

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

**Институт информационные технологии и прикладной
математики**

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Информационный поиск»

Студент:	Е.М. Стифеев
Преподаватель:	А.А. Кухтичев
Группа:	М8О-109М-21
Дата:	11.10.21
Оценка:	
Подпись:	

Москва, 2021

Лабораторная работа №1 «Добыча корпуса документов»

Необходимо подготовить корпус документов, который будет использован при выполнении остальных лабораторных работ:

- Скачать его к себе на компьютер. В отчёте нужно указать источник данных.
- Ознакомиться с ним, изучить его характеристики. Из чего состоит текст? Есть ли дополнительная мета-информация? Если разметка текста, какая она?
- Разбить на документы.
- Выделить текст.
- Найти существующие поисковики, которые уже можно использовать для поиска по
- выбранному набору документов (встроенный поиск Википедии, поиск *Google* с использованием ограничений на *URL* или на сайт). Если такого поиска найти невозможно, то использовать корпус для выполнения лабораторных работ нельзя!
- Привести несколько примеров запросов к существующим поисковикам, указать недостатки в полученной поисковой выдаче.

В результатах работы должна быть указаны статистическая информация о корпусе:

- Размер «сырых» данных.
- Количество документов.
- Размер текста, выделенного из «сырых» данных.
- Средний размер документа, средний объём текста в документе.

1. Описание

Для выполнения поставленной задачи требуется написать веб-скрапер. После длительных размышлений над тематикой документов было принято решение «обкачать» сайты с кинотематикой. И, конечно же, в качестве первого кандидата был выбран крупнейший сборник подобных документов ру-сегмента интернета *kinopoisk* (<https://www.kinopoisk.ru/>).

Рассмотрим структуру страницы о фильме на примере <https://www.kinopoisk.ru/film/64021/> (х/ф «Зомби по имени Шон»). Открыв её в браузере, наблюдаем следующую интересующую нас информацию для скрапинга:

- Название фильма («Зомби по имени Шон»);
- Альтернативное название («Shaun of the Dead»);
- Год производства (2004);
- Средняя оценка пользователей (7.4);
- Страна/страны производства фильма (Великобритания, Франция);
- Режиссёр/режиссёры (Эдгар Райт);
- Сценарист/сценаристы (Саймон Пегг, Эдгар Райт);
- Актёры (Саймон Пегг, Ник Фрост, ...);
- Жанры (ужасы, комедия);
- Описание («В жизни Шона все идет наперекосяк ...»);
- Рецензии («Пересматриваю эту пародию ритуально ...»).

Теперь нужно каким-то образом найти ссылки на все фильмы из базы. Для этого была найдена карта сайта <https://www.kinopoisk.ru/sitemaps/sitemap.xml> (ссылка на неё указана в <https://www.kinopoisk.ru/robots.txt>), откуда был извлечён список ссылок на все фильмы, хранящиеся на сайте. Получив такой список, можно «распарсить» каждую ссылку, выделив нужные сведения из html-кода, совмещая такие техники как:

- Получение кода *html*-кода страницы (*Scrapy*, *Selenium*)
- Исследование кода страницы через встроенный в браузер (в моём случае *Yandex Browser*) инспектор сайта, который помогает понять, какой блок *html*-кода, за какой элемент страницы отвечает.
- Средства анализа древовидной структуры страницы (*BeautifulSoup*) с целью извлечения нужных тэгов.
- Сохранение информации в локальный файл (в моём случае *json*).

В процессе выполнения работы автор столкнулся с такими проблемами, как:

- Динамически подгружающая секциями с отзывами.

Сначала был написан скрипт, использующий фреймворк *Selenium* для выполнения *java script*'а на странице, работающий по принципу: загрузить страницу, проскроллить в её конец, подождать пока на странице не появится нужный блок с отзывами (по появлению определённого тэга) и тогда извлечь отзывы из блока, либо пока не пройдёт, например, 5 секунд, и тогда завершить обработку страницы без отзывов.

Затем автор выяснил по карте сайта, что есть ссылки на страницы, на которых находятся только отзывы <https://www.kinopoisk.ru/film/64021/reviews/> (64021 – уникальный номер фильма). В связи с чем было принято решение обрабатывать по две страницы на фильм (фильм + отзывы) с секундной задержкой.

