

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

Информатика и основы программирования

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

Целью освоения дисциплины является формирование следующих компетенций:

– ПК-4 – разрабатывать программный код при решении задач автоматической обработки текстов;

– ОПК-7 – Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.1 Применяет способы формализации и алгоритмизации поставленных задач в сфере автоматической обработки текстов.

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

ИОПК-7.3 Применяет современные технические средства и информационно-коммуникационные технологии для решения задач профессиональной деятельности.

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 (ОПК-7), итоговый проект (продукт (ИПК-4.1), код (ПК-4), презентация), контрольная и самостоятельная работы (ИПК-4.2, ИОПК-7.3).

Итоговый проект включает блок вопросов по теоретической части изучаемой дисциплины (ОПК-7):

Какие основные компоненты включает в себя обработка естественного языка (Natural Language Processing, NLP)?

Каким образом компьютеры обрабатывают и интерпретируют естественный язык?

Какие алгоритмы используются в обработке естественного языка для выполнения различных задач, таких как разбор предложений или определение тональности текста?

Какие языки программирования и инструменты часто применяются в области обработки естественного языка?

Какие методы и техники применяются для предварительной обработки текстовых данных перед их анализом на естественном языке?

Каким образом машины могут выполнять задачи, связанные с пониманием и генерацией текста на естественном языке?

Какие проблемы и вызовы возникают при обработке естественного языка, и как они могут быть решены?

Как обработка естественного языка применяется в различных областях, таких как машинный перевод, анализ социальных медиа или автоматическое рецензирование текста?

Пример практических задач:

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

2. Реализуйте функцию на R, которая принимает на вход строку и возвращает ее длину в символах.

3. Создайте скрипт на R, который считывает текстовый файл, преобразует все буквы в нижний регистр и сохраняет результат в новом файле.

4. Напишите программу на R, которая анализирует текст и выводит наиболее часто встречающиеся слова в порядке убывания их частоты.

5. Реализуйте функцию на R, которая принимает на вход две строки (текст) и возвращает значение, показывающее степень их сходства (например, коэффициент Жаккара или косинусное сходство).

6. Создайте скрипт на R, который извлекает именованные сущности (например, имена людей, организации) из текста и выводит их список.

7. Напишите программу на R, которая преобразует текст в числовую матрицу, где каждая строка представляет собой вектор количества вхождений определенного слова.

8. Реализуйте функцию на R, которая принимает на вход текст и возвращает список наиболее значимых ключевых слов или фраз, используя алгоритм TF-IDF.

9. Создайте скрипт на R, который выполняет лемматизацию текста с использованием библиотеки `quanteda` или ПО `mystem`.

10. Напишите программу на R, которая проводит sentiment-анализ текста, определяя его тональность (положительную, отрицательную или нейтральную) с использованием словарей или машинного обучения.

Ответы:

Теоретические вопросы (ОПК-7):

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;
«отлично»	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
«хорошо»	Ответы обучающегося удовлетворяют тем же требованиям, что и для оценки «отлично», но допускается 1-2 ошибки, которые сам же исправляет.
«удовлетворительно»	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.
«неудовлетворительно»	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;

«отлично»	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;
«хорошо»	Основные требования к решению практических задач выполнены, но при этом допущены недочеты. В частности, недостаточно раскрыты навыки стиля, недостаточно комментариев
«удовлетворительно»	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
«неудовлетворительно»	Задача не решена, обнаруживается существенное непонимание проблемы

Итоговый проект (ИПК-4.1, (ПК-4, ИПК-4.2, ИОПК-7.3)

Итоговый проект включает разработку кода, демонстрацию и применение алгоритмов машинного обучения и искусственного интеллекта в задачах обработки естественного языка, презентацию и устный доклад относительно выбранной темы.

Выбор тем 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)

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
«отлично»	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
«хорошо»	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
«удовлетворительно»	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.
«неудовлетворительно»	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 во втором семестре состоит из трех частей.

Первая часть представляет собой тест из 20 вопросов, проверяющих ОПК-7, ПК-4, ИПК-4.1.

