***Программа итогового контроля по дисциплине***

***«****Методы машинного обучения в обработке естественного языка****»***

***на 2023/2024 учебный год***

*весенний семестр*

**Факультет***\_Информационных технологий\_\_\_*

**Кафедра***\_Информационные системы\_\_\_\_\_\_\_*

**Шифр и наименование образовательной программы *«****7M06101 – Вычислительная лингвистика»*

**Отделение**: *русский*

**Уровень образования** *магистр*

**Курс***: 1*

**Преподаватель**: *Карюкин Владислав Игоревич*

**Форма проведения итогового контроля** – \_*устно*\_\_

**Формат экзамена –** *офлайн*

***ТЕМАТИЧЕСКАЯ ПРОГРАММА ДИСЦИПЛИНЫ НА ОСНОВЕ ТЕМ МОДУЛЕЙ, ЛЕКЦИЙ, СЕМИНАРОВ***

1. Введение в обработку естественных языков
2. Основные этапы предобработки текстовых данных
3. Операции стемминга с текстовыми данными
4. Операции лемматизации с текстовыми данными
5. Операции векторизации текстовых данных
6. Подготовка текстовых данных для классификации моделями машинного обучения
7. Классификация текстовых данных моделями машинного обучения
8. Классификация текстовых данных нейронными сетями
9. Большие языковые модели BERT, GPT
10. Анализ и обработка текстов с помощью запросов ChatGPT
11. Основные методы анализа текстовых документов в HTML формате
12. Работа с парсингом HTML страниц с помощью BeautifulSoup и Scrapy
13. Создание программы парсинга текста с помощью Selenium web driver
14. Добавление функций поиска данных в социальных сетях
15. Основные этапы создания веб-приложения с моделями машинного обучения

***МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ВЫПОЛНЕНИЯ ЗАДАНИЯ ИТОГОВОГО КОНТРОЛЯ ПО ВЫБРАННОЙ ФОРМЕ***

**Стандартный экзамен:** *устно*

**Формат экзамена –** *офлайн***.**

Общее число экзаменационных вопросов по дисциплине: 15

Данная форма предназначена для итогового контроля по дисциплинам, которые формируют навыки студента излагать ответы и доказательства положений в устной форме, вести дискуссии с экзаменационной комиссией, обосновывать свою точку зрения, приводить аргументы и доводы, способствуют развитию коммуникативной компетенции студента. Данная форма позволяет установить непосредственный контакт между экзаменационной комиссией и студентом, в процессе которого студент демонстрирует уровень освоения учебного материала. Выполнение практических заданий по разработке аппаратного/программного обеспечения предполагает использование компьютеров, лабораторного оборудования для сборки и запуска аппаратных модулей.

Индивидуальный опрос предполагает обстоятельные, связные ответы студентов на вопрос, относящийся к изучаемому учебному материалу, поэтому он служит важным средством различия речи, памяти, мышления студентов. Чтобы сделать такую проверку более глубокой, необходимо ставить перед студентами вопросы, требующие развернутого ответа. Вопросы должны быть четкими, ясными, конкретными, емкими, иметь прикладной характер.

- цель и ожидаемые результаты выполнения задания

цель – изучение концепций обработки естественных языков, базовых принципов и задач NLP, включая обучение с учителем, обучение без учителя, глубокое обучение, включающее архитектуры сверточных нейронных сетей, рекуррентных нейронных сетей, трансформеров и больших языковых моделей, таких как BERT, GPT-3 и т.д.

результаты – понимать базовые и расширенные программы по обработке текстов, анализировать особенности методов стемминга и лемматизации текстов, применять библиотеки NumPy, Pandas, Matplotlib для работы с моделями машинного обучения, разрабатывать скрипты для парсинга текстов, создавать основной каркас веб-приложения на HTML, CSS и JavaScript.

- форма представления выполненного задания (шаблоны/структуры и т.д.)

Устный ответ на вопрос и представление программного кода

***ОСНОВНЫЕ ЭТАПЫ РАБОТЫ ПО ИНСТРУКЦИИ***

**Длительность**

Время на подготовку – 20 мин.

Время на ответ – 15 мин.

Экзаменационный билет содержит 3 вопроса: 2 вопроса по теории, 1 вопрос практическое задание. В каждом вопросе в скобках указывается соответствующая максимальная оценка, указанная в процентах.

*Организация проведения устного офлайн экзамена*

- при входе в аудиторию проведения экзамена обучающийся обязан предоставить экзаменатору удостоверение личности и поставить подпись в явочном листе;

- вставить и / или меняться местами, выходить из аудитории до завершения своего ответа на билет в ходе экзамена запрещено;

- при проведении устного экзамена экзаменационный билет выбирает сам экзаменующийся;

- по приглашению преподавателя обучающийся поочередно получает экзаменационный билет;

- при необходимости будет возможность подготовиться к ответу на вопросы экзаменационного билета во время подготовки. А если обучающийся полностью готов к вопросам экзаменационного билета, он может ответить сразу;

- экзаменатор имеет право с целью более глубокого выяснения уровня знаний обучающегося, задавать ему дополнительные вопросы, а также предлагать задачи и примеры в рамках вопросов экзаменационного билета.