- Капча.

При её появлении было принято решение пропускать обработку таких страниц, занося ссылку в отдельный файл с целью ручного распознавания в будущем (таких страниц было 1% от всех). Также можно воспользоваться таким инструментом как *tesseract* для автоматической обработки (это также требует ручной обработки определённого количества изображений).

В дальнейшем планируется пополнить список подобных сайтов и провести с ними аналогичную работу по извлечению нужных документов. Для текущего сайта можно ещё обработать новостную ленту.

2. Исходный код

Структура проекта:

- film_scraper
 - film_scraper
 - captcha (файлы со ссылками на страницы с капчей)
 - sitemap_captcha_links1.txt
 - jsons (выходные файлы)
 - films1.jsonlines
 - sitemaps (файлы со ссылками для обработки пауком)
 - sitemap_film_review1.txt
 - sitemap_film_review2.txt
 - sitemap_film1.txt
 - ...
 - sitemap_film18.txt
 - spiders (пауки)
 - __init__.py
 - FSpyder.py
 - __init__.py
 - items.py
 - middlewares.py
 - pipelines.py
 - setting.py (глобальные настройки)
 - scrapy.cfg

Итоговый «паук» требует следующие этапы настройки:

- Нужно создать txt-файл со списком ссылок для обработки
(\film_scraper\film_scraper\sitemaps\sitemap_film_review1.txt)

```
sitemap_film_review1.txt x
1 https://www.kinopoisk.ru/film/1178971/
2 https://www.kinopoisk.ru/film/52818/
3 https://www.kinopoisk.ru/film/16747/
4 https://www.kinopoisk.ru/film/11910/
5 https://www.kinopoisk.ru/film/98888/
6 https://www.kinopoisk.ru/film/251602/
7 https://www.kinopoisk.ru/film/583647/
8 https://www.kinopoisk.ru/film/269869/
9 https://www.kinopoisk.ru/film/57980/
10 https://www.kinopoisk.ru/film/450732/
11 https://www.kinopoisk.ru/film/44550/
12 https://www.kinopoisk.ru/film/13943/
13 https://www.kinopoisk.ru/film/414661/
14 https://www.kinopoisk.ru/film/7656/
15 https://www.kinopoisk.ru/film/15579/
16 https://www.kinopoisk.ru/film/464933/
17 https://www.kinopoisk.ru/film/1309187/
18 https://www.kinopoisk.ru/film/297001/
19 https://www.kinopoisk.ru/film/35763/
20 https://www.kinopoisk.ru/film/20394/
```

- Далее, нужно настроить параметры в файле
film_scraper\film_scraper\settings.py:

```
sitemap_film_review1.txt settings.py
24
25 # Obey robots.txt rules
26 ROBOTSTXT_OBEY = True
27
28 # Global variables
29 NUM_SITEMAP = 1 # номер карты с ссылками
30
31 PATH_2_WORK_DIR = "D:/Мои документы/Лабы и рефераты/5 курс 1 семестр/Информационный поиск/Сбор документов/film_scraper/"
32 PATH_2_JSON = path.join(PATH_2_WORK_DIR, "film_scraper/jsons/films{:d}.jsonlines".format(NUM_SITEMAP))
33 PATH_2_LINKS = path.join(PATH_2_WORK_DIR, "film_scraper/sitemaps/sitemap_film_review{:d}.txt".format(NUM_SITEMAP))
34 PATH_2_CAPTCHA_LINKS = path.join(PATH_2_WORK_DIR, "film_scraper/captcha/sitemap_captcha_links{:d}.txt".format(NUM_SITEMAP))
35 PATH_2_DRIVER = "D:/Program Files/Chrome/chromedriver.exe"
36
37 IBEGIN = 40000 # Индекс начала обработки
38 IEND = 50000 # Индекс конца обработки
39
40 REVIEWS_OFFSET = 200 # Количество комментариев на одной странице
41
42 MINUTES_SLEEP_PER_HOUR = 5 # На сколько минут паук отключается в час
43
44 DOWNLOAD_DELAY = 1.1 # Задержка при скачивании
45
46 # Feed
47 FEED_EXPORT_ENCODING = 'utf-8'
48 FEED_FORMAT = 'jsonlines'
49
```

