

Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Филологический факультет

УТВЕРЖДЕНО:
Декан филологического факультета
И.В. Тубалова

Оценочные материалы по дисциплине

Язык программирования R

по направлению подготовки

45.03.03 Фундаментальная и прикладная лингвистика

Направленность (профиль) подготовки:
«Фундаментальная и прикладная лингвистика»

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
А.В. Васильева

Председатель УМК
Ю.А. Тихомирова

Томск – 2024

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

автоматической обработки текстов на естественном языке (ЕЯ)

Целью освоения дисциплины является формирование следующих компетенций:

ПК-4. Способен разрабатывать программный код при решении задач автоматической обработки текстов

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.2. Создает программный код с использованием языков программирования и манипулирования данными в сфере автоматической обработки текстов

2. Оценочные материалы текущего контроля и критерии оценивания

К числу форм контроля, оценивающих уровень достижения компетенций по текущей дисциплине, относятся тесты, устные вопросы, выполнение практических заданий (ПК-4, ПК-4.2), итоговый проект (продукт (ПК-4), код (ИПК-4.2), презентация), контрольная и самостоятельная работы (ПК-4, ПК-4.2).

Итоговый проект включает блок вопросов по теоретической части изучаемой дисциплины (ПК-4, ПК-4.2):

Пример практических задач:

1. Реализуйте код регуляризации
2. Выполните кросс-валидацию модели
3. Создайте алгоритм лемматизации
4. Выполнить анализ количественный анализ лексем
5. Примените методы машинного обучения
6. Выполните детектирование выбросов в R
7. Реализуйте метод главных компонент (PCA)
8. Выполните кластеризацию данных с помощью K-средних в R
9. Создайте алгоритм случайного леса (Random Forests) и примените его на

текстах

10. Выполните нормализацию данных в R?

Ответы:

Теоретические вопросы (ПК-4, ПК-4.2):

Критерии оценивания ответа на теоретический вопрос

Оценка	Критерии
	1. Понимание и логика высказывания изученного материала 2. Представление взаимосвязей процесса и взаимосвязи теоретических модулей изучаемого предмета 3. Полнота данных ответов; 4. Аргументированность данных ответов; 5. Правильность ответов на вопросы;

«зачтено»	Полно и аргументировано даны ответы по содержанию задания. Обнаружено понимание материала, может обосновать свои суждения, применить знания на практике, привести необходимые примеры не только по учебнику, но и самостоятельно составленные. Изложение материала последовательно и правильно Ответы обучающегося удовлетворяют тем же требованиям, что и для оценки «отлично», но допускается 1-2 ошибки, которые сам же исправляет. Обучающийся демонстрирует знание и понимание основных положений данного задания, но: 1) излагает материал неполно и допускает неточности в определении понятий или формулировке правил; 2) не умеет достаточно глубоко и доказательно обосновать свои суждения и привести свои примеры; 3) излагает материал непоследовательно и допускает ошибки.
«не зачтено»	Демонстрация незнания ответа на соответствующее задание, допускаются ошибки в формулировке определений и правил, искажающие их смысл, беспорядочно и неуверенно излагается материал; отмечаются такие недостатки в подготовке, которые являются серьезным препятствием к успешному овладению последующим материалом.

Критерии оценивания практической работы:

Оценка	Критерии
	1. Понимание и логика алгоритма работы 2. Наличие или отсудив ошибок в коде 3. Полнота решения практических задач 4. Своевременность выполнения; 5. Умения связать практический материал с теоретическим; 6. Понимание базовых формул обработки естественного языка и программирования;
«зачтено»	Основные требования к решению практических задач выполнены. Продемонстрированы умение анализировать алгоритмы и находить оптимальное количество решений, умение работать с информацией, в том числе умение затребовать дополнительную информацию, необходимую для уточнения реализации алгоритма, навыки разработки программного кода; Основные требования к решению практических задач выполнены, но при этом допущены недочеты. В частности, недостаточно раскрыты навыки стиля, недостаточно комментариев Имеются существенные отступления от решения практических задач. В частности отсутствуют навык и умения моделировать решения в соответствии с заданием, представлять различные подходы к разработке алгоритмов, ориентированных на конечный результат
«не зачтено»	Задача не решена, обнаруживается существенное непонимание проблемы

