Информационный Поиск

Домашнее задание 2

Срок сдачи: 11.11.2016, 23:59

1. **[10pt]** Рассмотренные методы исправления опечаток не работают напрямую при пропуске пробела (например, informationretrieval). Опишите, как исправлять такие опечатки (не обязательно на основе рассмотренных методов).

2. **[5pt]** Мы рассмотрели два типа методов для рекомендации запросов, аналогичных заданному запросу:

   a) Рекомендовать запросы, встречающиеся в одной сессии с заданным запросом.

   b) Рекомендовать те запросы, у которых множество кликнутых результатов сильно пересекается с аналогичным множеством для заданного запроса.

   Какие еще методы для рекомендации запросов вы можете предложить?

3. **[5pt]** Вы планируете использовать следующие методы поиска: модель векторного пространства с весами TF-IDF, BM25, языковую модель. Какую минимальную информацию должен содержать индекс, чтобы поддерживать эффективное использование этих методов? Какую информацию нет смысла хранить в индексе? Как ее нужно хранить? Дайте развернутый ответ.

4. **[10pt]** Отранжируйте документы из таблицы 1 по запросу "car insurance" с использованием модели векторного пространства и весов TF-IDF.

Слово	idf	tf		
		Документ 1	Документ 2	Документ 3
car	1.65	27	4	24
auto	2.08	3	33	0
insurance	1.62	0	33	29
best	1.60	14	0	17

Таблица 1: Частота слов в документах и обратная документная частота слов.

5. **[5pt]** Рассмотрим следующий запрос и три результата.

   Q information retrieval course

   D1 Information Retrieval and Web Search

   D2 Introduction to Information Retrieval

   D3 Text Retrieval and Search Engines

   Результаты 1 и 3 – это страницы соответствующих курсов, поэтому пользователь пометил их как релевантные. Результат 2 – это страница с книгой, поэтому пользователь пометил его как нерелевантный.

   Примените алгоритм Роккио и выпишите вектор запроса после учета обратной связи по релевантности. Элементы вектора перечислите в алфавитном порядке. Считайте, что компоненты векторов содержат только частоты слов (без обратной документной частоты и нормировки). Параметры алгоритма Роккио: $\alpha = 1, \beta = 0.75, \gamma = 0.25$.

6. **[10pt]** Выпишите формулу BM25 для длинных запросов. Опишите ее составляющие. Каким образом каждая составляющая влияет на ранжирование (т.е. что происходит с ранжированием результатов при изменении каждой из составляющих)?

7. **[10pt]** Пусть бинарная случайная величина X_t – это индикатор того, что слово t встречается в документе (т.е. $X_t = 1$, если слово t есть в документе, и $X_t = 0$, если слова t нет в документе). $P_t = P(X_t = 1 \mid d)$ – это вероятность того, что слово t встречается в документе d.

   Примените метод максимального правдоподобия (MLE) для формального вычисления P_t и покажите, что $P_t = \frac{tf(t,d)}{dl(d)}$, где $tf(t,d)$ – это частота слова t в документе d, а $dl(d)$ – это длина документа d.

8. **[5pt]** Рассмотрим коллекцию из двух документов.

   D1 A language model is a probability distribution over words or sequences of words.

   D2 A language model is used in many natural language processing applications.

   Выпишите сглаженную униграмную языковую модель для каждого документа. Используйте сглаживание Jelinek-Mercer с параметром $\lambda = 0.5$. Отранжируйте эти документы по запросу “many words”.