- Запустить «паука» командой с указанием пути выходного файла:
\$ scrapy runspider spiders/FSpider.py -o jsons/films1.jsonlines

Структура FSpyder.py:

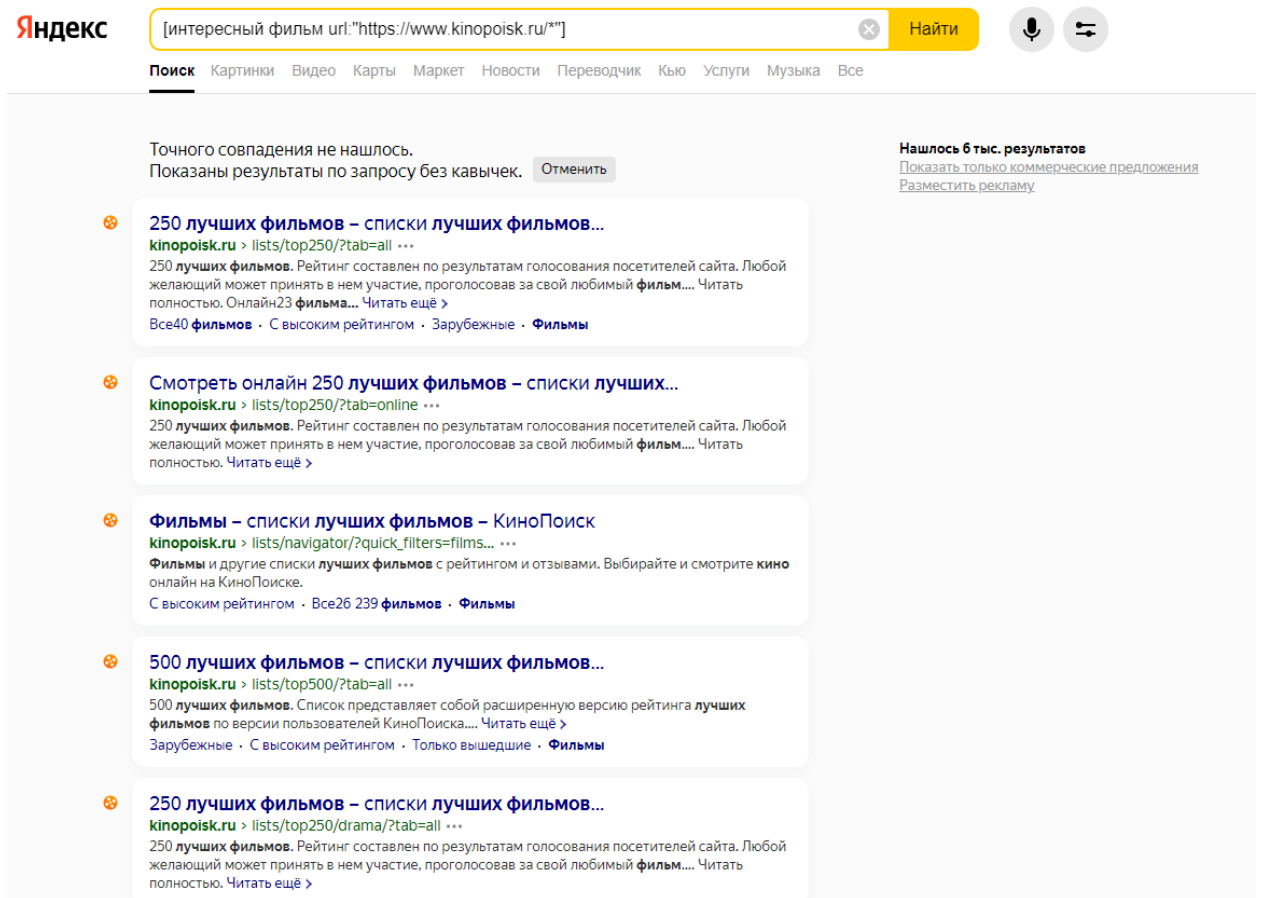
Сигнатура	Описание
<code>def extract_key(url)</code>	Извлечь уникальный номер фильма из ссылки на него
<code>def find_wrap(search)</code>	Безопасный механизм обработки поиска в <i>html</i> -коде (возврат пустой строки, если поиск неудачный)
<code>def fetch_headers(page, item)</code>	Извлечь основную информацию из страницы (всё, кроме отзывов) в объект <i>item</i>
<code>def fetch_reviews(item, driver, url, review_count, logger, captcha_fp)</code>	Извлечь отзывы в объект <i>item</i> . Функция умеет обрабатывать случаи, когда фильм имеет больше отзывов, чем <code>REVIEWS_OFFSET</code> (200 по умолчанию) через обработку нескольких страниц
<code>class FSpider(Spider)</code>	Класс паука
<code>def __init__(self, *args, **kwargs)</code>	-Извлечение ключей всех записей из текущего json-файла в множество (set) <i>film_ids</i> для предотвращения добавление неуникальных записей; -Извлечение списка ссылок на фильмы для обработки из файла и сравнение их с ключами; -Очистка файла с ссылками на страницы с капчей; -Инициализация веб-драйвера
<code>def parse(self, response)</code>	Основная функция обработки страницы. Извлечение <i>item</i> 'а.
<code>def closed(self, reason)</code>	Обработка остановки паука

