

Вопрос 1

Верно

Баллов: 1,00 из 1,00

Дано несколько текстов. Преобразуйте их в бинарные вектора.

Казнить нельзя, помиловать.

Казнить, нельзя помиловать.

Нельзя не помиловать.

Обязательно освободить.

При токенизации текстов был построен следующий словарь: казнить не нельзя обязательно освободить помиловать.

Ответ (вектора для приведённых выше текстов) запишите в следующем формате:

- Одна строка соответствует одному тексту.
- Значения признаков следует разделять пробелами
- Порядок значений признаков должен соответствовать порядку слов в словаре (приведён выше).

Перенос строки можно заменить на

//. Пример ответа для первых двух текстов (не забудьте их тоже вставить в ответ ;)):

```
1 0 1 0 0 1
```

```
1 0 1 0 0 1
```

Ответ: (штрафной режим: 0 %)

```
1 0 1 0 0 1
1 0 1 0 0 1
0 1 1 0 0 1
0 0 0 1 1 0
```

	Comment
✓	Верное решение. Так держать!

Прошли все тесты! ✓

Верно

Баллы за эту попытку: 1,00/1,00.

Вопрос 2

Выполнен

Баллов: 1,00 из 1,00

После токенизации всех документов в корпусе строится словарь, содержащий для каждого уникального токена количество его употреблений в корпусе. Затем из этого словаря удаляются самые **редкие** слова.

Как Вы думаете, зачем это может быть нужно?

Выберите один или несколько вариантов ответа.

Выберите один или несколько ответов:

- ☒ а. Чтобы уменьшить риск переобучения
- ☒ b. Чтобы сэкономить память, требуемую для размещения датасета и модели
- ☐ c. Чтобы убрать союзы, местоимения, предлоги
- ☒ d. Чтобы убрать слова, содержащие опечатки

Ваш ответ **верный**.

Вопрос **3**

Верно

Баллов: 1,00 из 1,00

Выпишите все символьные N-граммы для $N=\{2, 3\}$ для слова "язык".

В ответе необходимо указать все n-граммы через пробел, порядок n-грамм не важен. Регистр букв в ответе должен совпадать с регистром букв в исходном слове.

При желании, для выделения N-грамм можно написать несложную программу на Python.

Ответ: (штрафной режим: 0 %)

яз зы ык язы зык

	Comment
✓	Верное решение. Так держать!

Прошли все тесты! ✓

Верно

Баллы за эту попытку: 1,00/1,00.

Вопрос 4

Выполнен

Баллов: 1,00 из 1,00

Какие способы векторизации текста могут лучше подойти для задачи классификации сообщений из Twitter?

При использовании N-грамм можно выделять N-граммы сразу нескольких длин (например, 3,4,5-граммы).

Выберите один или несколько вариантов ответа.

Выберите один или несколько ответов:

- ☒ a. N-граммы символов, бинарный вектор
- ☐ b. N-граммы токенов, взвешивание по частоте
- ☐ c. Целые токены, бинарный вектор
- ☒ d. N-граммы символов, взвешивание по частоте
- ☐ e. Целые токены, взвешивание по частоте
- ☐ f. N-граммы токенов, бинарный вектор

Ваш ответ верный.

Вопрос Инфо

Вопрос 5

Выполнен

Баллов: 1,00 из 1,00

Предположение о независимости словоупотреблений - упрощение, которое мы допускаем, когда строим матрицы признаков по методу бинарных векторов или TF-IDF. Оно проявляется в том, что когда мы заполняем значение для некоторого слова (например, полёт), мы никак не меняем значения для других, сильно связанных с ним слов (например, синонимов - переезд, путешествие, поездка и т.п.).

На языке теории вероятностей предположение о независимости можно описать формулами

$$P(w_1|w_2, d) = P(w_1|d)$$

$$P(w_1, w_2|d) = P(w_1|d)P(w_2|d),$$

где $P(w|d)$ - вероятность встретить слово w в документе d .

Как Вы думаете, почему это может быть плохо?

Выберите один или несколько вариантов ответа.

Выберите один или несколько ответов:

- ☐ а. Модель с таким предположением вообще не работает
- ☒ b. Такая модель более чувствительна к качеству обучающей выборки по сравнению с моделью, которая учитывает отношения между словами
- ☒ с. Модель может хуже работать на новых текстах, содержащих синонимы слов из обучающей выборки, и не содержащих сами эти слова
- ☐ d. Это приводит к чрезмерному увеличению размерности пространства признаков

Ваш ответ верный.

Вопрос Инфо

Вопрос 6

Выполнен

Баллов: 1,00 из 1,00

Отметьте основные недостатки линейных моделей для классификации текстов, принимающих на вход разреженные вещественные вектора, извлечённые из документов через подсчёт отдельных токенов: бинарные вектора (one-hot) или TF-IDF.

Выберите один или несколько вариантов ответа.

Выберите один или несколько ответов:

- ☒ а. чувствительность к шуму (к опечаткам, случайным словам, редким метафорам)
- ☒ b. невозможно учитывать структуру фраз
- ☐ с. большой, по сравнению с моделями, работающими с 2-граммами токенов, размер признакового пространства
- ☐ d. высокая, по сравнению с нейросетями, вычислительная сложность
- ☒ e. предположение о независимости словоупотреблений
- ☐ f. нужна гигантская размеченная обучающая выборка

Ваш ответ верный.

[◀ 1.4 В общих чертах: Лингвистический анализ](#)

Перейти на...

[1.6 Прикладные задачи обработки текста и итоги ▶](#)