Диагностическое тестирование по дисциплине "Инструменты, подходы и методы обработки структурированных и неструктурированных данных"

Всего: 0/20

Для аттестации необходимо набрать 70% правильных ответов.

ФИО: *

Демьянцев Виталий Владиславович

1. ... - это процесс преобразования текстовых данных в числовой формат, который может быть обработан компьютером. *0 из 1

векторизация текста

2. Какой тип базы данных предназначен для хранения неструктурированных данных? *1 из 1

- [x] 1. NoSQL база данных
- [ ] 2. Реляционная база данных
- [ ] 3. Объектно-ориентированная база данных
- [ ] 4. Иерархическая база данных

3. Какой метод используется для обработки текстовых данных? * 1 из 1

- [ ] 1. Случайный лес
- [ ] 2. K-средних
- [ ] 3. DBSCAN
- [x] 4. Токенизация

4. В чем принципиальная необходимость превращения неструктурированных данных в структурированный формат? *1 из 1

- [x] 1. возможность компьютерной обработки.
- [ ] 2. возможность обработки человеком.
- [ ] 3. наглядность.
- [ ] 4. уменьшение объема.

5. Какой процент неструктурированных данных в Интернете? * 1 из 1

- [ ] 1. 0%
- [ ] 2. 20%
- [x] 3. 80%
- [ ] 4. 100%

6. Как в Python хранится вещественное число? * 1 из 1

- [ ] 1. float, занимает 32 бита (4 байта)
- [x] 2. float, int, по сути, double, занимает 64 бита (8 байта)
- [ ] 3. массив необходимой (переменной) длины из 32 битных целых чисел
- [ ] 4. два массива (мантисса и экспонента) необходимой длины из 32 битных целых чисел

7. Что используется для решения задачи линейной регрессии? * 1 из 1

- [ ] 1. Метод бисекции
- [ ] 2. Метод опорных векторов
- [x] 3. Метод наименьших квадратов
- [ ] 4. Логистическая функция потерь

8. Какой инструмент используется для обработки структурированных данных? *1 из 1

- [ ] 1. Графические редакторы
- [ ] 2. Текстовые редакторы
- [ ] 3. Операционные системы
- [x] 4. Базы данных

9. Каково количество возможных различных бинарных строк длины 8? *1 из 1

256

12. Какая информация о словах не сохраняется при использовании представления мешок слов (bag of words)? *1 из 1

- [ ] 1. количество разных слов в документе;
- [x] 2. порядок вхождений слов в документ;
- [ ] 3. число вхождений слов в документ;
- [ ] 4. число всех слов в документе.

10. Соотнесите определение и термин: *
Частично верный ответ не приносит баллов.

	Токенизация	Лемматизация	Стемминг	Баллы
разбиение текста на отдельные слова или словоформы	◉	○	○	0 из 0
приведение слова к его основе путем отбрасывания окончаний	○	○	◉	1 из 1
приведение слова к его нормальной форме	○	◉	○	0 из 0

11. Соотнесите описание и описываемый метод кластеризации: *
Частично верный ответ не приносит баллов.

	K-средних	Иерархическая кластеризация	DBSCAN	Баллы
находит плотные регионы в пространстве объектов и относит их к кластерам, а объекты, находящиеся вне плотных регионов, - выбросами	○	○	◉	1 из 1
строит иерархическое дерево кластеров, объединяя на каждом шаге два ближайших кластера	○	◉	○	0 из 0
разбивает данные на K кластеров, минимизируя сумму квадратов расстояний между точками и центроидами кластеров	◉	○	○	0 из 0

13. Соотнесите тип данных и описание с примером: *
Частично верный ответ не приносит баллов.

	Категориальные	Количественные	Бинарные	Баллы
можно измерить и представить числами (например, возраст, доход, количество товаров)	○	◉	○	1 из 1
могут принимать ограниченное количество значений из заданного набора (например, цвет, пол, тип автомобиля)	◉	○	○	0 из 0
могут принимать только два значения (например, да/нет, 0/1, муж/жен)	○	○	◉	0 из 0

14. Выберите правильный порядок применения методов предобработки данных. *0 из 1

- [ ] 1. Обработка пропущенных значений; Удаление выбросов; Нормализация данных; Кодирование категориальных признаков
- [ ] 2. Удаление выбросов; Нормализация данных; Кодирование категориальных признаков; Обработка пропущенных значений;
- [x] 3. Нормализация данных; Обработка пропущенных значений; Удаление выбросов; Кодирование категориальных признаков
- [ ] 4. Кодирование категориальных признаков; Обработка пропущенных значений; Удаление выбросов; Нормализация данных

15. Выберите правильный порядок применения методов обработки текстовых данных. *1 из 1

- [ ] 1. Стемминг / Лемматизация, Токенизация, Синтаксический анализ;
- [x] 2. Токенизация, Стемминг / Лемматизация, Синтаксический анализ
- [ ] 3. Токенизация, Синтаксический анализ, Стемминг / Лемматизация;
- [ ] 4. Синтаксический анализ, Токенизация, Стемминг / Лемматизация

16. Какие из следующих языков программирования наиболее часто используются для обработки и анализа данных? *1 из 1

- [x] 1. R
- [ ] 2. C++
- [ ] 3. Java
- [x] 4. Python

17. Какие из следующих методов можно использовать для обработки пропущенных значений в структурированных данных? *0 из 1

- [x] 1. Интерполяция
- [x] 2. Замена пропущенных значений на медианное значение столбца
- [x] 3. Удаление строк с пропущенными значениями
- [x] 4. Замена пропущенных значений на среднее значение столбца

18. Какие из следующих методов обработки текстовых данных используются для извлечения ключевых слов из текста? *0 из 1

- [ ] 1. Лемматизация
- [ ] 2. LDA
- [ ] 3. Стемминг
- [x] 4. TF-IDF

19. Как будет выглядеть векторное представление мешка слов для набора текстов: «Это были лучшие времена» и «Это было худшее время»? *0 из 1

- [ ] 1. [1 1 1 0 0], [1 1 0 1 1];
- [x] 2. [1 1 0 1 0 0], [1 0 0 1 1];
- [ ] 3. [1 1 1 0], [1 1 0 1];
- [ ] 4. [1 1 1 0 0], [1 1 0 0 1].

20. Какова доля правильных ответов модели, если для нее TP = 3, FP = 7, FN = 7, TN = 3? *1 из 1

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False negative (FN)	True Negative (TN)

Таблица 1: Матрица ошибок

- [x] 1. 30%
- [ ] 2. 35%
- [ ] 3. 50%
- [ ] 4. 70%