Вопрос 1

Выполнен

Баллов: 1,00 из 1,00

Термин "дистрибутивная семантика" происходит от двух других терминов:

- семантика - наука о смыслах
- дистрибутивная - от английского термина distribution (имеется в виду probability distribution, распределение вероятностей)

Методы дистрибутивной семантики пытаются извлечь смысл слов, анализируя распределение ...

Выберите один ответ:

a. материальных благ в обществе

b. вероятностей совместной встречаемости слов в рамках одного фрагмента текста (или вероятностей встретить одни слова в контексте других)

c. печенек по офису

Ваш ответ верный.

Вопрос Инфо

Вопрос 2

Выполнен

Баллов: 1,00 из 1,00

Вы построили матрицу совместной встречаемости с размером словаря 100000 (сто тысяч) и сжали её в две матрицы размерности 100000×300 и 300×100000. Сколько памяти Вы сэкономили этой операцией при том, что числа хранятся во float (4 байта)?

Ответ запишите в гигабайтах, округлив до целого значения ($1\text{ГБ} = 1024^3\text{Б}$).

Ответ: 37

Ваш ответ верный.

Вопрос Инфо

Вопрос 3

Выполнен

Баллов: 1,00 из 1,00

Зачем нужно сглаживание частот через логарифмирование?

Выберите один ответ:

a. Чтобы покрасоваться перед девочками (мальчиками)

b. Чтобы сделать распределение менее контрастным и сжать диапазон значений

c. Чтобы сделать распределение более контрастным и расширить диапазон значений

Ваш ответ верный.

Вопрос Инфо

Вопрос 4

Выполнен

Баллов: 1,00 из 1,00

Какую задачу решает Word2Vec Skip Gram?

Выберите один ответ:

a. Классификации - предсказания оценки студента за курс по обработке текстов

b. Классификации - предсказания соседних слов по заданному слову

c. Регрессии - предсказания количества совместных упоминаний какого-либо слова в контексте заданного слова

d. Классификации - предсказания тематики слова

Ваш ответ верный.

Вопрос Инфо

Вопрос 5

Выполнен

Баллов: 1,00 из 1,00

Общий алгоритм обучения FastText Skip Gram Negative Sampling выглядит следующим образом:

1. Очистить и токенизировать обучающую коллекцию документов
2. Построить словарь - подсчитать частоты всех целых токенов и N-грамм заданной длины (например, от 3 до 6 символов). При построении словаря раз в заданное число шагов прореживать словарь - удалить из словаря токены, набравшие с предыдущего прореживания меньше всего употреблений (или меньше заданного порога).
3. Проход по корпусу скользящим окном заданной ширины, для каждой позиции окна выполнять шаги 4-7.
4. Для текущего словоупотребления в центре окна выделить его N-граммы, содержащиеся в словаре (то есть только достаточно частотные N-граммы)
5. Вычислить вектор центрального токена, усреднив вектора целого токена (если он есть в словаре) и всех N-грамм, выделенных на шаге 4.
6. Выбрать случайным образом отрицательные слова (сделать negative sampling).
7. Обновить следущие вектора так, чтобы улучшить оценку правдоподобия:
    1. N-грамм, участвовавших в получении вектора центрального токена,
    2. контекстные вектора всех токенов в окне, кроме центрального,
    3. контекстные вектора отрицательных слов.
8. Повторять шаги 3-7 заданное число раз или до сходимости.

Отметьте N-граммы, вектора которых будут обновляться при обучении FastText SkipGram каждый раз, когда в качестве центрального слова будет выступать слово "бутявка".

Внимание! FastText учитывает само центральное слово как n-грамму, только если оно достаточно частотное. В этом задании у нас такой статистики нет, поэтому само слово "бутявка" будем считать достаточно частотным.

Выберите один или несколько ответов:

- [x] a. бутявка
- [x] b. бут
- [ ] c. воч
- [x] d. явка
- [ ] e. кав
- [x] f. вка

Ваш ответ верный.