «Форсаж 9», «ТИХОЕ МЕСТО 2», «Кролик Питер ..
https://www.kin...	Джон Красински организовал онлайн-выпускнойакончил работу над «ТИХИМ МЕСТОМ 2». Лента ожидалась в...
https://www.kin...	Вместе в кинотеатрах 2 апреля 2021-го. При...nt перенесла выход «ТИХОГО МЕСТА 2». Universal сдвин...	..боту над триллером «ТИХОЕ МЕСТО 2». Лента ожидалась в...
https://www.kin...	Джон Красински запустил передачу только с хоро...	..вет «Грани будущего 2» — продолжению фант... Датом на премьере «ТИХОГО МЕСТА 2» «Действие «Грани...
https://www.kin...	Эмили Блант надеется, что сиквел «Грани будуще...	..046777", "title": "«ТИХОЕ МЕСТО 2» — самый страшный ф...
https://www.kin...	«Тихое место 2» — самый страшный фильм лета.деть свет 26 марта, 2 апреля и 15 апреля ...залась от премьеры «ТИХОГО МЕСТА 2» 19 марта и пока ..
https://www.kin...	Из-за коронавируса отменены релизы «Мулан» икритики про сиквел «ТИХОГО МЕСТА», "body": "В сети ...нении, «Тихое место 2» — зрелищный, страш...
https://www.kin...	«Не так хорошо, как первая часть»: Что пишут крит...	..за отмены премьеры «ТИХОГО МЕСТА 2» Paramount потеряет...
https://www.kin...	Что происходит в прокате: «Форсаж» передвинул...	.. вернул себе первое МЕСТО в списке крупнейших...ае.Новые переносы: «ТИХОЕ место 2» в мае, «Форсаж 9» ..
https://www.kin...	Бокс-офис России: Горилла не терпит тишины	..енно с ним хоррора «ТИХОЕ МЕСТО». Притом что у посл...хого места» — всего 2 млн долларов, 34,5 ..
https://www.kin...	Опрос: 42% зрителей готовы вернуться в кинотеат...	.. 1984» — 34% / 25% «ТИХОЕ МЕСТО 2» — 27% / 19% «После...
https://www.kin...	Выход «Морбиуса» отложили на осень. СМИ пол...	..сно. Кроме того, на 2 апреля заявлена пре...ланирует продавать «ТИХОЕ МЕСТО 2» (оно должно выйт...
https://www.kin...	Американский бокс-офис: Затишье перед войной	..ендантами на первые МЕСТА. Так что вторую нед...пасает мир./n/n/n1 (2) «ТИХОЕ место», 22 млн долл...
https://www.kin...	Американский бокс-офис: Танос собрал камни иларов). На третьем МЕСТЕ находится «Черная П... млн долларов./n/n/n2 (NEW) «За бортом», ...«За бортом»/n3 (2) «ТИХОЕ мес...
https://www.kin...	Молчание (2019)	..ышали, как я ругала ТИХОЕ МЕСТО, и учли все мои зам...10", "Тихое место"-2/nМонстры, реагирующ...
https://www.kin...	Что смотреть дома: «Чудотворцы: Диний Запад», «...	..ры: Диний Запад», «ТИХОЕ МЕСТО 2», «Красный призрак...
https://www.kin...	7 лучших трейлеров недели: Зрелища Верховена,Каннском фестивале «ТИХОЕ МЕСТО 2»Сиквел кассового хи...
https://www.kin...	Что происходит в прокате: «Форсаж 9» набирает с...	..стартовал на первом МЕСТЕ еще в нескольких ст...а с Эммой Стоун и «ТИХОЕ место 2». Так Голливуд откр...
https://www.kin...	Трейлер комедии с Эдди Мерфи и еще 10 новосте...	..йана Тайри Генри в «ТИХОМ МЕСТЕ 2»Фильм рассказывает ..
https://www.kin...	Что происходит в прокате: В Китае новый рекорд.смогла заработать 2 млн долларов. Оказа... невыполнима 7» и «ТИХОЕ МЕСТО 2». Оба фильма выйд...
https://www.kin...	Что происходит в прокате: Наташа Романофф ста...	..лла против Конга», «ТИХОГО МЕСТА 2» и «Форсажа 9». Дис...
https://www.kin...	Тихое место 2 (2020)	..129900/", "title": "ТИХОЕ МЕСТО 2 (2020)", "alternati...

