Контрольные и тестовые вопросы по ПР8 «Анализ текста (Text Mining)» по вариантам с ответами

Вариант 1

1. Какая из техник анализа текста преобразует слова к их базовой словарной форме, учитывая контекст и грамматику?
   A) Стемминг
   B) Лемматизация
   C) Токенизация
   D) Векторизация

Ответ: **B**

2. Какой из подходов векторизации текста учитывает частоту слова в конкретном документе и обратную частоту документа в корпусе?
   A) Bag of Words (BoW)
   B) Term Frequency (TF)
   C) Word2Vec
   D) TF-IDF

Ответ: **D**

3. Как называется техника, используемая для разбиения текста на отдельные слова или токены?
   A) Стемминг
   B) Токенизация
   C) Лемматизация
   D) Векторизация

Ответ: **B**

4. Какую задачу решает метод Sentiment Analysis в Text Mining?
   A) Определение грамматических ошибок в тексте
   B) Анализ эмоциональной окраски текста
   C) Подсчет количества слов
   D) Удаление ненужных символов из текста

Ответ: **B**

5. Какой алгоритм машинного обучения традиционно используется в задачах классификации текстов с учетом частоты слов и простоты реализации?
   A) Support Vector Machine (SVM)
   B) Random Forest
   C) Naïve Bayes
   D) Logistic Regression

Ответ: **C**

6. Что из перечисленного наиболее эффективно визуализирует точность классификации текстов по различным классам?
   A) Scatter Plot
   B) Confusion Matrix
   C) Histogram
   D) Word Cloud

Ответ: **B**

7. Какое преимущество имеет использование стоп-слов при предобработке текста?
   A) Ускорение работы модели
   B) Сокращение словаря и уменьшение шума в данных
   C) Улучшение грамматической структуры текста
   D) Повышение точности классификации на коротких текстах

Ответ: **B**

8. Чем отличается стемминг от лемматизации в контексте Text Mining?
   Ответ: Стемминг сокращает слова до корневой части без учета грамматики и контекста, а лемматизация приводит слова к базовой форме, учитывая грамматические особенности.

9. В каких ситуациях предпочтительнее использовать векторизацию методом Word2Vec вместо TF-IDF?
   Ответ: Word2Vec предпочтительнее при необходимости сохранения семантических и контекстных связей между словами, особенно в задачах тематического моделирования или аналогий.

10. Почему Naïve Bayes эффективен при работе с текстовыми данными?
    Ответ: Наивный Байес эффективно работает с разреженными данными

и высокими размерностями, часто возникающими в текстовых данных, благодаря предположению независимости признаков.

Вариант 2

1. В каком случае токенизация может быть недостаточной для качественной обработки текста?
   A) Если текст содержит множество стоп-слов
   B) Если необходимо сохранить контекст словосочетаний
   C) Если текст написан одним регистром
   D) Если текст содержит мало слов

Ответ: **B**

2. Какой метод предобработки текста удаляет самые распространенные служебные слова?
   A) Стемминг
   B) Удаление стоп-слов
   C) Лемматизация
   D) Векторизация

Ответ: **B**

3. Какой алгоритм классификации наиболее устойчив к высокоразмерным и разреженным текстовым данным?
   A) Decision Tree
   B) K-Nearest Neighbors
   C) Logistic Regression
   D) Support Vector Machine (SVM)

Ответ: **D**

4. Какой подход помогает избежать влияния часто встречающихся, но малозначимых слов при анализе текста?
   A) Лемматизация
   B) Стемминг
   C) Векторизация TF-IDF
   D) Bag of Words (BoW)

Ответ: **C**

5. Какая метрика оценки классификации показывает долю правильно классифицированных объектов среди всех предсказаний одного класса?

A) Accuracy
B) Precision
C) Recall
D) F1-score

Ответ: **B**

6. Какой вид визуализации позволяет наглядно сравнивать частоту использования различных слов в тексте?
   A) Scatter Plot
   B) Histogram
   C) Confusion Matrix
   D) Pie Chart

Ответ: **B**

7. Каким образом можно улучшить модель классификации текстов без изменения алгоритма машинного обучения?
   A) Использование более сложных токенизаторов
   B) Улучшение процедуры предобработки текста (например, лемматизация)
   C) Увеличение количества стоп-слов
   D) Сокращение обучающей выборки

Ответ: **B**

8. Какие проблемы возникают при использовании Bag of Words (BoW) модели в анализе текста?
   Ответ: BoW не учитывает порядок слов и семантические связи, что может приводить к потере важного контекста и снижению качества классификации.

9. Что такое тематическое моделирование (Topic Modeling), и зачем оно применяется в Text Mining?
   Ответ: Тематическое моделирование — это метод автоматического выявления абстрактных тем в наборе документов, применяется для обнаружения скрытых структур и трендов в текстовых данных.

10. Чем отличаются задачи классификации текста от задач тематического моделирования?
    Ответ: Классификация текста предполагает предварительно заданные категории и метки, тогда как тематическое моделирование выявляет темы и структуры в тексте без заранее известных категорий.

