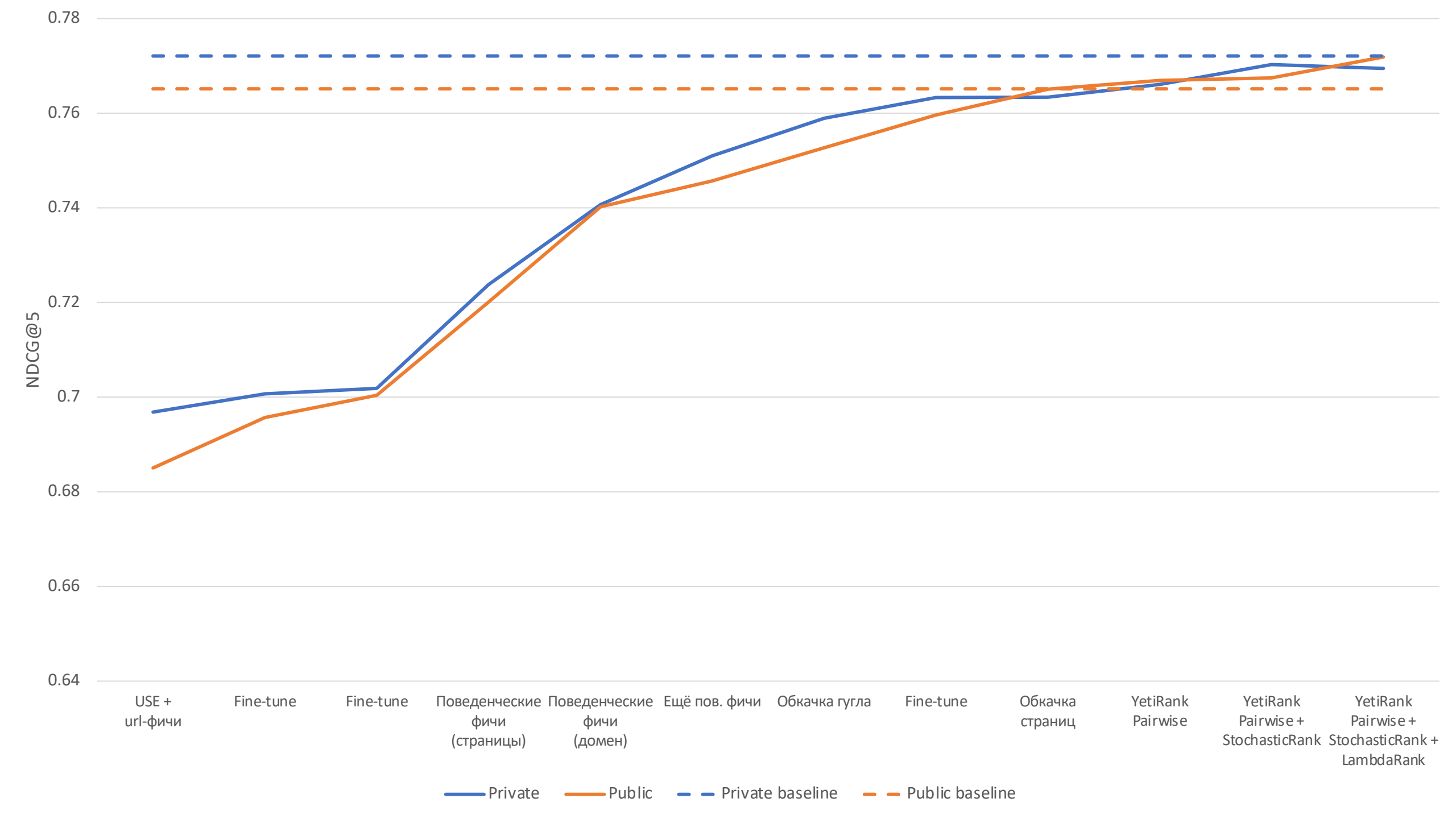
Финальный проект

или как пройти бейзлайн не до конца


0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
NDCG@5
USE + url-фичи
Fine-tune
Fine-tune
Поведенческие фичи (страницы)
Поведенческие фичи (домен)
Ещё пов. фичи
Обкачка гугла
Fine-tune
Обкачка страниц
YetiRank Pairwise
YetiRank Pairwise + StochasticRank
YetiRank Pairwise + StochasticRank + LambdaRank
Private
Public
Private baseline
Public baseline

Предобработка данных

- Запросы: pyaspeller; pymystem3; удаление стоп-слов nltk
- Заголовки: pymystem3

Дополнительные данные

- SERPы гугла
- Сами страницы с документами

URL и текстовые фичи

- Доменная зона (one-hot)
- Porn / forum / news (простой эмпирический классификатор)
- USE: склярное произведение: запрос (правленный / неправленный) и тайтл, запрос и текст
- USE: разность скоров запрос-тайтл / запрос-текст
- BM25 на тайтлах
- Количество одинаковых слов запрос-тайтл, количество уникальных слов запрос/тайтл
- Год в запросе и документе, совпадают ли года