Итоговый проект включает возможность выбора типа проекта: звучащая или письменная речь. В случае выбора звучащей речи, необходимо записать звуковой файл и осуществить частотный анализ на основе спектрограммы и осциллограммы. Доказать поставленную гипотезу методами АТР. Аргументировать ответ и сделать презентацию. Если выбран трек по работе с письменной речью, то необходимо осуществить следующие действия с текстом: лемматизация и токенизация; б) осуществить предобработку; в) применить один из методов векторизации

Выбор тем текстов и звучащих не ограничен и предусматривает возможность выбора индивидуальной образовательной траектории, связывающую проектную деятельность со смежными гуманитарными дисциплинами. Однако существуют ограничительные критерии реализации проекта, направленные на формирование вышеуказанных компетенций: 1. Сбор и структуризация данных. 2. Наличие двух классов в обучающем датасете. 3. Объем материала не менее 50 тыс текстов на каждый класс. 4. Наличие формальных метрик оценивания качества результата машинного обучения

Сбор и структуризация данных предусматривает свободу выбора темы. В случае, если выбор не был осуществлен, то предлагаются следующие рубрики:

1. Новостные сайты и пресс релизы (классификация и генерация текстов по рубрике и/или информационному агентству)
2. Отзывы к товарам или фильмам (Наличие 3 класса: позитивный, негативный, нейтральный отзывы)
3. Комментарии в социальных сетях
4. Стены сообществ «ВКонтакте» (классификация по рубриками)
5. Звучащая речь студентов, изучающих русский язык
6. Родная речь иностранцев с целью ее сравнения со своей

Критерии оценивания практической работы:

Оценка	Критерии
	1. Сформулирована гипотеза проекта 2. Собран и структурирован датасет 3. Четкая логика реализации алгоритмов обработки естественного языка в коде; 4. Правильность ответов на вопросы; 5. Наличие структурированной презентации 6. Полнота проекта

«зачтено»	Выполнены все требования к проекту: сформулирована гипотеза, создан прасер сайтов для сбора и структуризации информации, написан код для обработки естественного языка Выполнены все требования к составлению презентаций: дизайн слайдов, логика изложения материала, текст хорошо написан и сформированные идеи ясно изложены и структурированы Существуют незначительные ошибки в проекте, не влияющие на конечный результат. В частности, может быть низкий уровень формальных метрик, неточности в визуализации данных Основные требования к презентациям выполнены, но при этом допущены недочеты. В частности, имеются неточности в изложении материала; отсутствует логическая последовательность в суждениях; не выдержан объем презентации Существенные ошибки в коде, отсекает логика представления проекта, недостаточный объем датасета Имеются существенные отступления от требований к презентациям. В частности: тема освещена лишь частично; допущены фактические ошибки в содержании презентаций или при ответе на дополнительные вопросы.
«не зачтено»	Критические ошибки в коде, гипотеза не подвержена, недостаточное количество или отсутствие обучающей выборки Тема презентации не раскрыта, обнаруживается существенное непонимание проблемы

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Зачет в первом семестре состоит из трех частей.

Первая часть представляет собой тест из 20 вопросов, проверяющих ПК-4, ПК-4.2

Ответы на вопросы первой части даются путем выбора из списка предложенных, установления соответствия между объектами или формулирования однословного ответа на вопрос открытого типа.

Примерный тест

1. Что такое R?

- а) Язык программирования для разработки веб-приложений
- б) Графическая библиотека для создания визуализаций
- в) Инструмент для анализа данных и статистики
- г) Фреймворк для разработки мобильных приложений
- **Ответ: в) Инструмент для анализа данных и статистики**

2. Какой символ используется для присваивания значения переменной в R?

- а) =
- б) :
- в) ->
- г) +

- **Ответ: a) ==**

3. Каким образом можно создать вектор в R?