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

- Что такое информатика?
 - А) Наука об информации
 - Б) Программирование
 - В) Математическая дисциплина
 - Г) Язык программирования
 - Д) + Наука о обработке информации
- Какие компоненты включает в себя компьютерная система?
 - А) Процессор, клавиатура, мышь
 - Б) + Аппаратное и программное обеспечение
 - В) Монитор, принтер, сканер
 - Г) Клавиатура, оперативная память, жесткий диск
 - Д) Программы, файлы, папки

3. Что такое алгоритм?
- А) Портативное устройство
 - Б) + Последовательность действий для решения задачи
 - В) Язык программирования
 - Г) Система хранения данных
 - Д) Графический интерфейс пользователя
4. Какие языки программирования относятся к высокоуровневым?
- А) + Python, Java, C++
 - Б) HTML, CSS, JavaScript
 - В) Assembler, машинные языки
 - Г) SQL, C#, PHP
 - Д) MATLAB, R, Perl
5. Что такое переменная в программировании?
- А) Название программы
 - Б) + Область памяти для хранения данных
 - В) Модель компьютера
 - Г) Время выполнения программы
 - Д) Список команд
6. Что означает оператор IF в программировании?
- А) + Условный оператор
 - Б) Циклический оператор
 - В) Арифметический оператор
 - Г) Логический оператор
 - Д) Бинарный оператор
7. Какое предназначение у циклов в программировании?
- А) + Повторение выполнения блока кода
 - Б) Выбор ветвления программы
 - В) Выполнение арифметических операций
 - Г) Взаимодействие с пользователем
 - Д) Обработка текстовых данных
8. Что такое функция в программировании?
- А) + Блок кода, выполняющий определенную задачу
 - Б) Числовое выражение
 - В) Набор символов для обозначения операций
 - Г) Графический элемент интерфейса
 - Д) Командная строка
9. Какие типы данных используются в программировании?
- А) + Целочисленные, вещественные, символьные
 - Б) Булевский, строковый, пустой
 - В) Файловый, директорий, папка
 - Г) HTML, CSS, JavaScript
 - Д) SQL, C#, PHP
10. Что такое массив в программировании?
- А) + Упорядоченная коллекция элементов одного типа

- Б) Специальный тип данных для хранения текстовой информации
 - В) Графический элемент интерфейса
 - Г) Числовой идентификатор
 - Д) Команда для выполнения определенного действия
11. Каким образом происходит взаимодействие программы с пользователем?
- А) + Через ввод и вывод данных
 - Б) Путем выполнения арифметических операций
 - В) Посредством создания графического интерфейса
 - Г) При помощи сетевого подключения
 - Д) Через обмен сообщениями
12. Что такое алгоритмическая сложность?
- А) + Мера количества ресурсов, требуемых для выполнения алгоритма
 - Б) Способность программы противостоять взлому
 - В) Точность расчетов программы
 - Г) Способность программы работать с большим объемом данных
 - Д) Степень вложенности циклов в программе
13. Что такое рекурсия в программировании?
- А) + Вызов функцией самой себя
 - Б) Условное выполнение кода
 - В) Обработка исключительных ситуаций
 - Г) Взаимодействие с файловой системой
 - Д) Преобразование данных из одного типа в другой
14. Что такое объектно-ориентированное программирование?
- А) + Парадигма программирования, основанная на использовании объектов и классов
 - Б) Метод разработки программного обеспечения
 - В) Язык программирования для создания веб-сайтов
 - Г) Система хранения и организации данных
 - Д) Операционная система для мобильных устройств
15. Какая функция выполняется базовой единицей данных в реляционных базах данных?
- А) + Запись (строка таблицы)
 - Б) Таблица
 - В) Поле (столбец таблицы)
 - Г) База данных
 - Д) Индекс

Критерии оценивания тестирования:

Оценка	Критерии
	1. Полнота выполнения тестовых заданий; 2. Своевременность выполнения; 3. Правильность ответов на вопросы; 4. Самостоятельность тестирования
«отлично»	Выполнено более 85 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос
«хорошо»	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.
«удовлетворительно»	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.
«неудовлетворительно»	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

