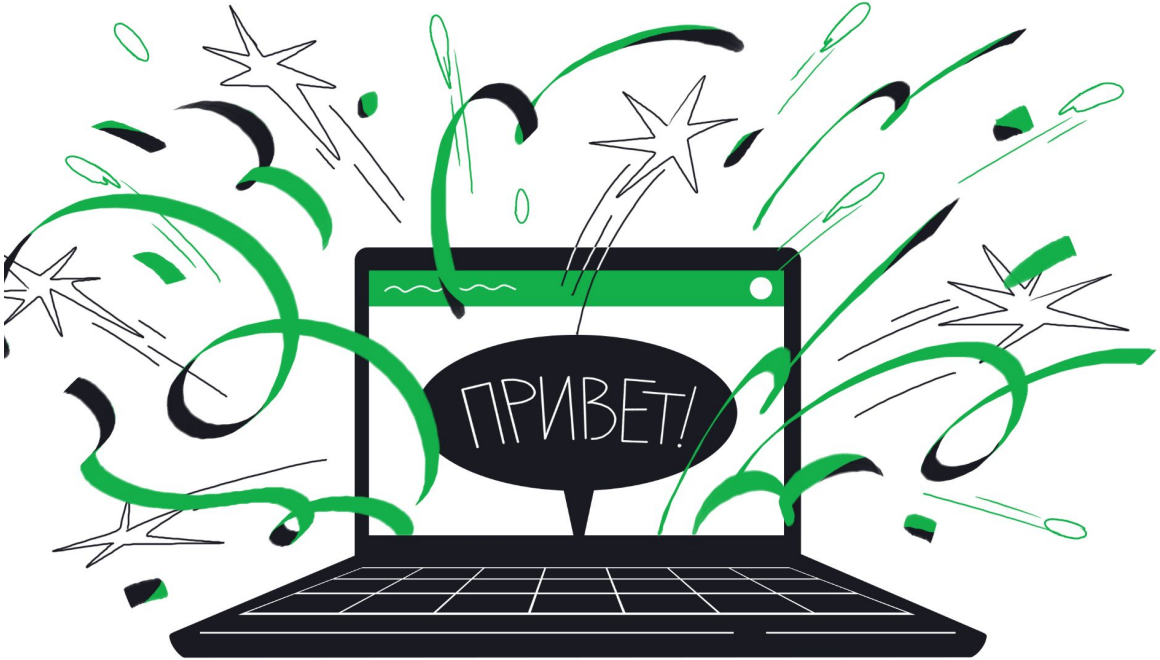
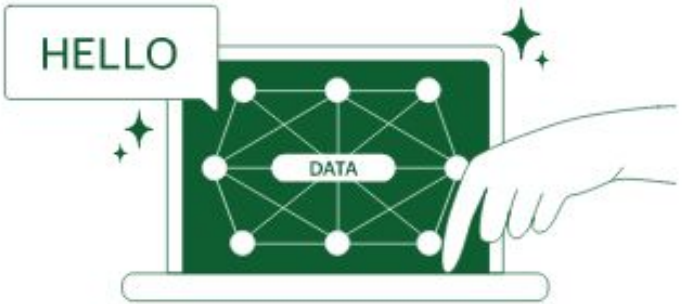
Мастерская №2

Вводный вебинар. Уточнение задачи



План встречи

- Знакомство с командой
- Что такое Мастерская, этапы, сроки
- Уточнение задачи
- Обсуждение плана работ
- Q&A



Эдуард Григорян
Team-lead проекта, эксперт и наставник
Яндекс Практикума

- Помогает с техническими вопросами, включая поиск оптимального решения и ошибок в коде;
- Проводит онлайн-встречи;
- На связи с 10 до 16 по МСК (пн-пт).

Тимлид проекта

Даша Голева
Проджект менеджер Мастерской Яндекс Практикума

- Помогает с организационными вопросами (дедлайны, приемом/отправкой проектов и т.д.;
- На связи с 10 до 19 по МСК, пн-пт (отвечает в течение часа).

Сопровождающий ПМ

Что такое Мастерская?

Это агентство внутри Яндекс Практикума, где студенты улучшают свои навыки и создают кейсы для портфолио, работая с реальными данными и задачами. Такие проекты высоко оцениваются работодателями, так как доказывают ваш интерес к индустрии и профессии, демонстрируют умение применять полученные знания на практике.

В рамках опроса потенциальных работодателей выяснили, что кандидатов с внеучебными проектами в портфолио считают более заинтересованными в профессии и им чаще отдают предпочтение при трудоустройстве.

