Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Макаренко Елена Николаевна
Должность: Ректор
Дата подписания: 27.12.2024 10:55:07
Уникальный программный ключ:
c098bc0c1041cb2a4cf926cf171d6715d99a6ae00adc8e27b55cbe1e2dbd7c78

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ
Начальник
учебно-методического управления
Платонова Т.К.
«25» июня 2024 г.

Рабочая программа дисциплины
Интеллектуальная обработка текстов

Направление 01.03.02 "Прикладная математика и информатика"
Направленность 01.03.02.02 "Математическое и программное обеспечение систем искусственного интеллекта"

Для набора 2021 года

Квалификация
Бакалавр

КАФЕДРА **Информационных систем и прикладной информатики**

Распределение часов дисциплины по семестрам

Семестр (<Курс>.<Семестр на курсе>)	**6 (3.2)**		Итого	
Недель	16			
Вид занятий	УП	РП	УП	РП
Лекции	6	6	6	6
Лабораторные	6	6	6	6
Итого ауд.	12	12	12	12
Контактная работа	12	12	12	12
Сам. работа	92	92	92	92
Часы на контроль	4	4	4	4
Итого	108	108	108	108

ОСНОВАНИЕ

Учебный план утвержден учёным советом вуза от 25.06.2024 г. протокол № 18.

Программу составил(и): доцент, Хаймин Е.С.

Зав. кафедрой: д.э.н., проф. Щербаков С.М.

Методический совет направления: д.э.н., профессор Тищенко Е.Н.

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1	освоение методов и технологий анализа, обработки и понимания текстовой информации с применением современных алгоритмов машинного обучения и искусственного интеллекта.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ОПК-2: Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач

ПК-2: Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта

В результате освоения дисциплины обучающийся должен:

Знать:

математические методы обработки, анализа и синтеза результатов при создании моделей интеллектуальных систем алгоритмы обработки текстов (соотнесено с индикатором ОПК-2.1)
методы и инструменты искусственного интеллекта (соотнесено с индикатором ПК-2.1)

Уметь:

использовать математические методы обработки, анализа и синтеза результатов при создании моделей интеллектуальных систем (соотнесено с индикатором ОПК-2.2)
реализовать алгоритмы машинного обучения; классифицировать задачи искусственного интеллекта (соотнесено с индикатором ПК-2.2)

Владеть:

навыками использования математических методов обработки, анализа и синтеза результатов при создании моделей интеллектуальных систем (соотнесено с индикатором ОПК-2.3)
методами разработки программного обеспечения, оценкой и валидацией результатов (соотнесено с индикатором ПК-2.3)

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Раздел 1. ПРЕДСТАВЛЕНИЕ И ВЫВОД ЗНАНИЙ

№	Наименование темы / Вид занятия	Семестр / Курс	Часов	Компетен-ции	Литература
1.1	Тема 1. «Введение в анализ текстов. Регулярные выражения» Развитие методов и технологий в области искусственного интеллекта. Анализ текстов — это обширная область, изучающая методы обработки, интерпретации и извлечения информации из текстовых данных. Анализ текстов позволяет выявлять скрытые закономерности, анализировать мнения, классифицировать документы и выполнять другие задачи, необходимые для извлечения понятной информации из обширных объемов данных Регулярные выражения (regex) — это мощные инструменты для поиска и манипуляции текстовыми строками на основе заданных шаблонов. Они позволяют находить, заменять и проверять наличие определённых последовательностей символов в текстах, что делает их незаменимыми при предобработке текстовых данных. / Лек /	6	2	ОПК-2, ПК-2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.2	Тема 2. «Текстовые модели. Мешок слов. Модель TF-IDF» Текстовые модели — это методы представления текстовой информации в числовом формате, позволяющие применять алгоритмы машинного обучения и анализа данных; помогают преобразовать текстовые данные в структуру, удобную для вычислительных методов, что является основным шагом в многих задачах обработки естественного языка (NLP). Мешок слов — это простая и широко используемая модель, которая игнорирует грамматическую структуру и порядок слов в тексте, рассматривая текст как набор слов. TF-IDF (Term Frequency-Inverse Document Frequency) — это более продвинутая модель, которая учитывает как частоту слова в документе, так и его распространенность в коллекции документов. / Лек /	6	2	ОПК-2, ПК-2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2

