



ПОИСК

АКАДЕМИЯ АНАЛИТИКОВ АВИТО  
ОДИН В ПОЛЕ / НИКИТА БАРАМИЯ

# Задача

- Дано: размеченный датасет для обучения:
  - `query_id`, `query_text` – идентификатор и текст запроса;
  - `query_category_id` , `query_microcat_id` – id категорий и подкатегорий;
  - `query_location_id` – местоположение пользователя;
  - `item_id`, `title`, `description`, `keywords` – характеристики объявления
  - `target` – целевая переменная, где 1 — айтем релевантен запросу, 0 — не релевантен
- Надо: научиться определять релевантность выдачи запросу

# Решения и метрики

| Метод                           | ROC-AUC      | MAP@10       | MAP@50       | NDCG@10      | NDCG@50      |
|---------------------------------|--------------|--------------|--------------|--------------|--------------|
| TF-IDF + cosine similarity      | 0.726        | -            | -            | -            | -            |
| Word2Vec with optuna params     | -            | 0.048        | 0.056        | 0.091        | 0.097        |
| rubert-tiny2 as-is              | -            | 0.013        | 0.016        | 0.032        | 0.035        |
| LaBSE-en-ru as-is (huge model)  | -            | 0.071        | 0.089        | 0.132        | 0.158        |
| rubert-tiny2 tuned + projection | <u>0.805</u> | <u>0.092</u> | <u>0.112</u> | <u>0.161</u> | <u>0.176</u> |

Во всех решениях кроме первого использовался Qdrant в качестве поискового движка на эмбедингах

# Структура финального решения

- Используются только `query_text` и `title`
- Препроцессинг:
  - `lambda text: ' '.join(text.lower().split())`
  - `('[I]' + text)` if `type_of_text == 'query'` else `('[Q]' + text)`
- Metric learning подход:
  - сблизить эмбединги `query_text` и `title` если релевантны и отдалить если нет:  $(y - \text{cos\_sim})^2$
- Финальный эмбединг: `rubert-tiny2 embeddings (312)` -> `projected embeddings (64)`
- Qdrant similarity search



# Демо решения

- <http://158.160.52.241:8080>

---

# Итог

- Реализован поисковый движок на основе LLM + Qdrant
- Что из критичного не сложилось / не успел сделать:
  - Тестирование нагрузкой
  - Большое внимание традиционным подходам (булев поиск, инвертированный индекс)
- Обработка галлюцинаций
- ...
- Спасибо за внимание! 😊