



ML Advanced

Поиск нечетких дублей

• REC

Проверить, идет ли запись

Меня хорошо видно
&& слышно?



Ставим “+”, если все хорошо “-”,
если есть проблемы

Тема вебинара

ML Advanced

Поиск нечетких дублей

Игорь Стурейко



Teamlead, главный инженер проекта – НИИгазэкономика

Опыт:

Более 15 лет занимался прикладной математикой и мат моделированием (Data Scientist) (Python, C++) в НИИ ПАО Газпром

Анализ временных рядов, эволюционное развитие сложных систем

+7 (916) 156-07-82 (whatsapp)

@stureiko (TG)

Правила вебинара



Активно
участвуем



Off-topic обсуждаем
в учебной группе



Задаем вопрос
в чат или голосом



Вопросы вижу в чате,
могу ответить не сразу

Условные обозначения



Индивидуально



Время, необходимое
на активность



Пишем в чат



Говорим голосом

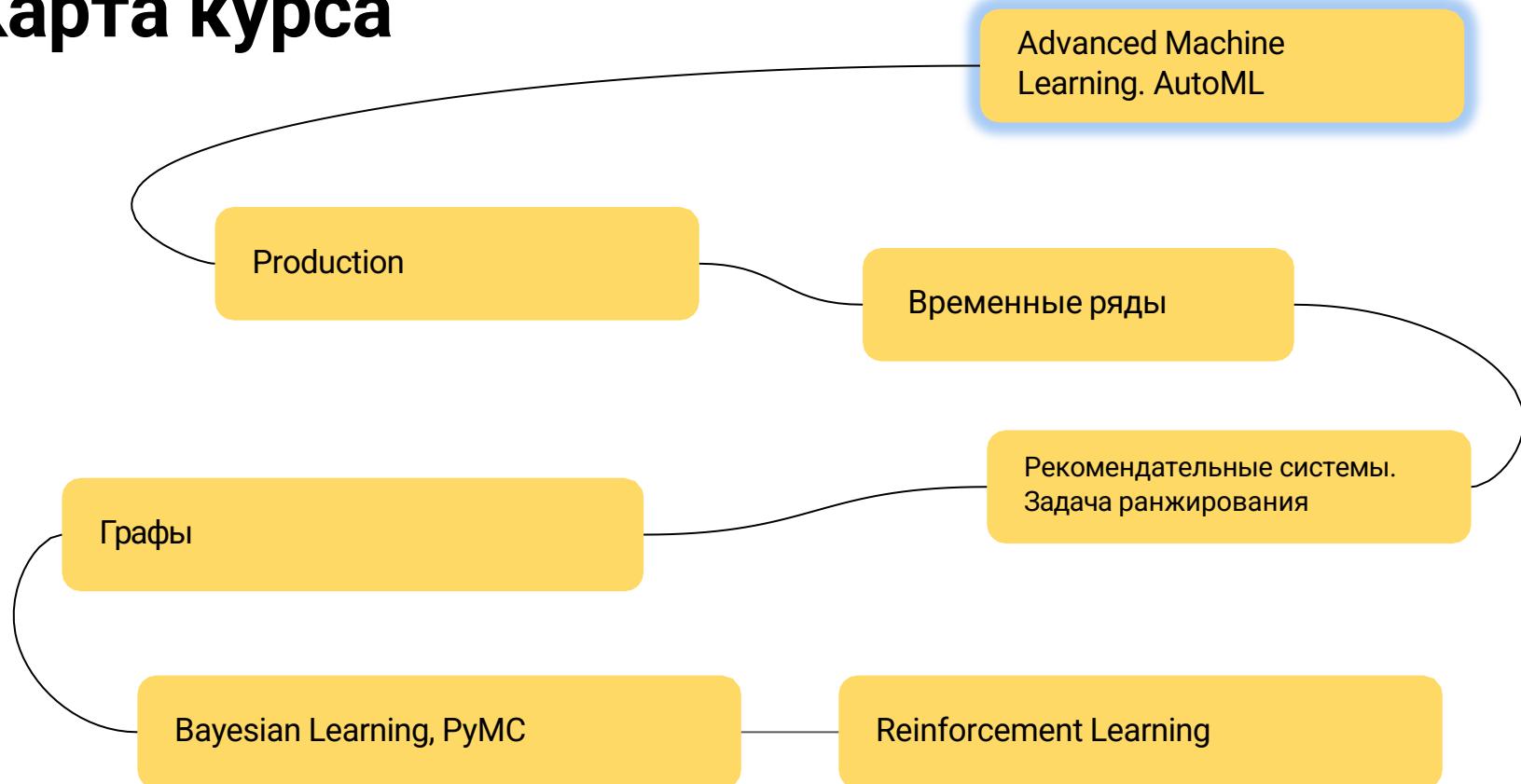


Документ



Ответьте себе или
задайте вопрос

Карта курса



Цели вебинара

