Контрольные и тестовые вопросы по ПР9

«Web Scraping и анализ данных в RapidMiner» по вариантам с ответами

Вариант 1

1. Какой оператор RapidMiner предназначен для извлечения данных из RSS-лент веб-сайтов?
   A) Read Database
   B) Read CSV
   C) Read RSS Feed
   D) Retrieve Data

Ответ: C

2. Зачем при обработке текста новостей используют токенизацию (tokenization)?
   A) Для удаления HTML-тегов из текста
   B) Для разделения текста на отдельные слова и фразы
   C) Для приведения текста к нижнему регистру
   D) Для расчета общей длины текста

Ответ: B

3. Какую роль играет оператор Filter Stopwords при анализе текста?
   A) Удаляет редко встречающиеся слова
   B) Приводит слова к нижнему регистру
   C) Удаляет слова, не несущие смысловой нагрузки
   D) Создает комбинации из слов (n-граммы)

Ответ: C

4. Какую задачу решает создание n-грамм (например, биграмм) при текстовом анализе?
   A) Уменьшает количество уникальных токенов
   B) Обеспечивает сохранение контекста словосочетаний
   C) Ускоряет процесс анализа текста
   D) Позволяет исключить редкие слова

Ответ: B

5. Что позволяет выявить частотный анализ текста при анализе новостных данных?
   A) Географическое расположение новостей
   B) Время публикации каждой новости
   C) Наиболее обсуждаемые темы и ключевые слова
   D) Авторов новостных статей

Ответ: C

6. Для каких целей чаще всего используется визуализация в виде облака слов (Word Cloud)?
   A) Для выявления частоты встречаемости слов в текстах
   B) Для демонстрации временных трендов новостей
   C) Для визуализации географического распространения новостей
   D) Для показа точного количества статей по категориям

Ответ: A

7. Какая визуализация наиболее удобна для анализа процентного распределения категорий новостей?
   A) Line Chart
   B) Bar Chart
   C) Scatter Plot
   D) Pie Chart

Ответ: D

8. Почему стоп-слова обычно удаляются из текстов перед частотным анализом?
   Ответ: Стоп-слова (предлоги, союзы, частицы) не несут смысловой нагрузки и могут искажать результаты анализа, занимая высокие позиции в частотном списке.

9. Какую информацию можно получить при помощи визуализации с помощью диаграммы рассеяния (Scatter Plot) для биграмм?
   Ответ: Можно выявить связи между парами слов и категориями новостей, показывая, в каких тематических областях чаще всего встречаются конкретные словосочетания.

10. Какие преимущества предоставляет web scraping с использованием RapidMiner в сравнении с ручным сбором данных?
    Ответ: Автоматизация процесса, возможность быстрого обновления

данных, сокращение времени на обработку и возможность одновременного анализа больших объемов информации.

Вариант 2

1. Какой тип данных обычно извлекается при использовании оператора Read RSS Feed в RapidMiner?
   A) Числовые значения
   B) Графические изображения
   C) Структурированные текстовые данные
   D) Видеоматериалы

Ответ: **C**

2. Какая функция оператора Transform Cases важна для анализа текста новостей?
   A) Удаление стоп-слов
   B) Приведение текста к единому регистру
   C) Генерация n-грамм
   D) Выделение имен собственных

Ответ: **B**

3. Зачем при анализе текста применяют фильтрацию токенов по длине (Filter Tokens by Length)?
   A) Для удаления слишком частых слов
   B) Для удаления коротких токенов, не несущих полезной информации
   C) Для сокращения объема текста
   D) Для повышения скорости обработки текста

Ответ: **B**

4. Какой оператор необходим для преобразования текста в частотный список слов в RapidMiner?
   A) Read RSS Feed
   B) Tokenize
   C) Process Documents from Data
   D) Aggregate

Ответ: **C**

5. Для каких задач при веб-скрейпинге используют регулярные выражения (RegEx)?

A) Очистка текста от стоп-слов
B) Извлечение конкретных данных по шаблону из текста
C) Создание n-грамм
D) Сортировка новостей по дате

Ответ: **B**

6. Какая визуализация наиболее информативна для оценки трендов по частоте упоминаний определенных слов со временем?
   A) Scatter Plot
   B) Pie Chart
   C) Line Chart
   D) Word Cloud

Ответ: **C**

7. В чем главное преимущество автоматизированного веб-скрейпинга по сравнению с ручным мониторингом сайтов?
   A) Более качественное извлечение информации
   B) Возможность анализа текстов на разных языках одновременно
   C) Автоматизация и масштабирование процесса получения данных
   D) Сохранение авторских прав на контент

Ответ: **C**

8. Почему RSS-ленты являются удобным источником данных для веб-скрейпинга?
   Ответ: RSS-ленты представляют информацию в стандартизированном структурированном формате (XML), что упрощает автоматическое извлечение данных и их последующую обработку.

9. Какую роль играет токенизация в процессе анализа текста?
   Ответ: Токенизация преобразует сплошной текст в отдельные единицы (слова или фразы), что позволяет проводить дальнейший анализ, такой как частотный анализ или тематическое моделирование.