1.3	Тема 3. «Текстовые модели. Векторное представление слов. Модель word2vec. Модель FastText» Векторное представление слов — это метод представления слов в виде непрерывных векторов в многомерном пространстве. Такой подход позволяет захватить семантические и синтаксические отношения между словами, что дает возможность эффективно использовать их в задачах обработки естественного языка (NLP) Word2Vec — это популярная модель, разработанная Google, которая обучает векторные представления слов с использованием нейронных сетей. Она включает два основных алгоритма: 1. CBOW (Continuous Bag of Words): Эта архитектура предсказывает текущее слово на основе контекста. Модель принимает векторные представления слов контекста и пытается угадать центральное слово. 2. Skip-gram: В отличие от CBOW, эта архитектура предсказывает слова контекста на основе текущего слова. Она работает лучше для редких слов и помогает создать качественные векторные представления. FastText — это улучшенная версия векторного представления слов, разработанная командой Facebook. Она была представлена в 2016 году и строится на принципах, сходных с word2vec, но с ключевыми различиями, которые увеличивают его эффективность, особенно для обработки языков с богатым морфологическим строением. / Лек /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.4	Тема 4. «Счетные языковые модели» Счетные языковые модели представляют собой класс моделей, которые основываются на принципах вероятностной оценки последовательностей слов в языке. Они используют статистические методы для вычисления вероятностей появления слов и их сочетаний. В отличие от нейронных сетей, такие модели часто являются более простыми и интерпретируемыми, но могут показывать низкую эффективность на больших объемах данных или в более сложных задачах. / Ср /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.5	Тема 5. «Сверточные нейронные сети для анализа текстов» Сверточные нейронные сети (CNN), изначально разработанные для обработки изображений, также нашли применение в области анализа текстов. Их использование в текстовой обработке можно объяснить тем, что они способны захватывать локальные паттерны и контексты, что делает их подходящими для задач классификации текстов, извлечения признаков и других приложений в области естественной обработки языка (NLP). / Ср /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.6	Тема 6. «Генерация текстов. Рекурентные нейронные сети» Рекуррентные нейронные сети (RNN) являются одним из основных инструментов для обработки последовательных данных, таких как тексты. Их уникальная структура позволяет эффективно работать с проблемами, связанными со временем и последовательностью, что делает их особенно подходящими для технологий генерации текстов / Ср /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.7	Тема 7. «QA-системы. Основы разработки чат-ботов» Разработка чат-ботов и QA-систем (систем вопросов и ответов) стала одной из наиболее актуальных тем в области искусственного интеллекта и обработки естественного языка (NLP). Чат-боты могут выполнять различные функции, начиная от простых задач, таких как предоставление информации о продуктах, до более сложных, включая обслуживание клиентов и техническую поддержку. Давайте рассмотрим основные аспекты разработки чат-ботов и систем QA. / Ср /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.8	Тема 8. «Оценка качества моделей обработки текстов» Оценка качества моделей обработки текстов является важным шагом в разработки и внедрении алгоритмов, направленных на анализ, генерацию, или понимание естественного языка. Различные задачи обработки текстов требуют применения	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2