- a) С помощью функции `matrix()`
 - b) С помощью функции `list()`
 - c) С помощью функции `vector()`
 - d) С помощью функции `data.frame()`
- **Ответ: c) С помощью функции `vector()`**

4. Какие основные типы данных поддерживает R?

- a) Числа, строки, булевы значения, списки
 - b) Числа, строки, даты, графы
 - c) Числа, векторы, матрицы, факторы
 - d) Числа, деревья, графы, функции
- **Ответ: a) Числа, строки, булевы значения, списки**

5. Как получить длину вектора в R?

- a) С помощью функции `length()`
 - b) С помощью функции `size()`
 - c) С помощью функции `count()`
 - d) С помощью функции `dim()`
- **Ответ: a) С помощью функции `length()`**

6. Какой оператор используется для объединения двух векторов в R?

- a) `+`
 - b) `-`
 - c) `*`
 - d) `/`
- **Ответ: a) `+`**

7. Что такое факторы в R?

- a) Операторы для выполнения математических операций
 - b) Структура данных для хранения категориальных переменных
 - c) Функции для работы с текстовыми данными
 - d) Специальные символы для работы с регулярными выражениями
- **Ответ: b) Структура данных для хранения категориальных переменных**

8. Как создать матрицу в R?

- a) С помощью функции `array()`
 - b) С помощью функции `data.frame()`
 - c) С помощью функции `list()`
 - d) С помощью функции `matrix()`
- **Ответ: d) С помощью функции `matrix()`**

9. Что такое список (`list`) в R?

- a) Упорядоченная коллекция элементов одного типа
 - b) Упорядоченная коллекция элементов разных типов
 - c) Неупорядоченная коллекция элементов одного типа
 - d) Неупорядоченная коллекция элементов разных типов
- **Ответ: b) Упорядоченная коллекция элементов разных типов**

10. Как создать функцию в R?

- a) С помощью ключевого слова function()
- b) С помощью ключевого слова def()
- c) С помощью ключевого слова create()
- d) С помощью ключевого слова new()
- **Ответ: a) С помощью ключевого слова function()**

Критерии оценивания тестирования:

Оценка	Критерии
	1. Полнота выполнения тестовых заданий; 2. Своевременность выполнения; 3. Правильность ответов на вопросы; 4. Самостоятельность тестирования
«зачтено»	Выполнено более 85 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос Выполнено более 70 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос; однако были допущены неточности в определении понятий, терминов и др. Выполнено более 54 % заданий предложенного теста, в заданиях открытого типа дан неполный ответ на поставленный вопрос, в ответе не присутствуют доказательные примеры, текст со стилистическими и орфографическими ошибками.
«не зачтено»	Выполнено не более 53 % заданий предложенного теста, на поставленные вопросы ответ отсутствует или неполный, допущены существенные ошибки в теоретическом материале (терминах, понятиях).

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

Первая часть. Решение тестовых задач:

Тест №1

1. Какой оператор используется для выполнения условного оператора в R?
 - a) if
 - b) for
 - c) while
 - d) switch
 - **Ответ: a) if**
2. Что такое цикл for в R и как его использовать?
 - a) Цикл для выполнения заданного количества итераций
 - b) Цикл для выполнения операций до выполнения условия
 - c) Цикл для выполнения операций некоторое количество раз
 - d) Цикл для выполнения операций в зависимости от условия
 - **Ответ: a) Цикл для выполнения за данного количества итераций**
3. Как получить справку о функции в R?
 - a) С помощью функции help()
 - b) С помощью функции info()

- c) С помощью функции doc()
- d) С помощью функции man()
- **Ответ: a) С помощью функции help()**

4. Каким образом можно выполнить чтение данных из файла в R?

- a) С помощью функции read.csv()
- b) С помощью функции read.table()
- c) С помощью функции read.json()
- d) С помощью функции read.xml()
- **Ответ: a) С помощью функции read.csv()**

5. Как сохранить результат работы в R в файл?