- Языки программирования применяются в филологических исследованиях в качестве
 - формальных
 - интроспективных
 - качественных
 - количественных
 - все из перечисленных
- К направлению филологических исследований, реализуемых на стыке литературоведения и информатики относится
 - Стилистика
 - Стилеметрия
 - Имагология
 - Фольклористика
- К областям применения приемов сентимент-анализа относятся
 - Морфологическая структура текста
 - Вариантов выражения общественного мнения в тексте
 - Своеобразие выражения эмоционально-оценочного фона художественного текста
 - Интонационно-профильный профиль текста
 - Все из перечисленных
- К этапам междисциплинарных экспериментальных исследований в лингвистике, в процедуре которых используется программирование относятся
 - Формулирование гипотезы
 - Отбор стимульного материала
 - Планирование дизайна эксперимента
 - Конструирование эксперимента в программной среде
 - Статистический анализ полученных данных
 - Интерпретация полученных результатов

5. К этапам создания лингвистически размеченных корпусов текстов, в процедуре которых используется программирование относятся

- А) Формальная предобработка данных
- Б) Создание пользовательского интерфейса
- В) Разработка системы разметки в корпусе
- Г) Разработка системы метаразметки в корпусе

6. Выберите утверждение, соответствующее языкам программирования:

- А) объектно-ориентированные
- Б) объектно-искусственные
- В) объектно-направленные
- Г) все вышеперечисленное

7. Выберите подходящее определение, характерное для термина язык программирования

- А) совокупность средств и правил перевода текста с естественного языка на формальный.
- Б) совокупность средств и правил перевода текста с формального языка на естественный.
- В) совокупность средств и правил представления алгоритма в виде, пригодном для выполнения вычислительной машиной.
- Г) совокупность систем формальных правил в двоичном коде, понятном для компилятора.

8. Какие языки программирования относятся к интерпретируемым?

- А) C++ и Pascal.
- Б) R и Python.
- В) C++ и Python
- Г) все вышеперечисленное.

9. Какие задачи решает компьютерная лингвистика (NLP)

- А) генерация текста.
- Б) разработка чат ботов.
- В) распознавание речи
- Г) все вышеперечисленное.

10. Какие языки программирования относятся к компилируемым?

- А) C++ и Pascal.
- Б) R и Python.
- В) C++ и Python
- Г) все вышеперечисленное.

Вторая часть. Решение практических задач

Примерные задания по 5 модулю:

Изучите код, инициализирующий лемматизацию массива текстов. Примените код на своих текстах, которые были скачаны в прошлом семестре

```
data_path = "D:/stem"
list.files(path = "D:/stem", "txt")
# all subjects
#subject_id = c(c(1:3))
subject_id = list.files(path = "D:/stem", "txt")
# stem_id = c(c(1:3))
# write_id = c(c(1:3))
#filename = paste0(data_path, "/", subject_id[2])
for ( id in subject_id ) {
#filename = paste0(data_path, "/dlg_m_", id, ".txt")
}
```

```

filename = paste0(data_path, "/", id)
print(filename)
text <- read.table(filename, header = FALSE, sep = "\t", encoding =
"russian",quote="","",comment=")
ldf <- lapply(as.matrix(text$V1), function(x)x)
text2 <- as.character(ldf)
#text <- as.character(text$V1)
text3 <- enc2utf8(text2)
cmd <- "D:/stem/mystem.exe -cnild -e UTF-8"
#if (!is.null(fixlist) && file.exists(fixlist))
# cmd <- paste(filename, + id)
#system(cmd, intern = TRUE, input = read.table(filename, header = TRUE, sep = "\t",
encoding = "UTF-8"))
# stem <- system(cmd, intern = TRUE, input = as.character(read.table(filename, header =
TRUE, sep = "\t", encoding = "UTF-8"), output = write(stem,"0_"))
stem <- system(cmd, intern = TRUE, input = text3)
#for(id in subject_id){
myfile<-paste0("0_", id)
# write.table(stem,myfile,sep="\t",row.names=FALSE, col.names=F,)
#}
write.table(stem, myfile, row.names=F,col.names=F,sep="\t")
}
Пример задания темы 8
#####Bayes classifier#####
# Loading package
library(e1071)
library(caTools)
library(caret)
# Fitting Naive Bayes Model
# to training dataset
classifier_auth <- naiveBayes(Author ~ ., data = author.numtr)

classifier_auth

# Predicting on test data'
y_pred <- predict(classifier_auth, newdata = author.numte)

# Confusion Matrix
cm <- table(author.numte$Author, y_pred)
cm
Примените алгоритм классификации Байеса на своих данных (частотной матрицы).

```

Информация о разработчиках

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