10. Какие данные нельзя получить напрямую из RSS-лент новостных сайтов?
    Ответ: Полные тексты статей, комментарии пользователей, мультимедийный контент (изображения и видео), которые обычно требуют дополнительного перехода по ссылке на сам сайт.

Вариант 3

1. Какой формат данных чаще всего используется в RSS-лентах?
   A) JSON
   B) XML
   C) CSV
   D) HTML

Ответ: **B**

2. Какой этап предварительной обработки текста включает удаление HTML-разметки и лишних символов?
   A) Tokenization
   B) Stemming
   C) Cleaning
   D) Lemmatization

Ответ: **C**

3. Какой способ обработки текста позволяет выявить словосочетания и устойчивые выражения в тексте?
   A) Stemming
   B) Tokenization
   C) Generate n-Grams
   D) Transform Cases

Ответ: **C**

4. В каком случае целесообразно применять облако слов (Word Cloud)?
   A) Для выявления динамики изменения частоты слов
   B) Для визуального представления наиболее частых слов
   C) Для расчета средних значений частоты слов
   D) Для анализа длины статей

Ответ: **B**

5. Какой подход используется для выявления трендов и популярных тем в текстовых данных?
   A) Регрессионный анализ
   B) Частотный анализ слов
   C) Кластерный анализ
   D) Корреляционный анализ

Ответ: **B**

6. Что такое стоп-слова в контексте текстового анализа?
   A) Слова, которые встречаются слишком редко
   B) Слова с высокой эмоциональной окраской
   C) Служебные слова, не несущие смысловой нагрузки
   D) Технические термины и аббревиатуры

Ответ: **C**

7. Что позволяет сделать анализ с использованием биграмм, чего невозможно добиться с помощью одиночных токенов?
   A) Уменьшить количество уникальных токенов
   B) Определить контекст употребления слов
   C) Ускорить процесс анализа
   D) Исключить служебные слова

Ответ: **B**

8. Почему при анализе текста новостей важно учитывать дату публикации?
   Ответ: Анализ по дате позволяет выявить временные тренды, сезонность и динамику популярности определенных тем или слов.

9. Какие ограничения могут возникнуть при автоматическом веб-скрейпинге новостных сайтов?
   Ответ: Юридические ограничения (авторские права), технические барьеры (CAPTCHA, блокировки), частые изменения структуры сайтов.

10. Какие основные шаги включает полный цикл веб-скрейпинга и анализа текстовых данных в RapidMiner?
    Ответ: Извлечение данных (web scraping), предварительная обработка текста (очистка, токенизация, удаление стоп-слов), частотный анализ слов и визуализация результатов.

Вариант 4

1. Для чего при анализе текстовых данных может использоваться метод TF-IDF?
   A) Для подсчета общего количества слов в тексте
   B) Для измерения важности слова в контексте конкретного документа

C) Для удаления коротких слов из текста
D) Для преобразования текста в нижний регистр

Ответ: **B**

2. Какой оператор в RapidMiner подходит для очистки текста от специфических паттернов с помощью регулярных выражений (Regex)?
   A) Replace Tokens
   B) Filter Tokens by POS
   C) Replace (Dictionary)
   D) Filter Tokens (by Content)

Ответ: **A**

3. Какой тип визуализации наиболее эффективно показывает зависимость частоты слова от времени публикации?
   A) Heatmap
   B) Scatter Plot
   C) Line Chart
   D) Bar Chart

Ответ: **C**

4. Какая техника текстового анализа позволяет свести разные грамматические формы слова к базовой?
   A) Tokenization
   B) Lemmatization
   C) n-Gram generation
   D) Transform Cases

Ответ: **B**

5. Какую проблему помогает решить приведение текста к нижнему регистру (lowercase transformation)?
   A) Исключение редких слов
   B) Исключение служебных слов
   C) Объединение одинаковых слов в разных регистрах
   D) Повышение скорости обработки текста

Ответ: **C**

6. Какой метод анализа текстов позволяет автоматически определить темы документов без предварительного задания ключевых слов?
   A) TF-IDF

B) Sentiment Analysis
C) Topic Modeling (тематическое моделирование)
D) Frequency Analysis

Ответ: **C**

7. Чем лемматизация (lemmatization) отличается от стемминга (stemming) при обработке текста?
   A) Лемматизация всегда быстрее стемминга
   B) Лемматизация учитывает грамматическое значение слова, а стемминг – нет
   C) Стемминг используется только для русского языка
   D) Стемминг позволяет точнее определить контекст слова

Ответ: **B**

8. Почему в текстовом анализе часто удаляют редко встречающиеся слова (редкие токены)?
   Ответ: Редкие токены часто не несут статистически значимой информации и могут усложнять анализ, увеличивая размерность данных и создавая шум.

9. В каких случаях предпочтительнее использовать n-граммы вместо отдельных слов при анализе текстов?
   Ответ: N-граммы предпочтительны, когда необходимо сохранить контекст и смысловые связи между словами, особенно в анализе мнений и тематическом моделировании.

10. Что такое sentiment analysis и для чего он применяется в текстовом анализе?
    Ответ: Sentiment analysis – это метод анализа текста для определения эмоциональной окраски (позитивной, негативной, нейтральной), применяется для изучения общественного мнения, анализа отзывов и оценки восприятия брендов или событий.