К концу занятия вы сможете

1. Рассмотреть методы быстрого, но нечеткого поиска ближайших соседей
2. Применить рассмотренные методы для выделения нечетких дублей из набора объектов

Смысл

Зачем вам это знать

1. Алгоритмы приближенного поиска
 2. Задачи нежесткого матчинга
-

Маршрут вебинара

Постановка задачи

Метрики и методы

Диаграмма Вороного

Faiss

SCANN

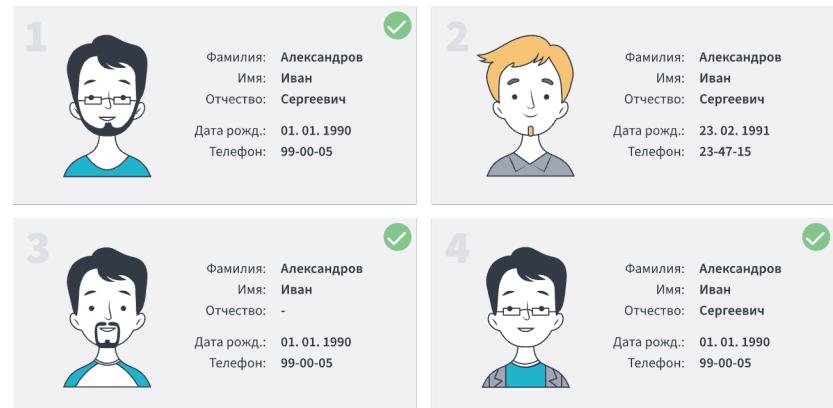


Нечеткие Дубликаты



Задачи поиска нечетких дублей

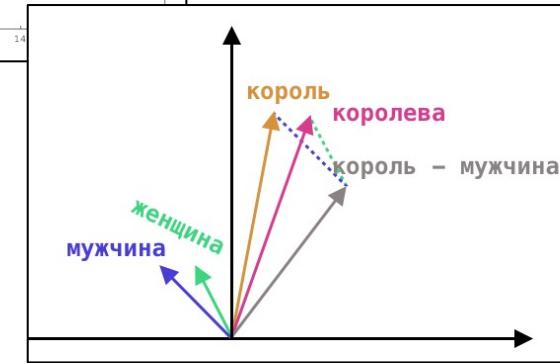
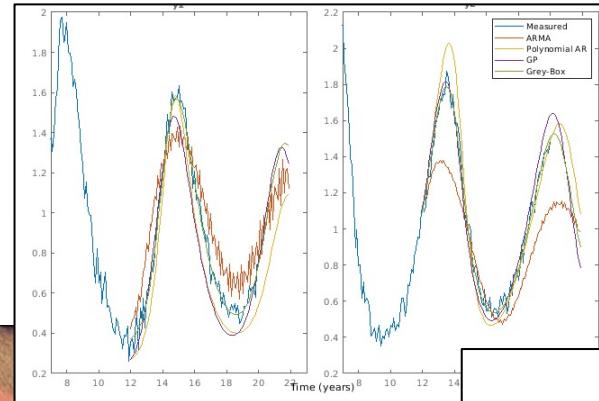
1. Shazam - поиск музыкального трека по зашумленному фрагменту.
2. Поиск дублирующихся постов “перезаливок” в UGC сервисах.
3. Детекция нарушения авторских прав.
4. Удаление дубликатов из датасета, чтобы поднять качество и скорость тренировки.
5. Распознавание лиц.
6. Наука: написание статей о том, что вы нашли дубликаты в популярном датасете.