Вариант 3

1. Что такое токен в контексте анализа текста?
   A) Целое предложение
   B) Единица текста (например, слово, символ)
   C) Стоп-слово
   D) Абзац текста

Ответ: **B**

2. Какой метод анализа текста наиболее эффективен для выявления настроения и тональности текстовых сообщений?
   A) Стемминг
   B) Лемматизация
   C) Sentiment Analysis
   D) Частотный анализ слов

Ответ: **C**

3. Что позволяет сделать техника Named Entity Recognition (NER) в задачах Text Mining?
   A) Определить тональность текста
   B) Выявить именованные сущности (имена, места, организации) в тексте
   C) Привести слова к базовым формам
   D) Удалить редкие слова

Ответ: **B**

4. Какой алгоритм лучше всего подходит для работы с текстами большой размерности и необходимости выделения признаков автоматически?
   A) Naïve Bayes
   B) Neural Networks (нейронные сети)
   C) Decision Trees
   D) K-Means

Ответ: **B**

5. Какая проблема возникает при использовании метода Bag of Words (BoW) в большом наборе документов?
   A) Низкая скорость обработки
   B) Потеря семантических и контекстных связей между словами

C) Сложность реализации
D) Высокие требования к памяти

Ответ: **B**

6. Какой этап предварительной обработки текста позволяет эффективно работать с разноязычными текстами?
   A) Векторизация
   B) Лемматизация с учетом языка
   C) Токенизация по символам
   D) Частотный анализ

Ответ: **B**

7. Какой тип графика удобен для представления распределения длин документов в корпусе текстов?
   A) Line Chart
   B) Histogram
   C) Scatter Plot
   D) Word Cloud

Ответ: **B**

8. Как влияет удаление стоп-слов на качество классификации текстов?
   Ответ: Удаление стоп-слов снижает размерность данных и шум, повышая точность модели классификации за счет исключения слов без смысловой нагрузки.

9. Почему важна нормализация текста (приведение к нижнему регистру) в задачах Text Mining?
   Ответ: Нормализация позволяет объединить одинаковые слова, записанные в разных регистрах, в одну категорию, уменьшая размерность и повышая точность анализа.

10. Какова основная цель использования методов визуализации в анализе текстовых данных?
    Ответ: Визуализация помогает наглядно интерпретировать результаты анализа, выявлять закономерности, тренды и проблемы в данных, упрощая понимание сложной информации.

Вариант 4

1. В каких случаях рекомендуется использовать алгоритм Support Vector Machine (SVM) при классификации текстовых данных?
   A) Когда требуется быстрая обработка большого количества данных
   B) Когда тексты имеют линейно разделимые классы с большим количеством признаков
   C) Когда данные содержат множество неструктурированных категорий
   D) Когда тексты короткие и имеют малое число уникальных слов

Ответ: **B**

2. Что такое «разреженная матрица» (sparse matrix) в контексте обработки текстовых данных?
   A) Матрица с большим количеством стоп-слов
   B) Матрица, в которой большинство элементов равны нулю
   C) Матрица, содержащая только наиболее частые слова
   D) Матрица, состоящая исключительно из биграмм

Ответ: **B**

3. Какая техника Text Mining позволяет определить части речи слов в тексте (существительные, глаголы и т.д.)?
   A) Стемминг
   B) Токенизация
   C) Лемматизация
   D) Part-of-Speech tagging (POS-tagging)

Ответ: **D**

4. Каким образом можно бороться с проблемой дисбаланса классов при анализе текстовых данных?
   A) Увеличением размера словаря
   B) Использованием техники SMOTE или другими методами балансировки данных
   C) Увеличением количества стоп-слов
   D) Использованием токенизации по символам

Ответ: **B**

5. Чем полезна метрика F1-score при оценке качества модели классификации текстов?

A) Она показывает точность классификации только для позитивного класса
B) Она учитывает как точность (precision), так и полноту (recall), особенно при дисбалансе классов
C) Она оценивает только полноту классификации
D) Она измеряет скорость работы модели

Ответ: **B**

6. Какой подход Text Mining позволяет автоматически группировать документы без заранее заданных категорий?
   A) Классификация
   B) Кластеризация
   C) Лемматизация
   D) Векторизация

Ответ: **B**

7. Какой алгоритм машинного обучения наиболее эффективен при обработке коротких текстов (например, твитов)?
   A) Decision Trees
   B) Random Forest
   C) Naïve Bayes с использованием сглаживания
   D) K-Nearest Neighbors

Ответ: **C**

8. Какую роль играет этап преобразования текста в числовой вектор (векторизация) в задачах анализа текста?
   Ответ: Векторизация преобразует текстовые данные в числовое представление, позволяющее применять алгоритмы машинного обучения для классификации, кластеризации и других аналитических задач.

9. Почему важно проводить этап токенизации в начале анализа текста?
   Ответ: Токенизация позволяет разделить текст на отдельные смысловые единицы, облегчая последующие этапы анализа, такие как удаление стоп-слов, стемминг и векторизацию.

10. Что такое эмбеддинги слов (word embeddings), и какое преимущество они имеют перед методом Bag of Words?
    Ответ: Эмбеддинги слов — это представления слов в виде плотных векторов, которые сохраняют семантические и контекстные связи

между словами, в отличие от метода Bag of Words, где контекст не учитывается.