Текстовая релевантность

Сульженко Родион

Модель 1 0.36402

Обработка страниц: текст лемматизирован, убраны некоторые стоп-слова

Обработка запросов: аналогично обработке страниц

Ранжирование: считается TF-IDF score

$$w_{t,d} = (1 + log\ tf_{t,d}) \cdot log \frac{N}{df_t}$$

$$Score(q, d) = \sum_{t \in q \cap d} tf.idf_{t,d}$$

Модель 2 0.37089

Обработка страниц: текст лемматизирован, убраны некоторые стоп-слова

Обработка запросов: аналогично обработке страниц

Ранжирование: TF-IDF score считается отдельно для заголовка и текста, финальный скор – их сумма.

$$w_{t,d} = (1 + log\ tf_{t,d}) \cdot log \frac{N}{df_t}$$

$$Score(q, d) = \sum_{t \in q \cap d} tf.idf_{t,d}$$

Модель 3 0.63754

Обработка страниц: текст лемматизирован, убраны некоторые стоп-слова

Обработка запросов: аналогично обработке страниц

Ранжирование: BM25 по двум зонам независимо: score считается отдельно для заголовка и текста, финальный скор – их сумма. Вес у заголовка – 1.5

$$score(Q,D) = \sum_{i=1}^{n} Idf(q_i) \frac{f(q_i,D)(k_1+1)}{f(q_i,D) + k_1(1 - b + b\frac{dl}{avgdl})}$$

$$\mathrm{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

- f(qi_d) – частота слова qi в документе D,
- K1 – 2.0,
- b = 0.75

Модель 4 0.68950

Обработка страниц: текст лемматизирован, убраны некоторые стоп-слова

Обработка запросов: исправлены опечатки с помощью YandexSpellchecker api (в т.ч. исправление раскладки), затем аналогично обработке страниц

Ранжирование: BM25 по двум зонам независимо: score считается отдельно для заголовка и текста, финальный скор – их сумма. Вес у заголовка – 1.5

	wrong	correct
0	жировики на спине можно ли применить пиявки	жировики на спине можно ли применять пиявки
1	rfr ghjgbcfnm ghjcnj flvbre	как прописать просто админку
2	можно ли ставить горчичник при ларингите	можно ли ставить горчичники при ларингите
3	каквернуть налог в аэропорту парижа	как вернуть налог в аэропорту парижа
4	что входитв стоимиость авиабилетов	что входит в стоимость авиабилетов
5	какоформить отказ от подписания акта о невыход...	как оформить отказ от подписания акта о невыхо...
6	сколько платят за кап ремонт в воркуте	сколько платят за капремонт в воркуте
7	что нужно для того что бы провести свет в гараж	что нужно для того чтобы провести свет в гараж
8	как устоновит ps cs6	как установить ps cs6

Модель 5 (финальная) 0.75695

Обработка страниц: текст лемматизирован, убраны некоторые стоп-слова

Обработка запросов: исправлены опечатки с помощью YandexSpellchecker api (в т.ч. исправление раскладки), затем аналогично обработке страниц

Ранжирование: Для каждого запроса в ранжировании участвуют не все документы, а только документы, указанные для него в sample_submission

BM25 по двум зонам независимо: score считается отдельно для заголовка и текста, финальный скор – их сумма. Вес у заголовка – 1.5

Модель 6 0.70717

Обработка страниц: текст лемматизирован, убраны некоторые стоп-слова

Обработка запросов: исправлены опечатки с помощью YandexSpellchecker api (в т.ч. исправление раскладки), затем аналогично обработке страниц

Ранжирование: пассажный скор.

Скользящее окно по тексту, скор - сколько максимум слов в окне совпадает со словами запроса. Нормируется делением на число слов запроса. (размер окна – число слов запроса)

Далее данная функция модернизировалась, чтобы учитывать заголовок и текст отдельно, а также количество найденных окон с максимальным числом общих токенов.

Результат – около 0.7

Пробовал также объединить пассажный скор с бм25 скором, но результата особо не дало (максимум 0.75742)

Что еще пробовал

- Понижать в релевантности страницы, распознанные с большой вероятностью как спам, с помощью классификатора спама из прошлого задания
- Учитывать при обработке текстов только те слова, которые встречаются в запросах
- Дополнять запрос синонимами к его словам
- Нормировать скор от бм25 и сочетать с другими моделями
- Биграммы

Спасибо за внимание!