Сходство

Виды сходства:

- текстовое,
- числовое,
- визуальное,
- семантическое



Сходство - это мера степени схожести, близости или схожести между двумя объектами, явлениями или явлениями. Сходство есть числовая мера, которая определяет, насколько два объекта близки друг к другу на основе определенной **метрики**.

Метрики сходства

Евклидово расстояние (Euclidean Distance)

геометрическое расстояние между двумя точками в n-мерном пространстве.

Манхэттенское расстояние (Manhattan Distance)

сумма абсолютных различий между координатами двух точек

Косинусное сходство (Cosine Similarity)

угол между векторами в n-мерном пространстве

Корреляция Пирсона (Pearson Correlation)

линейная зависимость между переменными

Жаккардово сходство (Jaccard Similarity)

мера пересечения множеств.

Расстояние Левенштейна (Levenshtein Distance)

различия между строками путем подсчета минимального количества операций

Расстояние Хэмминга (Hamming Distance)

различия между строками фиксированной длины путем подсчета несовпадающих символов.

Расстояние Махalanобиса (Mahalanobis Distance)

ковариация между признаками

Метрика Минковского (Minkowski Metric)

обобщенное расстояние

Корреляция касательной (Tanimoto Coefficient)

сравнение между двумя бинарными векторами.

Среднеквадратичное отклонение (MSE)

разницу между пикселями двух изображений.

Структурный сходство (Structural Similarity Index, SSIM)

сходство в структуре, контрасте и яркости между двумя изображениями.

Гистограммы цветов (Color Histograms)

гистограммы цветовых каналов.

Гистограммы текстур (Texture Histograms)

текстурные характеристики изображения.

Фреймовое сходство (Frame Difference)

видеообработка для измерения сходства между кадрами видео.

Wavelet-преобразование (Wavelet Transform)

высокочастотные и низкочастотные составляющие изображений.

Сравнение дескрипторов (Feature Descriptors)

такие как SIFT, SURF, и ORB, для извлечения и сравнения ключевых точек и их описаний.

Методы сиамских сетей

Оценивают векторное представление объектов

Структурные метрики (Structural Metrics)

структурные характеристики изображений, такие как сетка дескрипторов и глубокие связи.

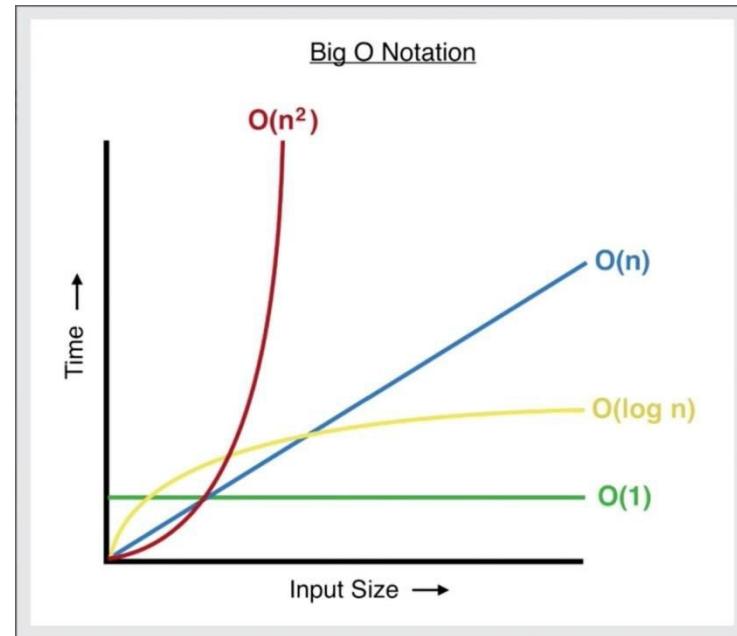


Наивное решение

Наивный алгоритм поиска нечетких дубликатов:

1. Для каждой пары x_1, x_2
2. Вычисляем сходство: $d(x_1, x_2)$
3. Проверяем: $d(x_1, x_2) < t$

Сложность задачи $O(n^2)$



Approximate Nearest Neighbor

Принцип: найдем K наиболее вероятных ближайших соседей.

