Тест начат	среда, 10 января 2024, 23:24
Состояние	Завершены
Завершен	среда, 10 января 2024, 23:30
Прошло времени	5 мин. 48 сек.
Оценка	5,00 из 5,00 (100%)

Вопрос Инфо

Вопрос 1

Выполнен

Баллов: 1,00 из 1,00

Почему для задачи тематической классификации предлоги, союзы и местоимения практически бесполезны?

Выберите один ответ:

a. Потому что они встречаются практически во всех текстах вне зависимости от тематики

b. Потому что они имеют слишком большой вес

c. Потому что предлоги могут привести к неправильной классификации

Ваш ответ верный.

Вопрос Инфо

Вопрос 2

Верно

Баллов: 1,00 из 1,00

Пусть в некотором языке есть N=3 слова - А, Б и В. Их ранги - 1, 2 и 3 (нумерация рангов начинается с 1). Найдите вероятности встретить каждое из этих слов в тексте при условии, что относительные частоты слов распределены по Ципфу с s = 2.

Представьте ответ в форме трёх чисел P(A) P(Б) P(В), разделённых пробелом, с точкой . в качестве десятичного разделителя, например, 0.1 0.2 0.3. Ответ округлите до не менее чем двух знаков после запятой. Напомним, $f(rank, s, N) = \frac{1}{rank^s \cdot \sum_{i=1}^{N} i^{-s}}$.

Ответ: (штрафной режим: 0 %)

```
0.73 0.18 0.08
```

Прошли все тесты! ✔

Верно

Баллы за эту попытку: 1,00/1,00.

Вопрос Инфо

Вопрос 3

Выполнен

Баллов: 1,00 из 1,00

Отметьте верные утверждения про TF-IDF и закон Ципфа.

Под значимостью в вариантах ответа понимается потенциальная полезность для задач тематической классификации - баланс частотности и специфичности.

- [ ] a. Слово считается значимым, когда оно одновременно частотное в документе и частотное в коллекции.
- [ ] b. Больший TF, как правило, соответствует глобально более значимым словам.
- [x] c. Больший IDF, как правило, соответствует специальной лексике и словам с опечатками.
- [ ] d. Больший IDF, как правило, соответствует общеупотребимым словам.
- [x] e. Слово считается значимым, когда оно одновременно частотное в документе и не очень частотное в коллекции.
- [x] f. Больший TF, как правило, соответствует общеупотребимым словам.
- [ ] g. Если слово часто встречается в документе, то оно гарантированно часто встречается и во всей коллекции (не считая предлогов, союзов и т.п.).
- [ ] h. Слово считается значимым, когда оно одновременно редкое в документе и частотное в коллекции.

Ваш ответ верный.

Вопрос Инфо

Вопрос 4

Выполнен

Баллов: 1,00 из 1,00

Допустим, Вы хотите строить матрицу признаков с помощью TF-IDF на биграммах токенов (N-граммах с N=2). Оцените, приблизительно, наибольшее количество уникальных биграмм в словаре для достаточно большой коллекции. Предполагайте, что в текстах используется 1000 уникальных токенов.

Выберите один ответ:

- a. 50000
- b. 1000
- c. 2000
- d. 1000000
- e. 10000

Ваш ответ верный.

Вопрос Инфо

Вопрос 5

Выполнен

Баллов: 1,00 из 1,00

Проанализируйте формулы для точечной взаимной информации и отметьте верные утверждения.

$$pmi(l, w) = log\frac{p(w,l)}{p(w)p(l)} = log\frac{p(l|w)}{p(l)} = log\frac{p(w|l)}{p(w)}$$

Для правильного решения этого задания можно рассмотреть несколько крайних случаев и подставлять соответствующие значения вероятностей в эти формулы: когда события l и w всегда происходят вместе, когда одно происходит, а другое не происходит и т.п. К тому же, в качестве событий могут выступать не только факты встречаемости слов и меток ;)

- [x] a. Взаимная информация симметрична pmi(l, w) = pmi(w, l)
- [ ] b. Взаимная информация максимальна, когда события w и l независимы.
- [ ] c. Взаимная информация минимальна, когда события w и l независимы.
- [x] d. Взаимная информация максимальна, когда события w и l всегда появляются одновременно.
- [x] e. Взаимная информация минимальна, когда события w и l несовместны (если происходит одно, второе не может произойти).
- [x] f. Взаимную информацию можно применить для оценки совместной встречаемости слов pmi(w_1, w_2)

Ваш ответ верный.