ВЕКТОРНЫЙ ПОИСК V0

Предложение по улучшению поиска на Stepik.org

Innotiative.

Ермолаева Л. Я.

Сайфетдиарова А. Р.

Газизов Б. А.

Ершов И. Н.

Попов А. М.

Кратко:

Мы, участники студенческой команды Innotiative, в наши школьные годы активно обучались на платформе Stepik.org и переодически испытывали трудности с поиском нужного курса. Во время обучения в Университете Иннополис, мы создавали проекты с “умными” поисковыми системами (векторный поиск, Deep Learning) и заметили преимущества применения векторного поиска в поисковой системе на Stepik.

Мы использовали открытое API Stepik и написали векторный поиск с курсами платформы, а также провели небольшое сравнительное исследование.

Результаты исследования: Векторный поиск значительно обходит поиск на платформе по релевантности результатов первой страницы. Преимущество векторного поиска прослеживается как на сложных и специфичных запросах, так и на более общих и кратких. Заметно повышение суммы стоимости релевантных курсов первой страницы и смещение наиболее подходящих курсов, в том числе платных, в первые 5-10 курсов поиска. Ограничения исследования: 10 поисковых запросов, рассматривались результаты только первой страницы.

Протестировать векторный поиск с фронтендом (визуалом) платформы Stepik по ссылке:

https://stepik.skillsnavigator.ru/

Лид студенческой команды **Innotiative:** Лана Ермолаева: tel: 8 (906) 141-16-24, email: ermolanaeva@gmail.com, tg: @oELYAo

Содержание документа:

Что такое векторный поиск?

Векторный поиск - это технология, которая позволяет компьютеру искать по смыслу, а не только по словам. Тексты и другие данные преобразуются в специальные числовые представления (*векторы*), которые отражают смысловое содержание. Система сравнивает насколько близки смыслы между запросом и объектами в базе и выдает результат, распределяя по важности параметров.

Как он работает?

1. Преобразуем текст в числа

Название, сложность и резюме курса преобразуются в набор чисел - смысл текста

```
[0.8, -0.3, 0.6, 0.1, ...]  (768 измерений, значения от -1 до 1)
```

```
Модель трансформера: cointegrated/LaBSE-en-ru
```

2. Храним данные как векторы

Каждый курс превращается в вектор и хранится в базе Qdrant

```
ID: 12345, Vector: [...], Payload: {название, цена, автор}
```

3. Преобразуем запрос в вектор

Запрос пользователя, например ”мобильная разработка” также переводится в вектор

Упрощенная демонстрация:

```
Запрос: “Python для начинающих” [0.1, -0,2, 0.3, ...]
```

```
Курс: “Основы разработки на Python” [0.12, -0.18, 0.32, ...]
```

```
Косинусное сходство ~0,95 (очень схоже)
```

4. Ищем с учетом важности параметров

Система сравнивает этот вектор с векторами курсов и находит ближайшие по указанной важности параметров:

```
Векторное сходство (20), количество учеников (1), цена курса (1)
```

5. Возвращаем результаты

Qdrant возвращает 200 результатов. Система сортирует по указанным параметрам и делит на 20 курсов для каждой страницы. Результаты готовы для отображения на сайте

Сравнительный анализ и его результаты

Методы: Мы проверили 10 тестовых запросов, отражающих основные формы и темы популярные на платформе. Результаты первой страницы каждого запроса были автоматически перенесены в таблицы, вручную проверены на релевантность и поделены на релевантные (без заливки) и нерелевантные (красная заливка). Мы посчитали количество релевантных курсов и общую сумму стоимости релевантных курсов первой страницы. Для наглядности выделили лучшие показатели зеленым цветом. Выжимка из таблиц далее в документе.

PDF и XLSX файла с таблицами вы можете найти на Яндекс Диске:

https://disk.yandex.ru/d/odhG8ZOc5UnglA

Выжимка результатов:

Запрос	Релевантность 1й страницы, курсы		Стоимость релевантных 1й страницы, ₽	
	Stepik.org	Векторный	Stepik.org	Векторный
Создание чат-ботов для бизнеса на Python	3	15	11390	22680
Разработка на Kotlin	12	20	62312	42160
react практические проекты	3	12	13000	28779
разработка игр на unity	17	20	202360	228269
Основы SQL для аналитиков	4	20	10220	26388
Технический английский	2	16	599	25447
Анатомия человека	5	7	0	0
SEO продвижение	7	12	10690	5449
UX UI дизайн	17	18	42190	43220
обучение тестированию мобильных приложений	4	12	11379	12238

Результаты: Во всех тестовых кейсах заметно улучшение релевантности выдачи. Векторный поиск показал значительное преимущество в релевантности запросов с большим количеством слов, такими как “Создание чат-ботов для бизнеса на Python”, “обучение тестированию мобильных приложений”, “Основы SQL для аналитиков”, а также с точными запросами: “react практические проекты”, “Технический английский”. Улучшение релевантности выдачи позитивно влияет на общую стоимость релевантных курсов. В случаях “Разработка на Kotlin” и “SEO продвижение” поиск платформы Stepik показал более общие и дорогие курсы, что сказалось на повышении суммы релевантных курсов. В этих случаях векторный поиск выдал более точные и популярные платные и бесплатные курсы, уступив в общей стоимости релевантных курсов поиску Stepik, но выиграв в релевантности и разнообразии.

Вычислительные ресурсы для векторного поиска:

Векторизованы все 10453 курса

Векторные измерения: 768

Вес векторной базы: 655MB

Модель для векторизации: cointegrated/LaBSE-en-ru

Вес модели для векторизации: 200-300MB

Реализация поиска занимает следующие вычислительные мощности, **если нода одна (если один сервер):**

RAM 2GB

vCPUs 0.5 vCPU

Disk Space 8 GiB

При наличии нескольких нод (серверов) рекомендуем обратиться к калькулятору Qdrant и изменить “Replication Factor” на нужное количество : https://cloud.qdrant.io/calculator?__hstc=265983056.4c72d1a26a750b97fc018534fdd1b08e.1750769441368.1756330886588.1756395467451.9&__hssc=265983056.1.1756395467451&__hsfp=%223283356429%22&qdrant_tech_ajs_anonymous_id=446b87bc-3b5d-4c76-81b5-aad4fa3a5d7d&provider=aws&storageOptimized=true&quantization=None&replicas=1&vectors=10000&dimension=768&sparseVectors=10000&sparseElements=768&storageRAMCachePercentage=35&region=ap-northeast-1

Польза для Stepik как для бизнеса:

Улучшение релевантности поиска поможет:

- Повысить конверсию в покупателей
- Снизить отток
- Сократить время до первого целевого действия
- Повысить удовлетворённость пользователей

Наша версия векторного поиска подбирает релевантные курсы, отдавая приоритет платным и популярным. Она одновременно повышает бизнес-метрики за счёт акцента на монетизируемых курсах и улучшает пользовательский опыт благодаря точным рекомендациям.

Попробуйте векторный поиск:

https://stepik.skillsnavigator.ru/

Каналы связи:

Лид студенческой команды Innotiative: Лана Ермолаева

тел.	email	telegram
8 (906) 141-16-24	ermolanaeva@gmail.com	@oELYAo