В Мастерской вы прокачаете:

- Навыки решения реальных задач
- Формулирование и уточнение постановки задачи
- Умение работать в условиях неопределенности
- Навыки поиска и апробации различных подходов и методов решения задачи
- Командное взаимодействие в процессе работы над проектом
- Тайминг и планирование работ

Мастерская №2

Важно

- Мастерская это не обучение - это практика.
- Для проекта мы возьмем реальный датасет из открытого источника.
- Будьте готовы поработать с “сырыми” данными и уделить время их предобработке.
- В этой Мастерской не будет реального клиента, но это никак не отразится на ценности результата для вас и потенциального рекрутера. Наша цель - показать Hard skills.

Мастерская №2

Важно

Особенность IT-индустрии состоит в непрерывном развитии: обучении, а также поиске новых инструментов для каждого клиента и для каждой задачи. По этой причине компании рассматривают самостоятельных и проактивных кандидатов. Тех, кто готов развиваться и развивать продукт, не останавливаясь на имеющемся спектре знаний.

Для прокачки этих качеств в Мастерскую встроены моменты, для которых нужно искать какую-то часть информации и пути решения, работать с новыми инструментами - проявить свою самостоятельность.

Однако, если у Вас что-то не получается, мы обязательно подхватим и поможем.

Мастерская №2

Важно

- Ищите новые подходы, изучайте новые методы и инструменты
- Не стесняйтесь обращаться за помощью, спрашивать и помогать друг другу, если знаете ответ
- Погрузитесь в задачу, начните и поймите свои пропуски в знаниях для дальнейшего их улучшения
- Не переживайте за конечный результат и его качество
- Помните: единственно правильного решения в проекте нет

Исходные данные. Уточнение задачи

что у нас есть на входе, и что ждут от нас на выходе

Мастерская №2

Задача мэтчинга (соответствия)

Дано:

Два множества объектов: A и B. Каждый объект в множестве описывается какими-то признаками.

Желаемый результат:

Для каждого объекта из множества A найти один или несколько объектов из B, которые близки к нему по некоторой заданной метрике.

*A и B могут быть одним и тем же множеством

**Можем и не найти ни одного соответствия.

Мастерская №2

Задача мэтчинга (соответствия)

В каких задачах применим мэтчинг:

1. Текстовый поиск: A - запросы, B документы в сети Интернет
2. Поиск по фотографиям
3. Поиск похожих товаров
4. “С этим товаром часто покупают…”
5. “Похожие товары”

Мастерская №2

Задача мэтчинга (соответствия)

Подготовили для вас описание мэтчинга и некоторые инструкции применимо к нашей задаче:
https://mushenokf.notion.site/60b8fca216134f10893ee15c2a7c78ca?pvs=4

Обратите внимание, что данные достаточно объемные (более 2 Гб). Также существует уменьшенная (~10% от исходной) версия датасета. Можете попробовать свои силы на ней, а затем переходить к полной версии (но не обязательно, работа на уменьшенной версии также будет зачтена).

Полная версия датасета: https://disk.yandex.ru/d/BBEphK0EHSJ5Jw

Уменьшенная версия датасета: https://disk.yandex.ru/d/YQEIc_cNQQLSOw

Мастерская №2

Исходные данные

base.csv - анонимизированный набор товаров. Каждый товар представлен как уникальный id (0-base, 1-base, 2-base) и вектор признаков размерностью 72.

train.csv - обучающий датасет. Каждая строчка - один товар, для которого известен уникальный id (0-query, 1-query, …) , вектор признаков И id товара из base.csv, который максимально похож на него (по мнению экспертов).

validation.csv - датасет с товарами (уникальный id и вектор признаков), для которых надо найти наиболее близкие товары из base.csv

validation_answer.csv - правильные ответы к предыдущему файлу.

Мастерская №2

Исходные данные

Задача:

- разработать алгоритм, который для всех товаров из validation.csv предложит несколько вариантов наиболее похожих товаров из base;
- оценить качество алгоритма по метрике accuracy@5

Формула:

- **Метрика.**
  - Необходимо максимизировать метрику `accuracy@5`, которая для каждого объекта вычисляется по формуле:

$$accuracy@5 = 100 * \frac{\text{кол-во верно определённых похожих объектов из 5 возможных}}{5}$$