Сильный выигрыш по скорости за счет потери в Recall.

ANN алгоритм поиска нечетких дубликатов

1. Для каждого элемента x_1
2. Находим K ближайших соседей $[x_2, x_3, x_4, x_5, \dots, x_K]$, вычисляем $d(x_1, x_i)$
3. Проверяем: $d(x_1, x_i) < t$

Сложность на построение индекса: $O(N)$

Сложность на поиск одного дубликата: $O(n \ll N)$

Новая проблема:

- False Positive: элемент не дубликат, распознан как дубликат
- False Negative: элемент дубликат, не попал в K ближайших соседей

Методы поиска нечетких дублей

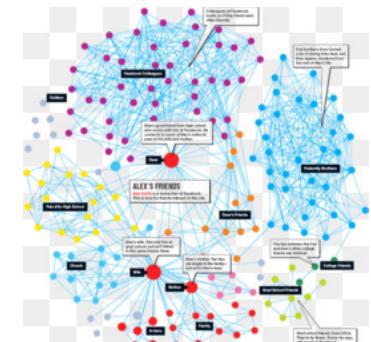
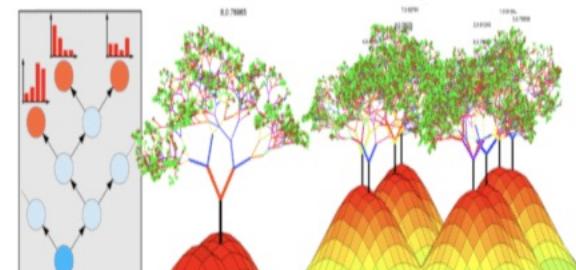
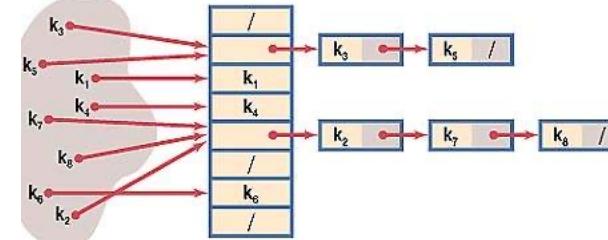
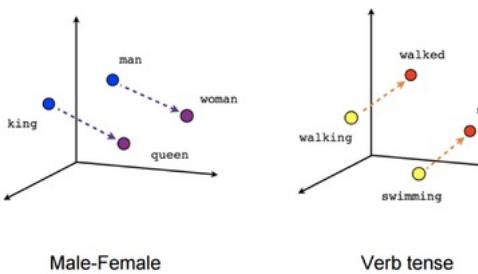
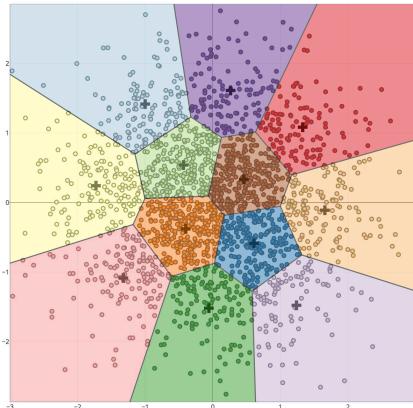
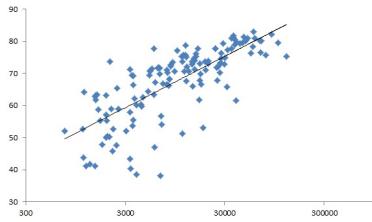
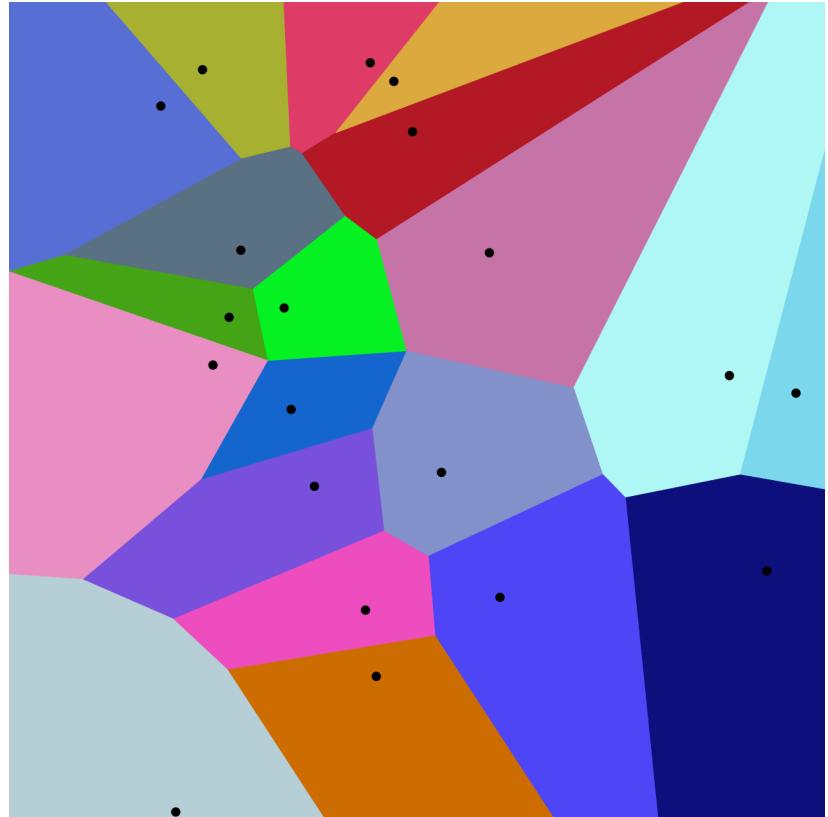