1/1

Извлекая ссылки из ответов и сравнивая их с ссылками, полученными Yandex'ом можно высчитать требуемые метрики.

Метрики

Для просчёта метрик были взяты формулы из статьи [2].

Пусть

$$r: E \rightarrow [0,1],$$

функция переводящая документ $d \in E$ в число – релевантность запроса. В моём случае множество значений было решено ограничить значениями $\{0,1\}$ (нерелевантен/релевантен);

$$\pi: E \rightarrow [1, K],$$

функция сопоставляющая документу $e \in E$ позицию в отсортированном по убыванию косинус-веса списке найденных документов, ограниченных уровнем K .

Тогда

$$p@K = \frac{1}{K} \sum_{k=1}^K r(\pi^{-1}(k));$$

$$DCG@K = \sum_{k=1}^K \frac{r(\pi^{-1}(k))}{\log_2 k + 1};$$

$$IDCG@K = \sum_{k=1}^K \frac{1}{\log_2(k + 1)};$$

$$NDCG@K = \frac{DCG@K}{IDCG@K};$$

$$ERR@K = \sum_{k=1}^K \frac{1}{k} r(\pi^{-1}(k)) \prod_{i=1}^{k-1} (1 - r(\pi^{-1}(i))).$$

Поскольку число запросов больше чем 1, в конце проводится усреднением по всем метрикам для множества запросов Q .

1. Исходный код

Структура проекта

- include
 - algebra.h (примитивные операции с математическими векторами)
 - create_index.hpp (создание, чтение индекса)
 - defs.hpp (подключение внешних библиотек, макросы)
 - docs_parse.hpp (извлечение полей из корпуса и индекса)
 - search.hpp (поиск)
 - token_parse.hpp (функции для преобразования токенов в термы)
- python
 - lemmatizator.py (лемматизация документа)
 - lemmatizator_setup.py (компиляция lemmatizator.py в exe-файл)
 - request_parse.py (лемматизация запроса)
 - request_parse_setup.py (компиляция request_parse.py в exe-файл)
- io
 - answers.txt
 - requests.txt
- src
 - main.cpp (точка входа в консольный интерфейс)

Проект был написан с помощью Microsoft Visual Studio 2019 эксклюзивно для ОС семейства Windows.

Запуск

Консольное приложение поддерживает флаги запуска:

- -i 'путь к корпусу'
- -o 'путь к индексу'
- -t 'путь к директории с блочным индексом'
- -m 'путь к директории с эталонами для метрик'
- -p кол-во_процессов_для_распараллеливания
- -create : создать блочный индекс
- -merge : выполнить слияние блочного индекса
- -clear : очистить папку с временными файлами после слияния
- -search : выполнить поиск
- -metric : высчитать метрики

Пример создания блочного индекса из корпуса:

```
$ ./prog.exe -p 4 -create -i "..\..\Корпус" -o -t "tmp"
```

Вывод

```
[INFO] Создание индекса для блоков
```

```
[INFO] Thread 0 processing block 1/13 : D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус\films13.txt
```

```
[INFO] Thread 1 processing block 2/13 : D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус\films9.txt
```

```
[INFO] Thread 2 processing block 3/13 : D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус\films10.txt
```

```
[INFO] Thread 3 processing block 4/13 : D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус\films12.txt
```

```
[INFO] Block 1 has 232216 terms
```

```
<...>
```

```
[INFO] Block 10 has 597482 terms
```

```
[INFO] Block 11 has 669497 terms
```

```
[INFO] Block 12 has 921883 terms
```

```
[INFO] Block 13 has 1383148 terms
```

```
[INFO] Создание очередей термов: 13 блок из 13
```