	различных методологий оценки, исходя из специфики работы. / Ср /				
1.9	Лабораторное задание 1 «Регулярные выражения» Изучение базовых концепций регулярных выражений. Наблюдение за применением регулярных выражений в реальных сценариях (поиск, замена, валидация). Разработка и тестирование регулярных выражений для различных задач. / Лаб /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.10	Лабораторное задание 2 «Мешок слов. Модель TF-IDF» Изучение структуры текстовых данных и их представление в виде модели мешка слов. Понимание и применение модели TF-IDF для оценки важности слов в документах. Освоение инструментов для реализации обработки текстов с помощью Python и библиотек, таких как scikit-learn. / Лаб /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.11	Лабораторное задание 3 «Векторное представление слов. Модель word2vec. Модель FastText» Ознакомление с концепцией векторного представления слов. Изучение алгоритмов Word2Vec и FastText для создания векторных представлений слов. Проведение практических экспериментов по обучению и использованию моделей Word2Vec и FastText. / Лаб /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.12	Лабораторное задание 4 «Счетные языковые модели» Понять основную концепцию счетных языковых моделей. Изучить различные типы таких моделей и их применение в обработке естественного языка (NLP). Реализовать счетные языковые модели и проанализировать их эффективность / Ср /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.13	Лабораторное задание 5 «Сверточные нейронные сети для анализа текстов» Понять структуру и принципы работы сверточных нейронных сетей (CNN) в контексте анализа текстов. Ознакомиться с предобработкой текстовых данных для обучения моделей. Реализовать модель сверточной нейронной сети для решения задачи классификации текстов. Проанализировать результаты и оценить эффективность модели. / Ср /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.14	Лабораторное задание 6 «Генерация текста. Рекурентные нейронные сети» Изучить принципы работы рекуррентных нейронных сетей для генерации текстов. Ознакомиться с процессом подготовки данных для обучения RNN. Реализовать модель RNN для генерации текста на основе заданного исходного текста. Оценить качество сгенерированного текста и сделать выводы о работе модели. / Ср /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.15	Лабораторное задание 7 «QA-системы. Основы разработки чат-ботов» При разработке чат-бота важно определить его цели и задачи. Функциональность: Определите, что бот должен делать. Будет ли это просто отвечать на вопросы, предоставлять рекомендации, обрабатывать заказы или выполнять другие задачи? Целевая аудитория: Исследуйте, кто будет использовать бота и какие у них потребности. Это поможет в создании более персонализированного опыта. Контент и данные: Решите, какую информацию бот должен обрабатывать. Если это QA-система, то нужно определить, откуда будут поступать данные (базы данных, API и т.д.). / Ср /	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.16	Лабораторное задание.8 «Оценка качества модели» Определить, как хорошо модель генерирует текст, схожий по	6	2	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2

	стилю и смыслу с тренировочными данными. Измерить производительность модели с помощью количественных показателей. Провести качественную оценку сгенерированного текста. Сравнить модель с другими подходами, если это возможно. / Ср /				
1.17	Выполнение индивидуальных заданий по теме «Анализ текстов с использованием моделей машинного обучения». / Ср /	6	72	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2
1.18	Зачет / Зачёт /	6	4	ОПК-2, ПК -2	Л1.1, Л1.2, Л1.3, Л2.1, Л2.2

4. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Структура и содержание фонда оценочных средств для проведения текущей и промежуточной аттестации представлены в Приложении 1 к рабочей программе дисциплины.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Основная литература

	Авторы,	Заглавие	Издательство, год	Колич-во
Л1.1	Павлов С. И.	Системы искусственного интеллекта: учебное пособие	Томск: Томский государственный университет систем управления и радиоэлектроники, 2011	https://biblioclub.ru/index.php?page=book&id=208939 неограниченный доступ для зарегистрированных пользователей
Л1.2	Семенов А., Соловьев Н., Чернопрудова Е., Цыганков А.	Интеллектуальные системы: учебное пособие	Оренбург: Оренбургский государственный университет, 2013	https://biblioclub.ru/index.php?page=book&id=259148 неограниченный доступ для зарегистрированных пользователей
Л1.3	Крюкова, А. А.	Интеллектуальные технологии в бизнесе: методические указания к практическим и лабораторным работам	Самара: Поволжский государственный университет телекоммуникаций и информатики, 2013	https://www.iprbookshop.ru/71835.html неограниченный доступ для зарегистрированных пользователей

