Финальный конкурс по ранжированию

Сульженко Родион

PowerPoint

Признаки (поведенческие)

- DocAvgTime', 'HostAvgTime',
- 'HostCtr', ‘UrlCtr’, 'UrlShows', ‘HostShows’
- 'HostAvgPos', 'UrlAvgPos',
- 'sDBN_clicks', 'sDBN_satisfied', 'sDBN_shows‘, 'sDBN_a‘, 'sDBN_s'

Признаки (поведенческие)

- DocAvgTime', 'HostAvgTime',
- 'HostCtr', ‘UrlCtr’, 'UrlShows', ‘HostShows’
- 'HostAvgPos', 'UrlAvgPos',
- 'sDBN_clicks', 'sDBN_satisfied', 'sDBN_shows‘, 'sDBN_a‘, 'sDBN_s'

Признаки (текстовые лексические)

- ‘Bm25Okapi’ по заголовкам
- ‘Bm25Plus’ по заголовкам
- ‘Bm25L’ по заголовкам
- ‘Bm25Plus’ по текстам (первые 1к слов)

Признаки (поведенческие)

- DocAvgTime', 'HostAvgTime',
- 'HostCtr', ‘UrlCtr’, 'UrlShows', ‘HostShows’
- 'HostAvgPos', 'UrlAvgPos',
- 'sDBN_clicks', 'sDBN_satisfied', 'sDBN_shows‘, 'sDBN_a‘, 'sDBN_s'

Признаки (текстовые)

- ‘Bm25Okapi’ по заголовкам
- ‘Bm25Plus’ по заголовкам
- ‘Bm25L’ по заголовкам
- ‘Bm25Plus’ по текстам (первые 1к слов)

Признаки (семантические)

- ‘universal-sentence-encoder-multilingual-qa/3’

По заголовкам

Признаки (поведенческие)

- DocAvgTime', 'HostAvgTime',
- 'HostCtr', ‘UrlCtr’, 'UrlShows', ‘HostShows’
- 'HostAvgPos', 'UrlAvgPos',
- 'sDBN_clicks', 'sDBN_satisfied', 'sDBN_shows‘, 'sDBN_a‘, 'sDBN_s'

Признаки (текстовые)

- ‘Bm25Okapi’ по заголовкам
- ‘Bm25Plus’ по заголовкам
- ‘Bm25L’ по заголовкам
- ‘Bm25Plus’ по текстам (первые 1к слов)

Признаки (семантические)

- ‘universal-sentence-encoder-multilingual-qa/3’

По заголовкам

Модель

- Голосование:
  - CatboostRanker
  - LgbmRanker x2

Что еще пробовал

- Разные USE
- BERT embeddings
- Пробовал аналогичные признаки для переведенных на англ. заголовках
- TfIdf (оказался значительно хуже Bm25)
- Словограммы и буквограммы
- Модель несовпадения слов
  http://www.machinelearning.ru/wiki/images/a/a6/2018_617_VikulinVA.pdf
- Разбивать Url на токены -> в векторайзер
- Блендинг на разных подмножествах признаков

$$sim(q_i, q_j) = \frac{1}{2} \cdot \Big(\sum_{i \in q_i, i \notin q_j} w'_{i,q_i} + \sum_{j \in q_j, j \notin q_i} w'_{j,q_j} \Big).$$

Что не успел

- Посчитать статистики по текстам целиком
- Биграммы по текстам
- Использовать репутацию сайтов (например, яндекс ИКС) в качестве признака
- Тематический анализ (LSA)