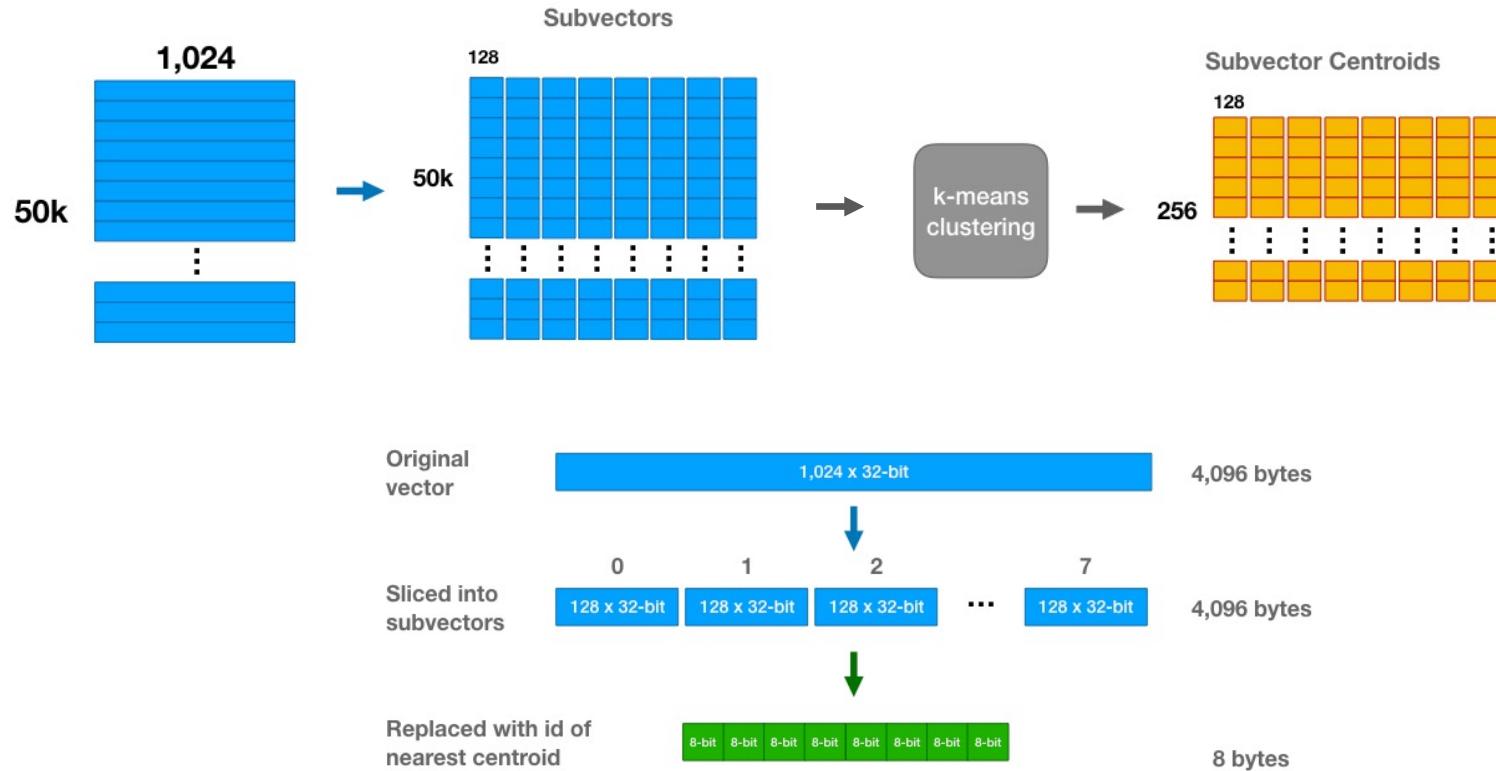
Диаграмма Вороного

1. Алгоритм построения диаграммы Вороного «в лоб». Сложность: $O(n^4)$;
2. Алгоритм построения диаграммы Вороного путём пересечения полуплоскостей. Сложность: $O(n^2 \log(n))$;
3. Алгоритм Форчуна построения диаграммы Вороного на плоскости. Сложность: $O(n \log(n))$;
4. Рекурсивный алгоритм построения диаграммы Вороного. Сложность: $O(n \log(n))$.

[Описание реализации алгоритмов](#)



Квантилизация



Вопросы?