3. Выводы



На данный момент удалось обработать 60492(фильмы)+60492(рецензии)=120984 страниц с фильмами и рецензиями, суммарным объёмом > 31 ГБ. Количество полученных документов соответственно составляет 60492, объёмом 1.5 ГБ, суммарной длиной в символах 807'172'519. Средний размер документа в символах: 13343. Средний размер документа: 26 Кбайт.

Примеры запросов:

- Интересный = лучший




- А ещё эта фамилия режиссёра

Яндекс Найти  


Поиск Картинки Видео Карты Маркет Новости Переводчик Кью Услуги Музыка Все

Точного совпадения не нашлось.
Показаны результаты по запросу без кавычек. Отменить


Нашлось 6 тыс. результатов
[Разместить рекламу](#)




Учитель (2020, сериал, 1 сезон) – смотреть онлайн...
[kinopoisk.ru > series/1290226/](#) ...
Жанр: драма. Режиссёр: Ханна Фидель, Эндрю Нил. В ролях: Кейт Мара, Ник Робинсон, Эшли Цукерман. Невзрачная школьная учительница из глубинки влюбляется в развитого не по годам ученика, что приводит к катастрофическим последствиям.
★★★★☆ 6,7 из 10 — 12 тыс. оценок · 2020 · США




Учитель (2020, сериал, 2 сезона) – трейлеры, даты...
[kinopoisk.ru > series/1347449/](#) ...
Жанр: триллер, драма, криминал. Режиссёр: Корай Керимоглу. В ролях: Илькер Калели, Джерен Морай, Афра Сарачоглу. В лицей приходит новый **учитель**, который намеревается выяснить детали загадочного самоубийства одной из учениц. Она была успешной...
2 ч 0 мин · 2020 · Турция



Учитель! (2017) – КиноПоиск
[kinopoisk.ru > film/1008271/](#) ...
Жанр: драма, мелодрама. Режиссёр: Такахиро Мики. В ролях: Манами Хига, Судзу Хиросэ, Тома Икута. Косаку Ито преподаёт всемирную историю в старшей школе. Он создаёт впечатление достаточно холодного человека, хотя на самом деле у него...
★★★★☆ 6,7 из 10 — 292 оценки · 1 ч 53 мин · 2017 · Япония



Учитель – КиноПоиск
[kinopoisk.ru > film/922789/](#) ...
Жанр: боевик, драма. Режиссёр: Сюй Хаофэн. В ролях: Ляо Фань, Сун Цзя, Цзян Вэньли. Желая открыть свою школу винчуня, Чань Ши приезжает в Тяньцзинь - город, известный своими боевыми искусствами. Несмотря на то, что Чань Ши...
★★★★☆ 6,2 из 10 — 240 оценок · 1 ч 49 мин · 2015 · Китай



