Для изучения мной был выбран сайт:

https://www.chitai-gorod.ru/catalog/books/nauka_i_tekhnika-9170/

Проект состоит из 4-х частей:

- Scrapy_part
- Synonyms
- IndexBooks
- Relev+ML+nDCG

В первом проекте скрапим сайт при помощи команды: «scrapy crawl pycoder -o output.json»

На выходе мы получим json.

Во втором проекте при помощи предобученной модели word2vec_model составляем словарь синонимов по получившемуся в Scrapy_part json-у.

Файл с синонимами можно найти в каталоге Synonyms. Название файла – synonym1.txt.

В проекте IndexBooks мы строим индекс. Ниже приведен скриншот работы checkIndexSearch.java.

```
Выберите действие:
1 - Поиск по автору
2 - Поиск по названию
3 - Поиск по цене
0 - Выход
2
Введите название: орел
Поиск по названию "орел":

Вокруг света на "Коршуне" == Станюкович К.
Жизнь животных. В 10 т. Т.6: Птицы. Л - Я. От ласточки до ястреба == Брем А.
Полетели соколы искать золото Тьмутороканя. "Слово о полку Игореве". Новый перевод и комментарии == Колтунова И.
Птицы. Лебеди, гуси, цапли, выпи, коршуны, зимородки, поганки, кулики, чайки, крачки, пастушки и многие другие в своей естественной среде.
```

На скриншоте можно увидеть, что по запросу «орел» книг нет. Тем не менее в выдаче есть результат, поскольку такие слова как коршун или птица являются синонимами слову «орел».

Для сохранения результата в json и поиска одновременно по трём полям нужно запустить checkChooses.java. Ниже приведён скриншот его работы.

```
Выберите действие:
1 - Поиск по автору
2 - Поиск по названию
3 - Поиск по издательству
4 - Поиск по всем трем полям
0 - Выход
2
Введите название: река
Введите имя файла для сохранения результата поиска: query
Поиск по названию "река":

Река, выходящая из Эдема == Докинз Р.
Москва-река в пространстве и времени == Озерова Н.
Зоопланктон равнинных малых рек == Крылов А.
От Кяхты на истоки Желтой реки == Пржевальский Н.
К реке. Путешествие под поверхностью == Лэнг О.
100 великих рек и озер мира == Ломов В.
Бентос лососевых рек Урала и Тимана == Шубина В.
Ах, что такое движется там по реке… == Гройсман В., Гройсман Я., Кириллова Ж., Абаева Г.
Фитопланктон Нижней Волги. Водохранилища и низовье реки == Трифонова И. (ред.)
О чем знает река: Романовские места на Яузе == Домашнева Н.
Обратные задачи математического моделирования неустановившегося движения воды в реках == Романов А.
Гидронимия бассейна реки Мсты. Свод названий и анализ микросистем == Васильев В.
```

В последнем проекте мы вычисляем метрику nDCG. Сначала мы вычисляем её по данным индекса. Затем вычисляем её при помощи регрессионной модели. Из скриншота ниже можем наблюдать, что результат улучшился по сравнению с результатом индекса.

```
nDGC для каждого запроса отдельно:
 [0.9501228274335413, 0.8960705772104712, 0.9966614662913998, 0.8675958648195072, 0.9958969263892178,
Общий nDCG для оценки результатов поиска с помощью индекса:  0.9475338414720744
Предсказанный результат:  [3.40416836 1.71302321 3.57174341 3.57482313 4.85265727 1.71302321
 4.25613211 1.71302321 4.28814857 3.57482313 1.71302321 3.68440462
 1.71302321 3.57482313 3.60907038 5.51373526 1.71302321 2.41491973
 1.71302321 3.53543023 2.51361364 1.71302321 3.18324494 3.3527469
```

•

•

•

```
 4.02246285 1.71302321 3.80540616 1.71302321 2.41491973 3.54492504
 3.57482313 2.97809426 2.41491973 3.57482313 3.57482313 3.57482313
 1.71302321 3.57482313 3.5702607  4.03084493 1.71302321 3.57482313
 3.16358609 3.57482313 3.57482313 1.71302321 1.71302321 3.52170893
 3.18163872 3.57482313 3.98836216 3.27394498 1.71302321 3.49191272
 2.51361364 3.88885745 4.34005741 1.71302321 1.71302321 3.21746009
 3.57482313 3.57482313 2.49014143 3.57482313 1.71302321 3.70835381
 6.31606074 6.00565048 2.64230726 2.41491973 3.35990854 2.41491973
 1.71302321 3.57482313 3.57482313 3.0447749  3.19162538 7.07887206
 2.64230726 2.51361364 3.54086917 2.64230726 1.71302321 1.71302321
 3.29265307 1.71302321 1.71302321 2.90492597 1.71302321 2.49184263
 1.71302321 3.15916186 2.64230726 3.69052066 1.71302321 2.51361364
 1.71302321 2.41491973 2.51361364 3.17378637 3.57482313 2.28994261
 6.29152806 3.34804185 1.71302321 3.16721154]
nDCG после применения ML:  0.9884424935100669
```