5.2. Дополнительная литература

	Авторы,	Заглавие	Издательство, год	Колич-во
Л2.1	Кухаренко Б. Г.	Интеллектуальные системы и технологии: учебное пособие	Москва: Альтаир|МГАВТ, 2015	https://biblioclub.ru/index.php?page=book&id=429758 неограниченный доступ для зарегистрированных пользователей
Л2.2		Прикладная информатика: журнал	Москва: Университет Синергия, 2023	https://biblioclub.ru/index.php?page=book&id=699833 неограниченный доступ для зарегистрированных пользователей

5.3 Профессиональные базы данных и информационные справочные системы

Гарант https://www.garant.ru/
Консультант +
Национальная электронная библиотека (НЭБ) - https://rusneb.ru/
Документация Python – https://python.org

5.4. Перечень программного обеспечения

Операционная система РЕД ОС
LibreOffice
IDlE Python
Jupyter Notebook
Pycharm
PostgreSQL

5.5. Учебно-методические материалы для студентов с ограниченными возможностями здоровья

При необходимости по заявлению обучающегося с ограниченными возможностями здоровья учебно-методические материалы предоставляются в формах, адаптированных к ограничениям здоровья и восприятия информации. Для лиц с нарушениями зрения: в форме аудиофайла; в печатной форме увеличенным шрифтом. Для лиц с нарушениями слуха: в форме электронного документа; в печатной форме. Для лиц с нарушениями опорно-двигательного аппарата: в форме электронного документа; в печатной форме.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Помещения для всех видов работ, предусмотренных учебным планом, укомплектованы необходимой специализированной учебной мебелью и техническими средствами обучения:
- столы, стулья;
- персональный компьютер / ноутбук (переносной);
- проектор;
- экран / интерактивная доска.

Лабораторные занятия проводятся в компьютерных классах, рабочие места в которых оборудованы необходимыми лицензионными программными средствами и выходом в Интернет.

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Методические указания по освоению дисциплины представлены в Приложении 2 к рабочей программе дисциплины.

Приложение 1

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

1.1. Показатели и критерии оценивания компетенций:

ЗУН, составляющие компетенцию	Показатели оценивания	Критерии оценивания	Средства оценивания
ОПК-2: Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач			
З. математические методы обработки, анализа и синтеза результатов при создании моделей интеллектуальных систем алгоритмы обработки текстов	знает основные математические методы обработки, анализа и синтеза результатов при создании моделей интеллектуальных систем алгоритмы обработки текстов	полнота и содержательность ответа умение приводить примеры	Вопросы к зачету (1-32), опрос (1-10), лабораторные задания (1-8), индивидуальное задание (1-10)
У. использовать математические методы обработки, анализа и синтеза результатов при создании моделей интеллектуальных систем	выполняет задания, отвечает на вопросы, умеет применять полученные знания на практике	полнота и содержательность ответа умение приводить примеры умение самостоятельно находить решение поставленных задач	Вопросы к зачету (1-32), опрос (1-10), лабораторные задания (1-8), индивидуальное задание (1-10)
В. навыками использования математических методов обработки, анализа и синтеза результатов при создании моделей интеллектуальных систем	проводит обобщенный анализ информации и обработку данных на различных задачах анализа и обработки текста	полнота и содержательность ответа умение приводить примеры умение самостоятельно находить решение поставленных задач	Вопросы к зачету (1-32), опрос (1-10), лабораторные задания (1-8), индивидуальное задание (1-10)
ПК-2: Способен классифицировать и идентифицировать задачи искусственного интеллекта, выбирать адекватные методы и инструментальные средства решения задач искусственного интеллекта			
З. методы и инструменты искусственного интеллекта	знает основные понятия и определения, методы, алгоритмы и технологии	полнота и содержательность ответа умение приводить примеры	Вопросы к зачету (1-32), опрос (1-10), лабораторные задания (1-8), индивидуальное задание (1-10)
У. реализовать алгоритмы машинного обучения; классифицировать задачи искусственного интеллекта	выполняет задания, реализовывает алгоритмы машинного обучения для различных задачах анализа и обработки текста	полнота и содержательность ответа умение приводить примеры умение самостоятельно находить решение поставленных задач	Вопросы к зачету (1-32), опрос (1-10), лабораторные задания (1-8), индивидуальное задание (1-10)
В. методами разработки программного обеспечения, оценкой и валидацией результатов	проводит обобщенный анализ информации и обработку данных, оценивает результаты	полнота и содержательность ответа умение приводить примеры умение самостоятельно находить решение поставленных задач	Вопросы к зачету (1-32), опрос (1-10), лабораторные задания (1-8), индивидуальное задание (1-10)