Ставим “+”,
если вопросы есть



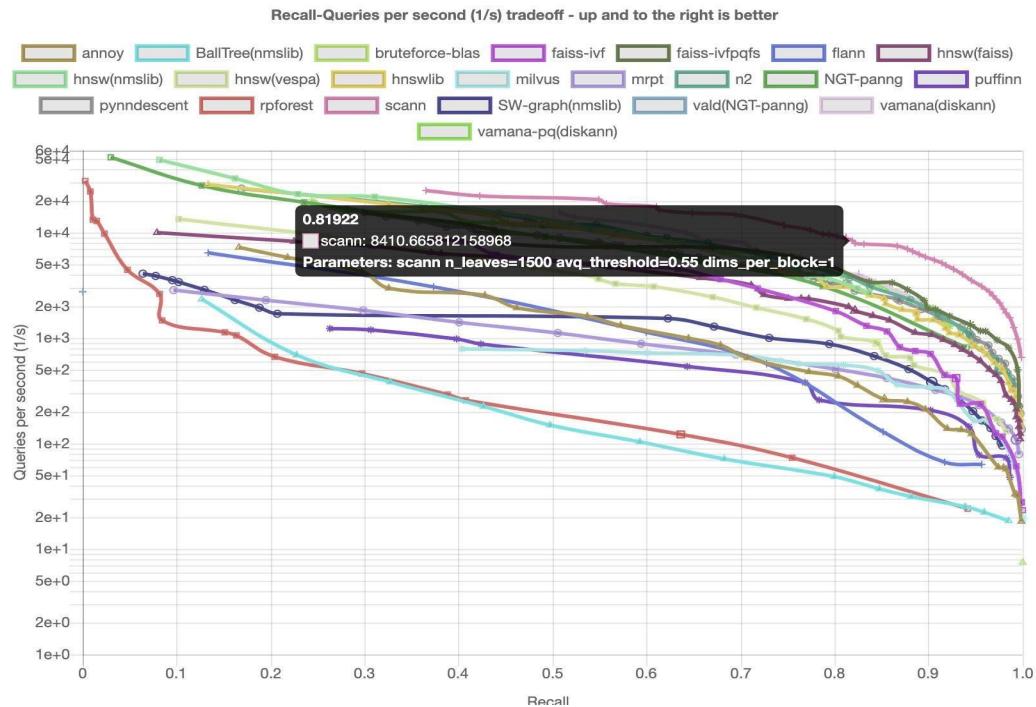
Ставим “-”,
если вопросов нет

Алгоритмы

Посмотреть кто сейчас самый самый

- faiss
 - scann
 - annoy

<http://ann-benchmarks.com>



FAISS = Facebook AI Research Similarity Search

Основная идея

- Набор данных «векторизуется»
- Индекс квантилизуется
- Выбирается метрика сходства
- По индексу происходит поиск k соседей

Поскольку индекс квантилизирован то будут найдены «приблизительные» соседи

Гиперпараметры

- Выбор метрики схожести для индекса
- Метод построения «приближенного» индекса
- Метод квантилизации

Примеры применения faiss:

<https://habr.com/ru/companies/okkamgroup/articles/509204/>

<https://habr.com/ru/companies/avito/articles/488658/>

Квантилизация индекса

<https://mccormickml.com/2017/10/13/product-quantizer-tutorial-part-1/#exhaustive-search-with-approximate-distances>

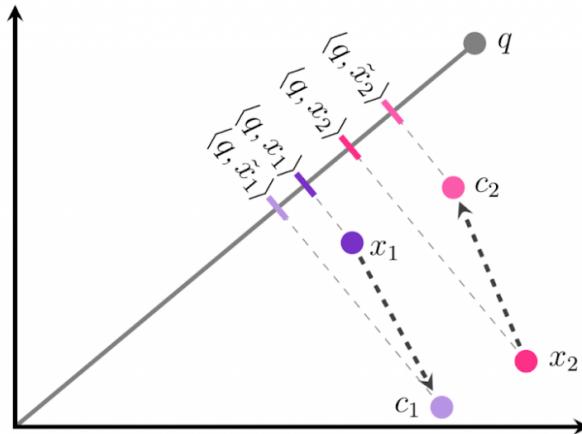
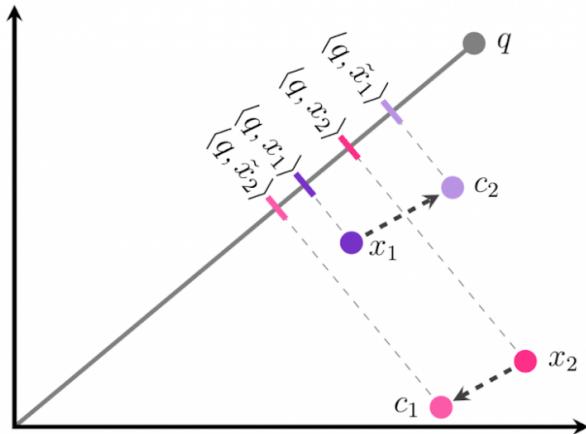
Гайд по выбору индекса

