Решение конкурса по ранжированию

Шилов Валентин

Public score: 0.78486
Private score: 0.78966

Подготовка данных

- Коррекция запросов через Яндекс спеллер
- Нормализация заголовков, запросов и доков
  - Pymorphy2
- Разделение файлов с контентом доков на файлы по запросам
- Подсчёт поведенческих признаков

Поведенческие факторы

- Усреднённые фичи по документам и доменам
- Фичи для документов по каждому запросу
- Фичи для документов для ближайшего запроса

  - CTR
  - Среднее время, проведённое на странице
  - Сколько раз был кликнут последним
  - ~~PBM~~
    - Бустинг на CTR по позициям, а так же на долях кликов (сколько раз кликнут первым, вторым, … / число кликов)
  - Минимальное расстояние от запроса из логов до интересующего запроса
    - (расстояние Левенштейна, число общих слов)
  - Среднее число кликов до
  - Среднее число кликов после
  - Cascade model

Классические текстовые факторы

- Tfidf (слова, словарные нграммы, буквенные нграммы)
  - Косинусное расстояние между заголовком документа и запросом
  - Косинусное расстояние между заголовком документа и заголовком ранее найденного топового документа
  - Косинусное расстояние между запросом и содержимым документа
  - Косинусное расстояние между содержимым документа и содержимым ранее найденного топового документа
  - Добавление синонимов
- BM25F

Семантические факторы

- FastText
  - заголовок документа <-> запрос
- Universal Sentence Encoder
  - (universal-sentence-encoder-multilingual-large, universal-sentence-encoder-multilingual-qa)
- Bert
  - Slavic BERT
  - RuBERT
  - Multilingual BERT
- Word2vec

CATBOOST

YetiRank , YetiRankPairwise
learning_rate: 0.3
iterations 1500
depth: 5

LightGbm

YetiRank
n_estimators=1000
learning_rate=0.05
max_depth=60

Public score: 0.78486
Private score: 0.78966

Feature Importance

USE_query_answer 0.0014121655472196792
content_tfidf_prev_top 0.0012800984544819594
host_[nafter] 0.0008331748547898421
host_[nlast] 0.0007681707342075406
host_[nbefore] 0.0006412293517727985
bm_features@3 0.0006239362354457656
tf_features@0 0.0004994392099945122
doc_click[nlast] 0.0004630720782545783
host_pos_done_clicks@8 0.0004137773966830238
host_pos_done_clicks@14 0.00039347154205482404
tf_ngram_features_prev@2 0.00038750350570704306
content_tfidf@3 0.0003874106351922091
fasttext_norm 0.0003819986551691912
host_max_pos_show 0.00036900633063741l4
fq_dom_features@23 0.0003412027989799782
doc_log(1.0+self.nclicks) 0.0003331871723959434
content_tf@0 0.0003241621496963276

Спасибо за внимание!