1.2 Шкалы оценивания для зачета:

Текущий контроль успеваемости и промежуточная аттестация осуществляется в рамках накопительной балльно-рейтинговой системы в 100-балльной шкале:

50-100 баллов (зачтено),

0-49 баллов (не зачтено).

2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Вопросы к зачету

1. Что такое анализ текстов и какие задачи он решает?
2. Объясните, что такое регулярные выражения и как они применяются в обработке текстов.
3. Приведите примеры регулярных выражений для поиска email адресов и телефонных номеров в тексте.
4. Как регулярные выражения могут помочь при предварительной обработке текстов?
5. Что такое модель «мешок слов» и как она строится?
6. Как работает модель TF-IDF? Объясните, как вычисляются TF и IDF.
7. В чем основные преимущества и недостатки модели «мешок слов» по сравнению с другими методами представления текстов?
8. Как можно использовать TF-IDF для индексирования и поиска документов?
9. Что такое векторное представление слов и в чем его преимущества?
10. Как работает модель word2vec? Опишите различия между методами Skip-gram и Continuous Bag of Words.
11. В чем отличие модели FastText от word2vec? Как FastText учитывает морфологию слов?
12. Как векторные представления слов могут использоваться в задачах обработки естественного языка?
13. Что такое счетные языковые модели и как они применяются в анализе текстов?
14. Объясните, чем отличаются n-граммные модели от моделей на основе более сложных языковых моделей.
15. Как вычисляется вероятность n-граммы в контексте языковых моделей?
16. Какие существуют методы сглаживания для улучшения n-граммных моделей?
17. Как сверточные нейронные сети (CNN) могут быть использованы для обработки текстов?
18. Объясните, как конструкция сверточных слоев может быть применена для извлечения признаков из текстовых данных.
19. Какие преимущества имеют CNN по сравнению с традиционными методами анализа текстов?
20. Какие задачи анализа текстов можно решать с помощью сверточных нейронных сетей?
21. Как рекуррентные нейронные сети (RNN) помогают в генерации текстов?
22. Чем модели LSTM и GRU отличаются от обычных RNN и какие преимущества они предоставляют?
23. Опишите процесс обучения генеративной модели текста с помощью RNN.
24. Как можно оценить качество сгенерированного текста?
25. Что такое QA-система и как она работает?
26. Какие ключевые этапы разработки чат-бота вы можете выделить?
27. Как можно интегрировать чат-бота с существующими системами данных?
28. Объясните, какие методы используются для обработки пользовательских запросов в QA-системах.
29. Какие метрики используются для оценки качества моделей обработки текстов?
30. Объясните, как рассчитывается точность, полнота и F-мера.
31. Как проходит процесс валидации моделей в контексте обработки текстов?
32. Какие методы можно использовать для улучшения качества существующих моделей?

Зачетное задание включает в себя один теоретический вопрос из представленного перечня и одно практико-ориентированное задание из подраздела «Лабораторные задания».

Критерии оценивания:

- 50-100 баллов («зачтено») – изложенный материал фактически верен, наличие глубоких исчерпывающих знаний в объеме пройденной программы дисциплины в соответствии с поставленными программой курса целями и задачами обучения; правильные, уверенные действия по применению полученных знаний на практике, грамотное и логически стройное изложение материала при ответе, усвоение основной и знакомство с дополнительной литературой; наличие твердых и достаточно полных знаний в объеме пройденной программы дисциплины в соответствии с целями обучения, правильные действия по применению знаний на практике, четкое изложение материала, допускаются отдельные логические и стилистические погрешности, обучающийся усвоил основную литературу, рекомендованную в рабочей программе дисциплины; наличие твердых знаний в объеме пройденного курса в соответствии с целями обучения, изложение ответов с отдельными ошибками, уверенно исправленными после дополнительных вопросов; правильные в целом действия по применению знаний на практике;