Учитель (сериал, 2 сезона) – трейлеры, даты премьер...
[kinopoisk.ru > series/1118208/](#) ...
Жанр: криминал. Режиссёр: Лукаш Палковский, Мацей Бохняк. В ролях: Мачей Штур, Катажина Дабровска, Петр Гловацкий. **Учитель** приезжает в провинциальный городок и становится участником расследования жестокого убийства местной ученицы. Постепенно он...
★★★★☆ 7,1 из 10 — 156 оценок · 55 мин · 2016 · Польша

- Герой сериала

Яндекс [нацуки субару url:"https://www.kinopoisk.ru/film/*"] Найти

Поиск Картинки Видео Карты Маркет Новости Переводчик Кью Услуги Музыка Все

Точного совпадения не нашлось.
Показаны результаты по запросу без кавычек. Отменить

Нашлось 5 тыс. результатов
[Разместить рекламу](#)

Нацуки Субару, самопровозглашенный рыцарь
kinopoisk.ru > film/971114/episodes/ ...
Перезапуск **Нацуки Субару**. Natsuki Subaru's Restart. 16 мая 2016. Эпизод 8. ... Self Proclaimed Knight, **Natsuki Subaru**. 26 июня 2016. Эпизод 14. [Читать ещё >](#)

Персонаж (2021) – трейлеры, даты премьер – КиноПоиск
kinopoisk.ru > film/4373329/ ...
Жанр: криминал, триллер. Режиссёр: Акира Нагаи. В ролях: Масаки Судэ, Фукасэ, Мицуки Такахата. Начинаящий мангака Кэйго Ямасиро очень неплохо рисует, но так как сам является довольно милым, никак не может создать правдоподобного...
[Трейлеры](#) · [Сборы](#) · [5 актёров](#) · [Слова](#)
2 ч 5 мин · 2021 · Япония

После. Глава 3 (2021) – КиноПоиск
kinopoisk.ru > film/1406473/ ...
Жанр: драма. Режиссёр: Кастиль Лэндон. В ролях: Джозефина Лэнгфорд, Хиро Файнс-Тиффин и др.. Встреча с притягательным бунтарем Хардином разделила жизнь Тессы на «до» и «после». Их судьбы кажутся неразрывно связанными, но Тесса...
[В главных ролях](#) · [Связи](#) · [Премьеры](#) · [Трейлеры](#) · [Слова](#)
★ ★ ★ ★ ☆ 5,1 из 10 — 7 тыс. оценок · 1 ч 39 мин · 2021 · США, Болгария

Истребитель демонов: Поезд «Бесконечный»...
kinopoisk.ru > film/1347949/ ...
Жанр: аниме, мультфильм, боевик. Режиссёр: Харуо Сотодзаки. В ролях: **Нацуки** Ханаэ, Ёсицугу Мацуока, Сатоси Хино. Завершив оздоровительные тренировки в Доме бабочки, Тандзиро и его друзья отправляются выполнять новое задание. На поезде...
[Премьеры](#) · [Сборы](#) · [Связи](#) · [В главных ролях](#) · [Трейлеры](#) · [Студии](#)
★ ★ ★ ★ ☆ 7,8 из 10 — 21 тыс. оценок · 1 ч 57 мин · 2020 · Япония

Мемуары Ванитаса (2021, сериал, 2 сезона) — актёры...
kinopoisk.ru > film/4478547/cast/ ...
Мемуары Ванитаса (2021). Актёры, режиссер, продюсеры и другие участники съемочной группы.

В ходе выполнения лабораторной работы я научился анализировать и скачивать *html*-страницы в автоматическом режиме, с целью извлечения нужных сведений. Это может пригодиться для сбора статистики по сайту, создание агрегаторов, таких как, например, агрегатора цен на авиабилеты, поиска уязвимостей, получение корпуса документов для машинного обучения и прочее.

Список литературы

- [1] Райан Митчелл *Современный скрапинг веб-сайтов с помощью Python*.
2-е межд. издание. — СПб.: Питер, 2021 — 528 с. (ISBN 978-5-4461-5)