Определение тематики статьи на habr.ru

Выполнили:
Сергеев Константин
Кондрашов Константин
Рубин Даниил

Постановка задачи

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Статьи на habr.ru содержат:

- заголовок
- текст
- картинки
- теги
- **хабы**
- комментарии

Пример хабов:

1. IT-компании
2. Физика
3. Ноутбуки
4. Веб-дизайн

Наша задача заключается в реализации нейронной сети, которая будет автоматически предлагать хабы на основе написанной автором статьи

Автоматическое проставление хабов позволит:

- улучшить опыт пользователей
- экономить время авторов

Сбор данных

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Задача: собрать данные свежих статей

https://habr.com/ru/articles/{id}/

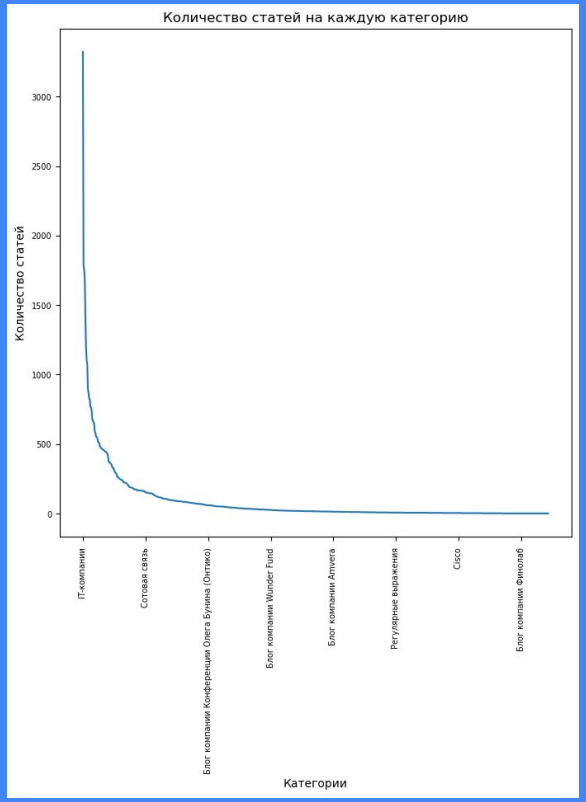
Собираем: название, текст, дату публикации, теги, **хабы**

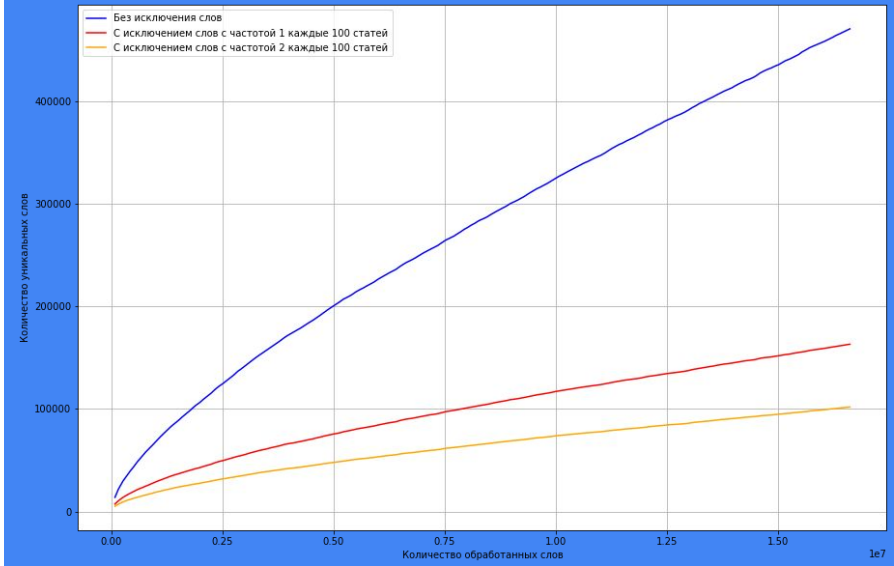
Стек:

- Python3
- requests
- BeautifulSoup
- S3 + pickle

Собрано:
~20K статей за 2023
~10K статей за 2018-2022

Разведочный Анализ Данных





Используемые метрики

Для оценки качества рекомендаций хабов мы используем метрики:

- Precision@k - показывает долю правильно предсказанных хабов среди рекомендованных;
- Recall@k - показывает долю правильно предсказанных хабов среди релевантных;
- F1@k - гармоническое среднее recall@k и precision@k,

$$F1@k = (2 * precision@k * recall@k) / (precision@k + recall@k).$$

Основной метрикой в нашем проекте была выбрана recall@k, так как нам важно предлагать пользователям все необходимые хабы, даже если при этом некоторые рекомендации могут быть излишними.