- 0-49 баллов («не зачтено») – ответы не связаны с вопросами, наличие грубых ошибок в ответе, непонимание сущности излагаемого вопроса, неумение применять знания на практике, неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Задания для опроса

Вариант 1: Определение понятий

Опишите, что такое анализ текстов. Какие основные задачи ставятся перед анализом текстов, и как регулярные выражения применяются для их решения?

Вариант 2: Работа с регулярными выражениями

Напишите регулярное выражение для поиска всех URL в заданном тексте. Приведите примеры текстов, в которых можно применить ваше регулярное выражение.

Вариант 3: Модель мешка слов

Объясните, как формируется модель «мешок слов». Приведите пример, как эта модель может быть использована для классификации текстов.

Вариант 4: TF-IDF вычисления

Дайте определение понятиям TF и IDF. Показать на практике, как вы можете вычислить TF-IDF для слова "анализ" в наборе документов.

Вариант 5: Сравнение текстовых моделей

Сравните модель «мешок слов» и TF-IDF. Каковы преимущества и недостатки каждой из моделей? В каких случаях какую модель следует использовать?

Вариант 6: Векторные представления слов

Объясните, как работает модель word2vec. Как различаются методы Skip-gram и Continuous Bag of Words на практике? Приведите примеры использования каждого метода.

Вариант 7: Модель FastText

В чем преимущество модели FastText по сравнению с word2vec? Как FastText обрабатывает морфологию слов, и как это сказывается на качестве векторных представлений?

Вариант 8: Счетные языковые модели

Опишите, что такое счетные языковые модели. Как они могут быть использованы для предсказания следующего слова в предложении? Приведите пример реализации.

Вариант 9: Применение сверточных нейронных сетей

Как сверточные нейронные сети могут быть применены для анализа текстов? Опишите, как они могут обрабатывать текстовые признаки, используя архитектуру CNN.

Вариант 10: Генерация текстов

Опишите, как рекуррентные нейронные сети используются для генерации текстов. Каковы основные шаги в процессе обучения RNN для генерации связного текста? Приведите пример задачи.

Критерии оценивания:

9-10 б. – ответы на все вопросы даны верно;

7-8 б. – один из ответов с неточностями;

5-6 б. – 2 ответа с неточностями;

3-4 б. – 3 ответа с неточностями;
1-2 б. – нет ответа на один вопрос.
0 б – нет ответов на вопросы или все ответы неверные.

Максимальное количество баллов за опрос – 10.

Лабораторные задания

Лабораторное задание №1
«Регулярные выражения»
Изучение базовых концепций регулярных выражений. Наблюдение за применением регулярных выражений в реальных сценариях (поиск, замена, валидация). Разработка и тестирование регулярных выражений для различных задач.
Лабораторное задание №2
«Мешок слов. Модель TF-IDF»
Изучение структуры текстовых данных и их представление в виде модели мешка слов.
Понимание и применение модели TF-IDF для оценки важности слов в документах.
Освоение инструментов для реализации обработки текстов с помощью Python и библиотек, таких как scikit-learn.
Лабораторное задание №3
«Векторное представление слов. Модель word2vec. Модель FastText»
Ознакомление с концепцией векторного представления слов.
Изучение алгоритмов Word2Vec и FastText для создания векторных представлений слов.
Проведение практических экспериментов по обучению и использованию моделей Word2Vec и FastText.
Лабораторное задание №4
«Счетные языковые модели»
Понять основную концепцию счетных языковых моделей.
Изучить различные типы таких моделей и их применение в обработке естественного языка (NLP).
Реализовать счетные языковые модели и проанализировать их эффективность
Лабораторное задание №5
«Сверточные нейронные сети для анализа текстов»
Понять структуру и принципы работы сверточных нейронных сетей (CNN) в контексте анализа текстов.
Ознакомиться с предобработкой текстовых данных для обучения моделей.
Реализовать модель сверточной нейронной сети для решения задачи классификации текстов.
Проанализировать результаты и оценить эффективность модели.