- a) С помощью функции write.csv()
- b) С помощью функции write.table()
- c) С помощью функции write.json()
- d) С помощью функции write.xml()
- **Ответ: a) С помощью функции write.csv()**

6. Как создать график в R?

- a) С помощью функции plot()
- b) С помощью функции draw()
- c) С помощью функции chart()
- d) С помощью функции visualize()
- **Ответ: a) С помощью функции plot()**

7. Как построить гистограмму в R?

- a) С помощью функции hist()
- b) С помощью функции barplot()
- c) С помощью функции boxplot()
- d) С помощью функции pie()
- **Ответ: a) С помощью функции hist()**

8. Что такое пакеты в R и как их установить?

- a) Библиотеки с дополнительными функциями и возможностями
- b) Модули для подключения к базам данных
- c) Инструменты для работы с графическим интерфейсом
- d) Специальные средства для отладки программ
- **Ответ: a) Библиотеки с дополнительными функциями и возможностями**

9. Какой оператор используется для выполнения логического И в R?

- a) &&
- b) ||
- c) !
- d) &
- **Ответ: d) &**

10. Какие встроенные статистические функции доступны в R?

- a) mean(), median(), sd(), var()
- b) sum(), count(), min(), max()
- c) sort(), unique(), subset(), merge()
- d) plot(), hist(), boxplot(), barplot()
- **Ответ: a) mean(), median(), sd(), var()**

11. Что такое линейная регрессия в R?
- a) Метод для классификации данных
 - b) Метод для кластеризации данных
 - c) Метод для прогнозирования значений на основе линейной зависимости
 - d) Метод для анализа временных рядов
 - **Ответ: c) Метод для прогнозирования значений на основе линейной зависимости**
12. Как вычислить среднее значение вектора в R?
- a) С помощью функции `average()`
 - b) С помощью функции `mean()`
 - c) С помощью функции `median()`
 - d) С помощью функции `mode()`
 - **Ответ: b) С помощью функции `mean()`**
13. Какие методы доступны для работы с датами и временем в R?
- a) `as.Date()`, `as.POSIXlt()`, `strptime()`
 - b) `date()`, `time()`, `datetime()`
 - c) `parse()`, `format()`, `difftime()`
 - d) `year()`, `month()`, `day()`
 - **Ответ: a) `as.Date()`, `as.POSIXlt()`, `strptime()`**
14. Что такое пакет `dplyr` в R и для чего он используется?
- a) Пакет для работы с графическим интерфейсом
 - b) Пакет для работы с базами данных
 - c) Пакет для выполнения операций над данными (фильтрация, сортировка, группировка и т.д.)
 - d) Пакет для создания интерактивных диаграмм
 - **Ответ: c) Пакет для выполнения операций над данными (фильтрация, сортировка, группировка и т.д.)**

Вторая часть. Решение практических задач:

Задача 1. Напишите программу, которая подсчитывает количество слов в текстовом файле.

Задача 2. Реализуйте функцию для удаления стоп-слов из текста.

Задача 3. Создайте программу для лемматизации текста с использованием библиотеки `quanteda`.

Задача 4. Напишите скрипт, который выделяет ключевые слова из текста с помощью алгоритма TF-IDF.

Задача 5. Разработайте программу для определения тональности текста с использованием библиотеки `sentimentr`.

Задача 6. Реализуйте алгоритм поиска именованных сущностей в тексте с помощью пакета `openNLP`.

Задача 7. Напишите программу для определения частей речи в предложении с использованием библиотеки `udpipe`.

Задача 8. Разработайте скрипт для сегментации текста на предложения с использованием пакета `openNLP`.

Задача 9. Создайте программу для автоматической генерации краткого описания текста с использованием алгоритма `TextRank`.

Задача 10. Реализуйте алгоритм классификации текста на основе наивного Байесовского классификатора с помощью пакета `e1071`.

Информация о разработчиках

Степаненко Андрей Александрович, старший преподаватель кафедры общей, компьютерной и когнитивной лингвистики