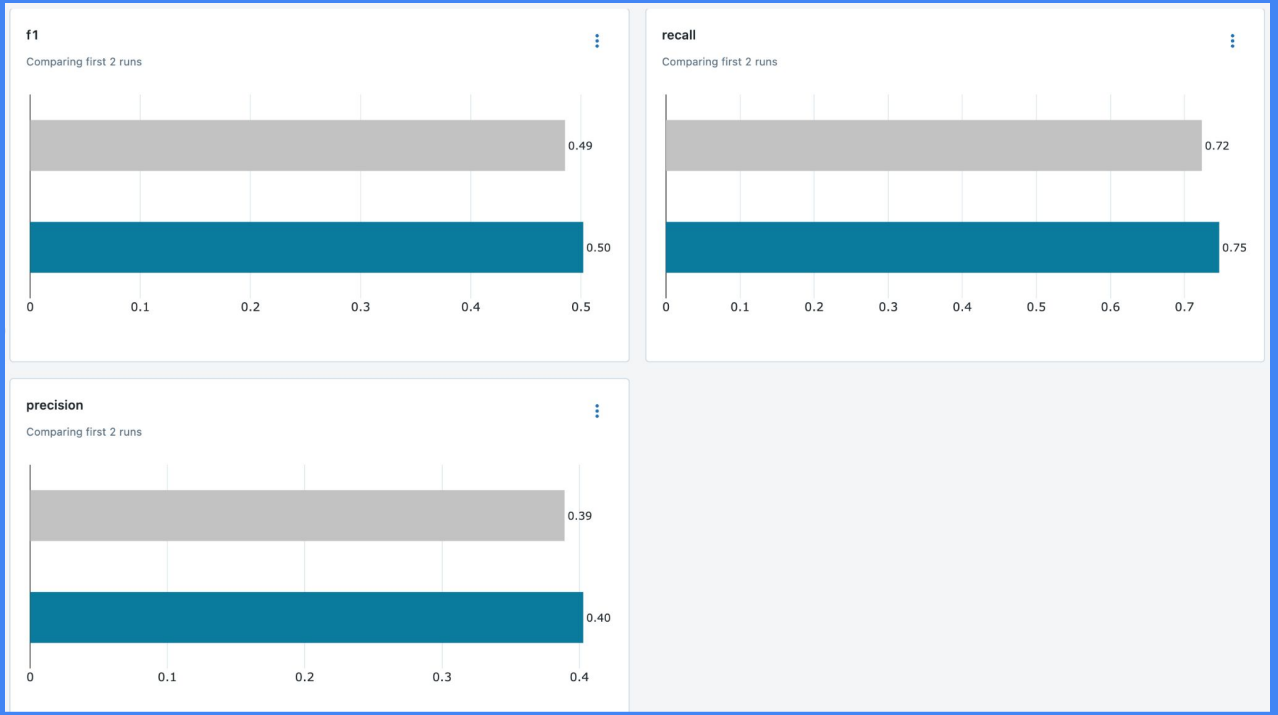
Линейная модель

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Сравнение линейной модели (однослойной нейронной сети) и нейронной сети с двумя слоями.

На вход модели подается вектор полученный из исходного текста путем векторизации, на выходе модель выдает вектор вероятностей принадлежности статьи хабам.