[INFO] Слияние docs_id: 13 блок из 13
[INFO] Слияние слопозиций термов
[INFO] Осталось термов: 0
[INFO] Очистка временных файлов
[INFO] Общее число термов в словаре = 2809203
Время выполнения = 145,5 sec, размер корпуса = 2,899 Gb, документов = 186109
Средняя скорость на документ = 0,782 ms
Средняя скорость на килобайт = 0,048 ms

Пример слияние блочного индекса:

```
$ ./prog.exe -p 4 -merge -clear -i "..\..\Корпус_index" -t "tmp"
```

Вывод

[INFO] Слияние блочного индекса
[INFO] Создание очередей термов: 13 блок из 13
[INFO] Слияние docs_id: 13 блок из 13
[INFO] Слияние слопозиций термов
[INFO] Осталось термов: 0
[INFO] Общее число термов в словаре = 1908410
Документов = 186109
[INFO] Время на слияние блочного индекса: 35 sec
[INFO] Вычисление статистики
Первый проход. Термов осталось: 0
Второй проход. Документов осталось: 0
[INFO] Вычисление статистики закончено

Пример просчёта метрик:

```
$ ./prog.exe -p 4 -metric -i "..\..\Корпус" -o "..\..\Корпус_index"  
-m "..\..\Корпус_metric"
```

Вывод

Чтение термов : [#####] 1908410/1908410
[INFO] Загружено 1908410 термов
Запрос: ' "тихое место 2" '
По запросу: ' " тихий место 2 " ' найдено 50 документов
Запрос: ' аватар скачать '
По запросу: ' !! аватар скачать ' найдено 474 документов
Запрос: ' "re zero" / 4 '
По запросу: ' " re zero " / 4 ' найдено 266 документов
Запрос: ' режиссёр назад в будущее '
По запросу: ' !! режиссер назад в будущее ' найдено 6233 документов
Запрос: ' фильм для интеллектуалов '
По запросу: ' !! фильм для интеллектуал ' найдено 1286 документов
Запрос: ' фильм сериал с самым большим рейтингом '
По запросу: ' !! фильм сериал с самый больший рейтинг ' найдено 4820 документов
Запрос: ' фильмы Макото Синкая '
По запросу: ' !! фильм макото синкай ' найдено 271 документов
Запрос: ' лучшие фильмы квентина тарантино '
По запросу: ' !! хороший фильм квентин тарантино ' найдено 1266 документов
Запрос: ' как звали главного героя коносубы '
По запросу: ' !! как звать главный герой коносуб ' найдено 85 документов
Запрос: ' самый лучший фильм '
По запросу: ' !! самый хороший фильм ' найдено 54490 документов
Запрос: ' сериалы с рейтингом 18+ '
По запросу: ' !! сериал с рейтинг 18 ' найдено 2061 документов
Запрос: ' высшая школа демонов '
По запросу: ' !! высокий школа демон ' найдено 1478 документов

Запрос: ' джокер '

По запросу: ' !! джокер ' найдено 1353 документов

Точность на уровне 30 = 0.241026

DCG на уровне 30 = 2.655877

nDCG на уровне 30 = 0.289893

ERR на уровне 30 = 0.732372

2. Выводы

Метрика \ Уровень	1	3	5	30
P	0.62	0.43	0.4	0.241
DCG	0.62	1.011	1.29	2.65
nDCG	0.62	0.47	0.44	0.29
ERR	0.62	0.71	0.71	0.73

В целом система работает удовлетворительно и со своими задачами справляется. Из улучшений остаётся: автоисправление опечаток (см. курсовую работу), автораспознавание вида запроса (булев, цитатный, нечёткий), ускорение ранжирования, ускорение индексации путём отказа от Natash'и (4 часа на полную индексацию корпуса в многопоточном режиме), сжатие индекса, прыжки по индексу, более умные сниппеты, зонный поиск и прочее.

В ходе выполнения лабораторной работы я научился считать метрики ранжирования для коллекции документов.

Литература

- [1] Кристофер Д.Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск. 2020, изд. Вильямс.
- [2] <https://habr.com/ru/company/econtenta/blog/303458/>