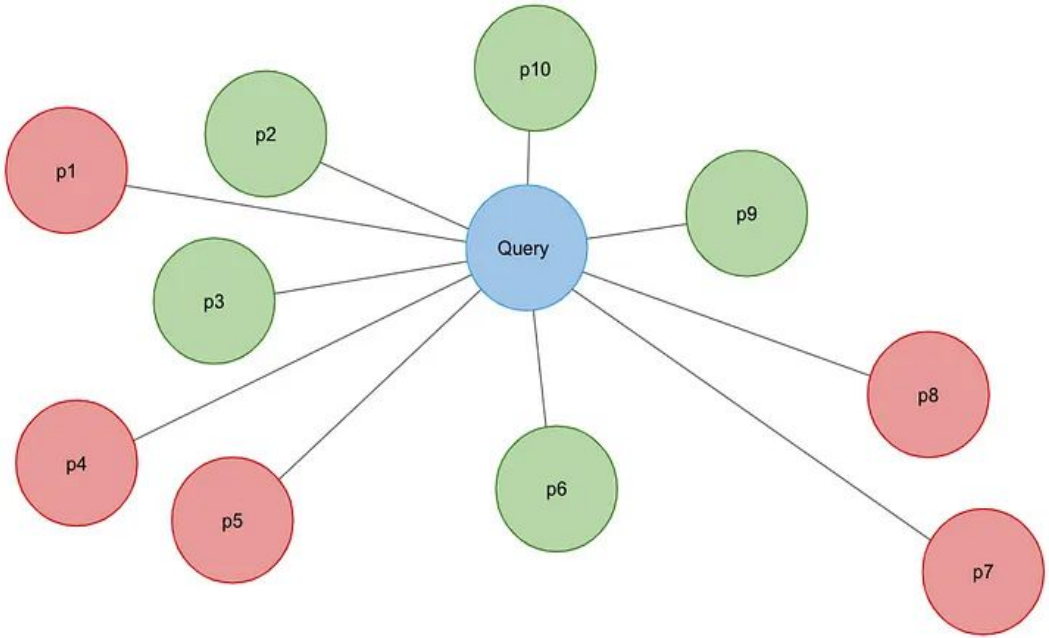
Мастерская №2

Итоги проекта

- репозиторий на GitHub
- README
- Jupyter Notebook с решением
- Google Colab

Мастерская №2

Приближенный поиск ближайших соседей

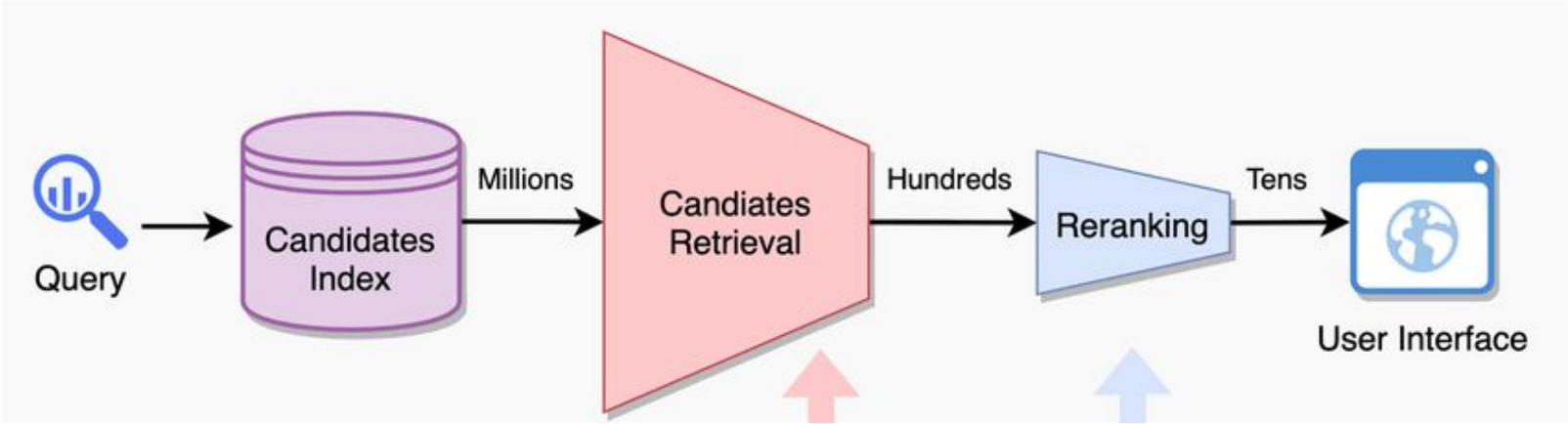


Библиотеки:

- FAISS
- Annoy
- Qdrant

Мастерская №2

Двухстадийный поиск



Мастерская №2

Итоги проекта

Deep Learning Stories

Векторные базы данных и стартап с Андреем Васнецовым

Мастерская №2

Полезные ссылки

- https://habr.com/ru/companies/vk/articles/338360/
- https://scikit-learn.org/stable/modules/neighbors.html#unsupervised-neighbors

FAISS:

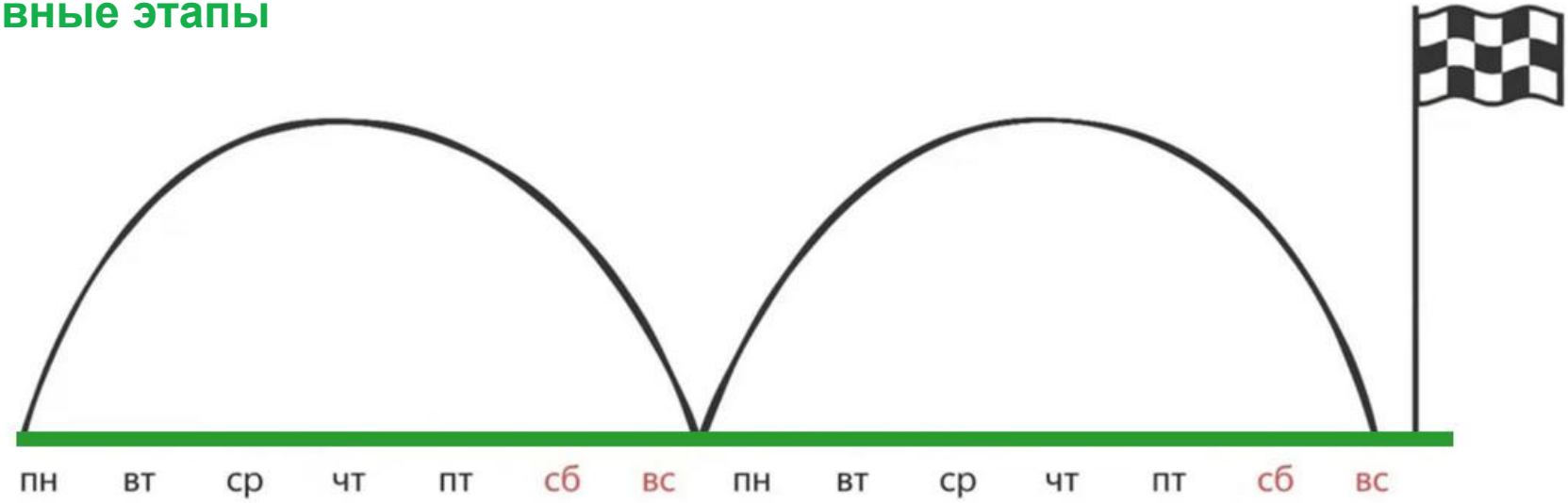
- https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/
- https://habr.com/ru/companies/okkamgroup/articles/509204/
- https://evogeek.ru/articles/298310/
- https://www.pinecone.io/learn/series/faiss/faiss-tutorial/
- https://towardsdatascience.com/understanding-faiss-619bb6db2d1a
- https://towardsdatascience.com/getting-started-with-faiss-93e19e887a0c

Annoy:

- https://erikbern.com/2015/09/24/nearest-neighbor-methods-vector-models-part-1
- https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html
- https://erikbern.com/2016/06/02/approximate-nearest-news.html
- https://github.com/spotify/annoy

Мастерская №1

Основные этапы



Вводный вебинар, постановка задачи

Самостоятельная работа, проведение Q&A

Демонстрация результатов проекта

Мастерская №1

График работы над проектом

- **22 июля 19.00** – Вводный вебинар;
- **26 июля 18.00** – Вебинар K Neighbors + Faiss + Q&A сессия;
- **29 июля 19.00** – Q&A сессия + вебинар + Faiss
- **5 августа — дедлайн по сдаче работ**
- * - финальная встреча, презентация лучших решений.

*** Дату финальной встречи определим позже**

Мастерская №1

Next steps…

1. Загрузить данные
2. Прочитать дополнительные материалы
3. Понять задачу
4. Повторить baseline
5. Провести EDA и повысить качество
6. Исследовать опции FAISS
7. Разработать ранжирующую модель

Мастерская №1

Рекомендации по оформлению

- Описание проекта, название, цели, исходные данные
- Краткие комментарии в коде
- Гипотеза-исследование-выводы
- Чистый код, PEP8
- Меняете данные – должно быть подтверждение ДО и ПОСЛЕ
- По итогу исследования можно подготовить короткий, но содержательный отчет.

Мастерская №1

Ревью работ

- Созданный проект необходимо разместить в **GitHub** (либо другом репозитории/диске/ Google Colab)
- Ссылку на работу вносим в специальную форму
- Проверенный проект мы пришлем вам на адрес электронной почты, указанный в форме
- По умолчанию предусмотрена 1 проверка проекта, но при наличии критических ошибок в работе, мы отдельно попросим вас прислать проект еще раз
- Проверка работ в Мастерской может занимать до 7 дней, но мы стараемся все сделать максимально быстро

Портфолио

Важным итогом Мастерской является созданный и качественно оформленный проект в вашем портфолио!

Необходимые советы и рекомендации вы найдете **здесь**

Ваши вопросы

Всем спасибо!

Не забываем:

- Задавать вопросы тимлиду;
- Обсуждать с сокурсниками сложности;
- Придерживаться плана и сроков.