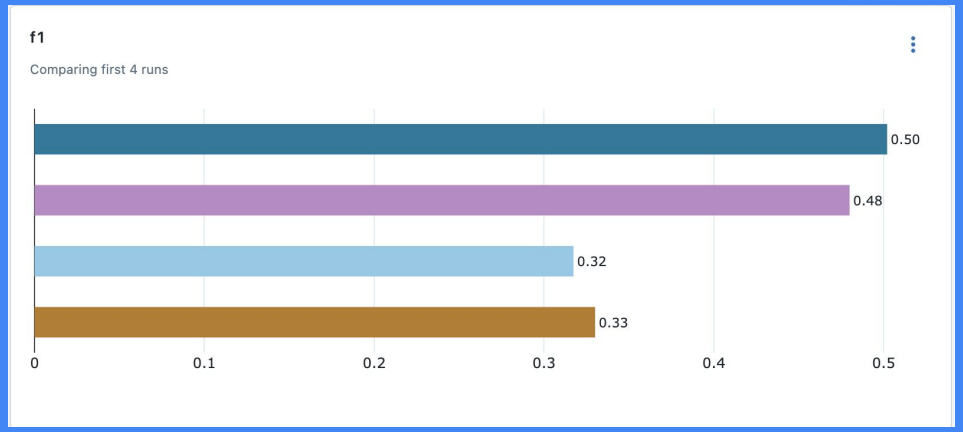
На основе метрик можно сделать вывод, что линейная модель (серый цвет на графиках) показывает себя хуже двухслойной нейронной сети (синий цвет).

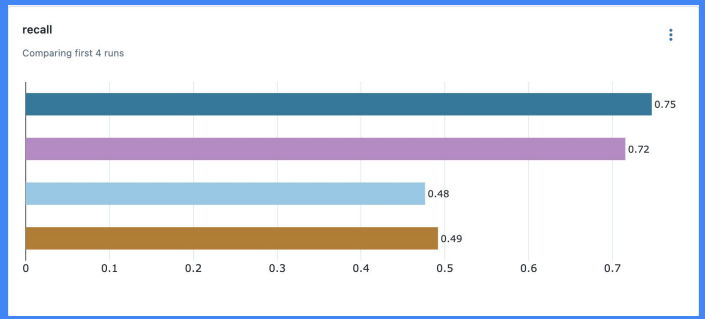


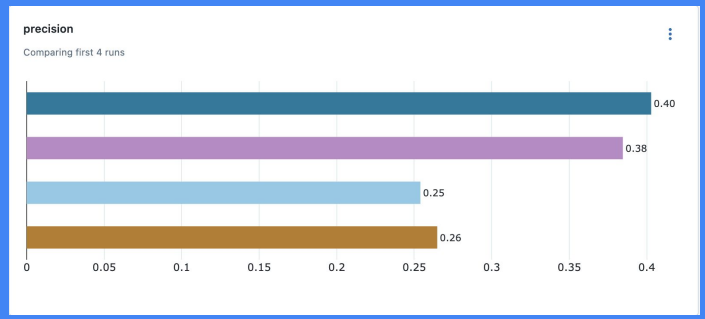
Векторайзеры

Нейронная сеть получает на вход векторизованный текст. Существуют различные способы векторизации текстов, и процесс векторизации как часть предобработки данных может влиять на качество итоговой модели.

Для экспериментов были выбраны 4 различных векторайзера, на основе метрик можно сделать вывод что лучше всего себя показывает tf-idf векторайзер, который векторизует слова на основе отношения частоты их встречаемости в конкретном документе к частоте их встречаемости во всех документах.







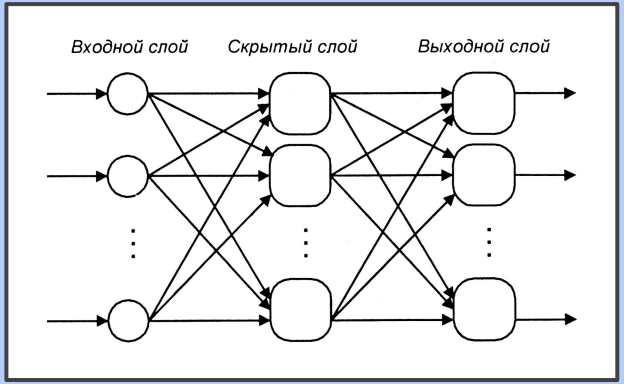
Лучшая модель

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Основная модель: двухслойная нейронная сеть

Вход: tf-idf вектор по нормам слов в статье

Выход: вектор скоров хабов



Стек:

- scikit-learn
- torch
- mlflow

Метрики:

Recall@5 - 0.745789

Precision@5 - 0.401015

F1@5 - 0.500357

Обучение:

20к статей за 2023

20 эпох

Размер модели:

(47 751 + 1)*(1 000) + (1 000 + 1)*(260) =

= 48 012 260 параметров