Финальный проект

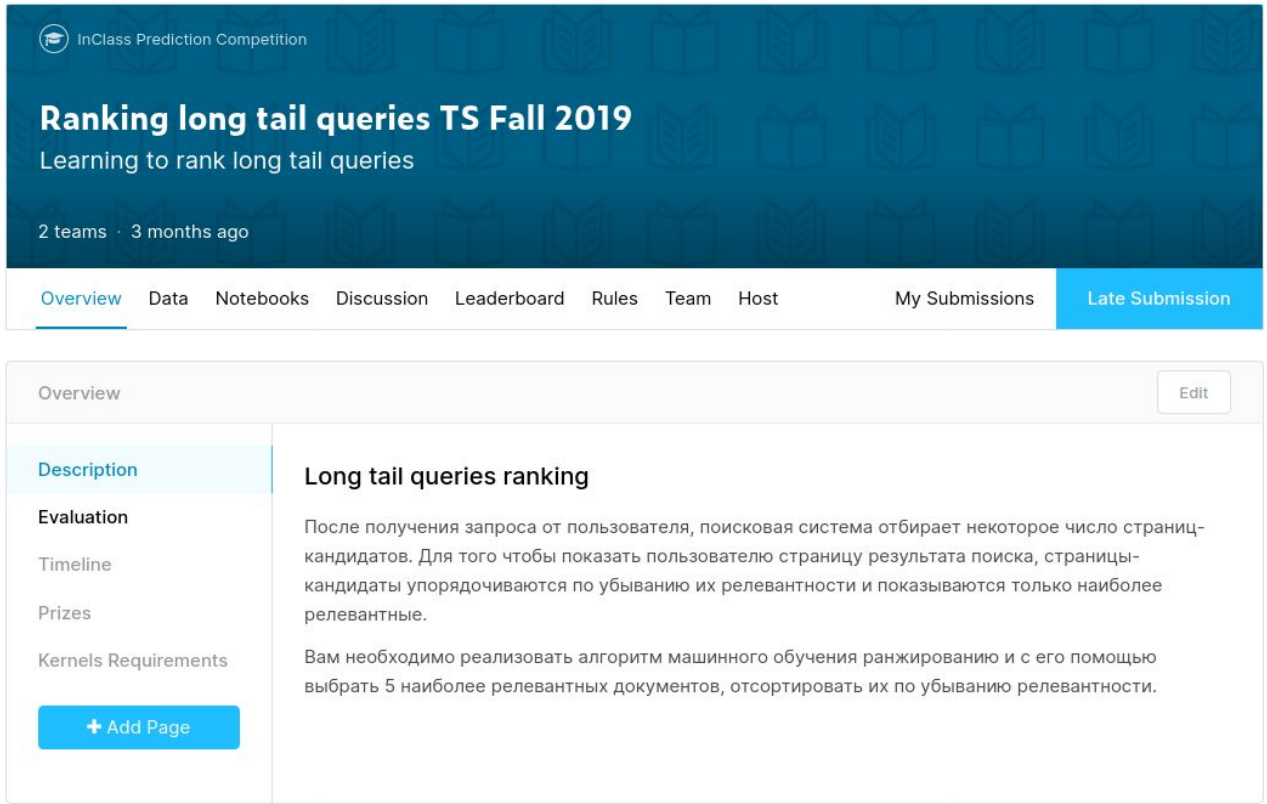
3 семест

Цели

Проверка основных компетенций:

- Data Science
- Information Retrieval
- Big Data

Конкурс по информационному поиску



Данные

Data Description

Edit

Data

- **docs.tsv.gz** - https://cloud.mail.ru/public/NBKs/fVUD7aRph
- **2017.tar** - https://cloud.mail.ru/public/ATwD/z2yotso3k

File descriptions

- **docs.tsv.gz** - files with docs content (DocId <tab> DocTitle <tab> DocContent)
- **2017.tar** - files with click data (Query Text @ Query Geo <tab> List of shown urls <tab> List of clicked urls <tab> Timestamps of clicks for clicked urls)
- **queries.tsv** - файл с запросами в формате (QueryId <tab> QueryText)
- **url.data** - файл с урлами в формате (DocumentId <tab> Url)
- **train.marks.tsv** - файл с оценками
- **sample.csv** - пример сабмита

Data fields

- **QueryId** - идентификатор запроса
- **DocumentId** - идентификатор документа
- **DocTitle** - заголовок документа
- **DocContent** - контент документа

Требования

Необходимо продемонстрировать свои навыки

- Программирования
- Обработки больших данных
- Нейронных сетей
- Ансамблирования
- Информационного поиска

Решение должно содержать

- Классические текстовые факторы
- Семантические факторы (без учителя)
- Семантические факторы с учителем
- Поведенческие факторы
- Факторы должны быть максимально посчитаны на хадупе
- Сглаженные поведенческие факторы
- Над этим всем набором факторов должна работать одна из моделей Learning to rank

Классические текстовые факторы

- Tfldf
- BM25
- BM25F
- Пассажи
- N-grams (языковые модели)
- Псевдорелевантностью
- Синонимы и переформулировки

Семантические факторы без учителя

- LSA, Word2vec, FastText (WMD)
- Doc2vec (поиск похожих доков)
- SentenceSpace, Skip Vectors (поиск похожих предложений)
- ELMO, USE, BERT

Поведенческие факторы

- Базовые кликовые факторы (CTR, длинные клики, позиция)
- Кликовые модели (Positional, DBN, UBM, Cascade)
- Кликовые шаблоны
- Временные факторы (время до клика, время после клика)

Сглаженные поведенческие факторы

- Коллаборативная фильтрация
- Функции переносимости факторов
- Коллаборативное ранжирование

Семантические модели с учителем

- Переводные модели
- Сети с ранним связыванием
- Сети с поздним связыванием
- Ансамбли