**Требования к сдаче экзамена:**

- согласно расписанию;

- можно отвечать на вопросы в любой последовательности.

- если будет обнаружено использование несанкционированных материалов или получения иных подсказок обучающимся, экзамен может быть аннулирован.

***ПОЛИТИКА ОЦЕНИВАНИЯ – РУБРИКАТОР ОЦЕНИВАНИЯ***

**Шаблон: РУБРИКАТОР КРИТЕРИАЛЬНОГО ОЦЕНИВАНИЯ ИТОГОВОГО КОНТРОЛЯ**

*(для форм стандартный устный / письменный)*

**Дисциплина**: Методы машинного обучения в обработке естественного языка **Форма:** стандартный устный. **Платформа:** ИС Univer

|  |  |  |
| --- | --- | --- |
| **№** вопроса  | **Балл**  **Критерий**  | **ДЕСКРИПТОРЫ**  |
| **«Отлично»**   | **«Хорошо»**    | **«Удовлетворительно»**    | **«Неудовлетворительно»**    |
| **90-100** % | **70-89** % | **50-69** % | * 1. %
 | **0-24** % |
| **1-2**  | Знание и понимание теории и концепции курса | На вопросы даны исчерпывающие ответы, обоснованы, проиллюстрированные наглядными примерами там, где это необходимо; Ответы изложены грамотным научным языком, показано полное понимание всех основных операций классификации текстовых данных моделями машинного обучения | На вопросы даны в целом верные ответы, но с отдельными неточностями, не носящими принципиального характера. Не полностью показано понимание всех основных операций классификации текстовых данных моделями машинного обучения, не все команды употреблены правильно, присутствуют отдельные некорректные утверждения и грамматические/ стилистические погрешности изложения. Ответы не проиллюстрированы примерами в должной мере. | Ответы на вопросы носят реферативный характер, верные выводы перемежаются с неверными. Упущены содержательные блоки моделей машинного обучения, необходимые дляполного раскрытия темы. Студент в целомориентируется в тематикеучебного курса, но испытывает проблемы сраскрытием конкретныхвопросов. | Ответы несоответствуют содержанию вопросов.Ключевые для учебногокурса понятия,содержащиеся в вопросах, трактуются ошибочно. | Ответы на вопросы отсутствуют;обнаружено незнаниеили непониманиестудентом большей илинаиболее важной частиучебного материала.Нарушение правил проведения итоговогоконтроля. |
| **3**  | Оценивание и анализ применимости выбранной методики к предложенной практической задаче, обоснование полученного результата  | Наличие способности к интеграции, обоснованности и анализу методов и технологии классификации текстов с помощью моделей нейронных сетей, ответы иллюстрируется примерами и наглядными материалами, написанием кода, демонстрирует умение вести диалог и вступать в научную дискуссию.  | Интеграция и анализ применения методов и технологии курса с последующим использованием наглядных материалов для закрепления своих рассуждений посредством классификации текстов с помощью моделей нейронных сетей с допущением незначительных ошибок при воспроизведении знаний; анализировать направление по вопросу экзаменационного билета. | Поверхностное обоснование команд и операторов моделей нейронных сетей, слабое применение основного объема материала в соответствии с программой обучения с затруднениями при его самостоятельном воспроизведении и требованием наводящих вопросов;  | Отсутствие обоснованности и анализа применения методов и технологии курса, проявление затруднения при предоставлении ответов на вопросы воспроизводящего характера.  | Отсутствие способности применять методологию курса при приведении примеров, использовании наглядных материалов; Нарушение Правил проведения итогового контроля.  |

***СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ***

# Python for Everybody: Exploring Data in Python 3 by Dr. Charles Russell Severance, Sue Blumenberg, Elliott Hauser, Aimee Andrion, 2016.

1. Natural Language Processing with Python and spaCy: A Practical Introduction, Yuli Vasiliev, 2021.
2. Machine Learning and Deep Learning in Natural Language Processing,
Anitha S. Pillai, Roberto Tedesco, 2023.
3. Natural Language Processing: A Machine Learning Perspective
Yue Zhang, Zhiyang Teng, 2021.
4. Natural Language Processing Projects: Build Next-Generation NLP Applications Using AI Techniques, Akshay Kulkarni, Adarsha Shivananda, Anoosh Kulkarni, 2021.
5. Learning Scientific Programming with Python, Christian Hill, 2021
6. Deep Learning for Natural Language Processing: Creating Neural Networks with Python. Palash Goyal, Sumit Pandey, Karan Jain, 2018