<https://github.com/facebookresearch/faiss/wiki/Guidelines-to-choose-an-index>



SCANN

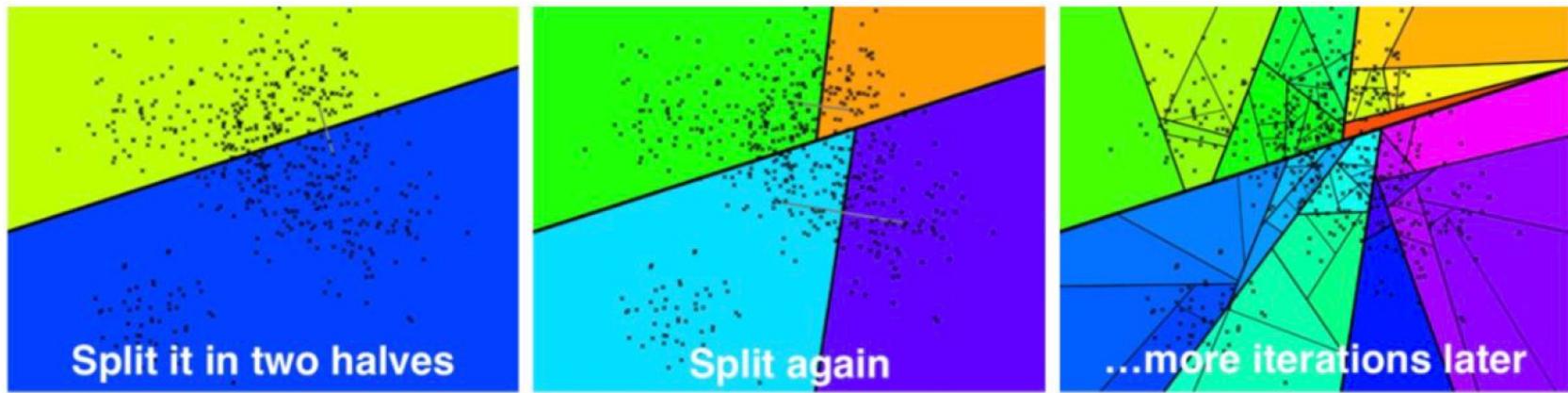
Анизотропная квантилизация



- Количество создаваемых листьев является квадратным корнем от количества данных
- Размер выборки обучения всегда должен быть больше, чем количество созданных листьев/кластеров

Annoy

Чтобы вычислить ближайших соседей, annoy разделяет набор точек пополам и делает это рекурсивно, пока каждый набор не будет иметь k элементов. Обычно k около 100.



Annoy может использовать статические файлы в качестве индексов. В частности, это означает, что вы можете передавать индекс между процессами.

Annoy также отделяет создание индексов от их загрузки, поэтому вы можете передавать индексы в виде файлов.

Вопросы?



Ставим “+”,
если вопросы есть



Ставим “-”,
если вопросов нет

Практика

Список материалов для изучения

1. Диаграмма Вороного

- <https://newtechaudit.ru/postroenie-diagrammy-voronogo-v-python-dlya-zadach-vizualizacii/>

2. ANN Algorithms benchmarks

- <https://medium.com/gsi-technology/how-to-benchmark-ann-algorithms-a9f1cef6be08>
- <https://medium.com/gsi-technology/ann-benchmarks-a-data-scientists-journey-to-billion-scale-performance-db191f043a27>

3. 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://github.com/spotify/annoy>
- <https://sds-aau.github.io/M3Port19/portfolio/ann/>

4. ScaNN

- https://scikit-learn.org/stable/auto_examples/neighbors/approximate_nearest_neighbors.html
- <https://medium.com/@kumon/similarity-search-scann-and-4-bit-pq-ab98766b32bd>
- <https://medium.com/@DataPlayer/scalable-approximate-nearest-neighbour-search-using-googles-scann-and-facebook-s-faiss-3e84df25ba>
- <https://medium.com/analytics-vidhya/scann-faster-vector-similarity-search-69af769ad474>

5. Faiss

- <https://habr.com/ru/companies/okkamgroup/articles/509204/>
- <https://habr.com/ru/companies/avito/articles/488658/>
- <https://mccormickml.com/2017/10/13/product-quantizer-tutorial-part-1/#exhaustive-search-with-approximate-distances>
- <https://github.com/facebookresearch/faiss/wiki/Guidelines-to-choose-an-index>



Вопросы?



Ставим “+”,
если вопросы есть



Ставим “-”,
если вопросов нет

Рефлексия

Рефлексия



С какими впечатлениями уходите с вебинара?



Как будете применять на практике то,
что узнали на вебинаре?

**Заполните, пожалуйста,
опрос о занятии
по ссылке в чате**

Спасибо за внимание!

Приходите на следующие вебинары

21.11 – Построение end-to-end пайплайнов и сериализация моделей

Игорь Стурейко



Teamlead, главный инженер проекта – НИИгазэкономика

Опыт:

Более 15 лет занимался прикладной математикой и мат моделированием (Data Scientist) (Python, C++) в НИИ ПАО Газпром

Анализ временных рядов, эволюционное развитие сложных систем

+7 (916) 156-07-82 (whatsapp)

@stureiko (TG)