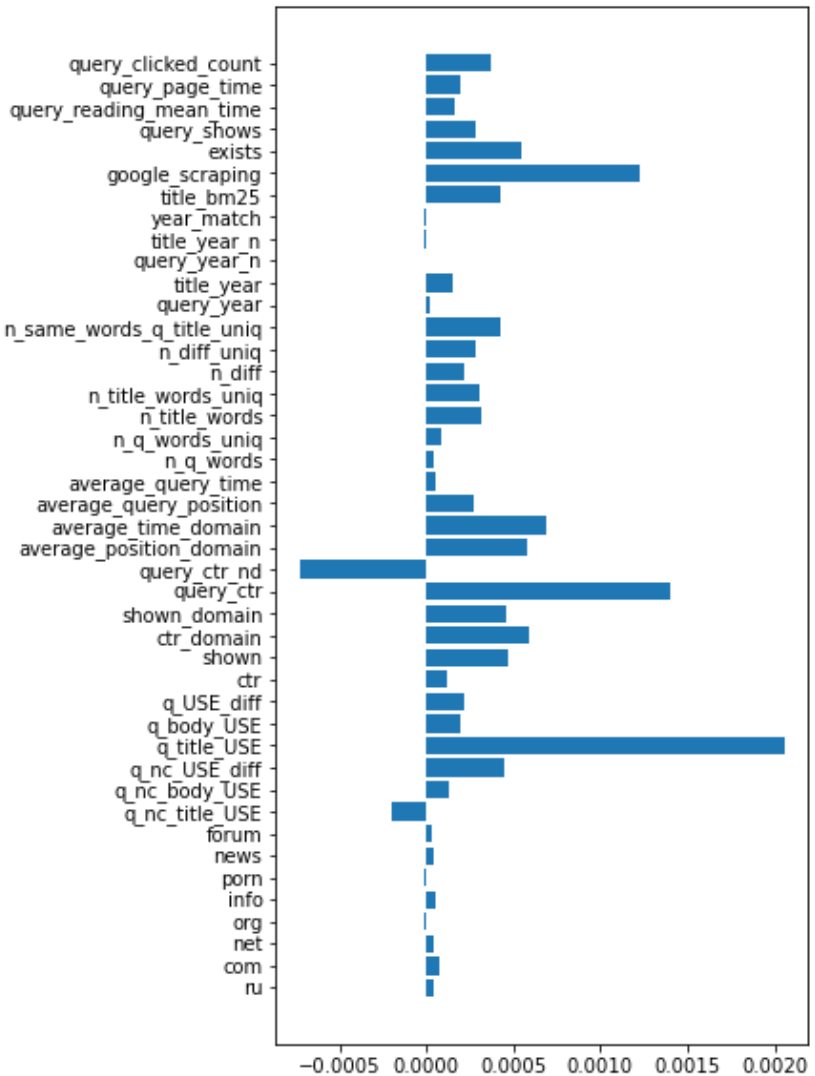
Поведенческие фичи

- По запросу, документу и домену:
  - Количество показов / кликов
  - Количество показов / кликов | запрос
  - Среднее время на SERPe
  - Среднее время перед кликом
  - Средняя позиция документа в SERPe

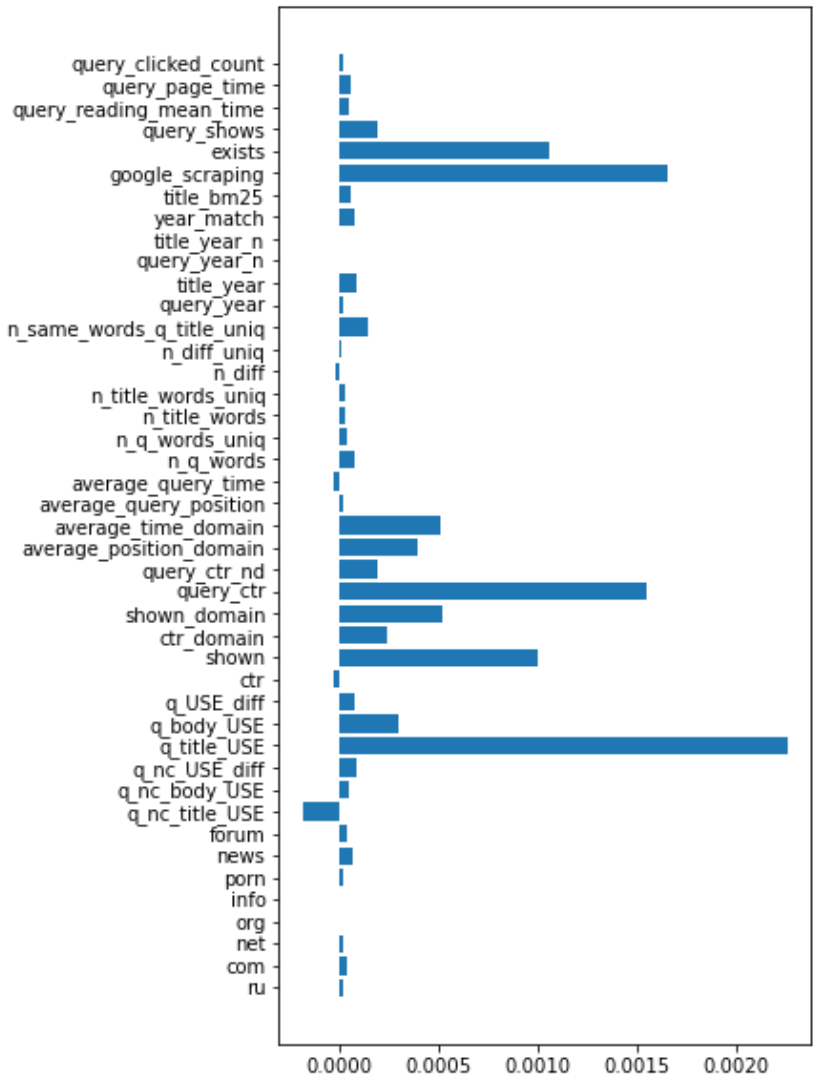
Хитрые фичи

- Наличие документа в выдаче гугла по запросу и его позиция
- Доступность страницы с документом и её размер в байтах

YetiRankPairwise



StochasticRank



LambdaRank