Лабораторное задание №6
«Генерация текста. Рекурентные нейронные сети»
Изучить принципы работы рекуррентных нейронных сетей для генерации текстов.
Ознакомиться с процессом подготовки данных для обучения RNN.
Реализовать модель RNN для генерации текста на основе заданного исходного текста.
Оценить качество сгенерированного текста и сделать выводы о работе модели.
Лабораторное задание №7
«QA-системы. Основы разработки чат-ботов»
При разработке чат-бота важно определить его цели и задачи.
Функциональность: Определите, что бот должен делать. Будет ли это просто отвечать на вопросы, предоставлять рекомендации, обрабатывать заказы или выполнять другие задачи?
Целевая аудитория: Исследуйте, кто будет использовать бота и какие у них потребности. Это поможет в создании более персонализированного опыта.

Контент и данные: Решите, какую информацию бот должен обрабатывать. Если это QA-система, то нужно определить, откуда будут поступать данные

Лабораторное задание №8

«Оценка качества модели»

Определить, как хорошо модель генерирует текст, схожий по стилю и смыслу с тренировочными данными.

Измерить производительность модели с помощью количественных показателей.

Провести качественную оценку сгенерированного текста.

Сравнить модель с другими подходами, если это возможно

Критерии оценивания (для каждого задания):

5 б. – задание выполнено верно;

4 б. – при выполнении задания были допущены неточности, не влияющие на результат;

3 б. – при выполнении задания были допущены ошибки;

0-2 б. – при выполнении задания были допущены существенные ошибки.

Максимальное количество баллов за все лабораторные задания – 40 (8 заданий по 5 баллов).

Индивидуальное задание

Задание 1: Анализ эмоциональной окраски текста

Выберите несколько статей из новостных источников и выполните анализ их эмоциональной окраски. Используйте библиотеки для обработки естественного языка, такие как NLTK или TextBlob, чтобы определить, положительная, нейтральная или отрицательная тональность статьи. Представьте результаты в виде графиков или таблиц.

Задание 2: Классификация текстов

Соберите набор документов на заданную тему (например, отзывы о продуктах, статьи о технологиях и т. д.) и постройте модель для классификации текстов по категориям. Используйте методы машинного обучения, такие как Naive Bayes или SVM. Оцените эффективность вашей модели с помощью метрик, таких как точность или полнота.

Задание 3: Выделение ключевых слов

Напишите программу, которая будет анализировать текст и выделять ключевые слова с использованием метода TF-IDF. Примените полученные ключевые слова для краткого резюмирования текста. Обсудите, насколько хорошо ключевые слова отражают содержание оригинального текста.

Задание 4: Построение модели «мешок слов»

Создайте модель «мешок слов» на наборе текстов (например, литературные произведения, статьи). Проведите анализ полученных векторов и визуализируйте их с помощью метода t-SNE или PCA. Обсудите, какие паттерны вы заметили в данных.

Задание 5: Генерация текста

Создайте простую модель генерации текста, используя рекуррентные нейронные сети (RNN) или GPT. Обучите модель на выбранном наборе текстов, а затем сгенерируйте новые примеры текста. Оцените качество сгенерированного материала и предоставьте рекомендации по его улучшению.

Задание 6: Анализ социальных медиа

Выберите платформу социальных медиа (например, Twitter, Instagram) и соберите данные о публикациях на заданную тему. Проведите аналитику на основе собранных данных, выделяя основные темы, хештеги и инфлюенсеров. Используйте визуализацию данных для представления результатов.

Задание 7: Перевод текста

Используйте API для машинного перевода (например, Google Translate или DeepL) и проведите анализ качества перевода. Сравните переведенные тексты с оригинальными, определите распространенные ошибки, и дайте рекомендации по улучшению переводов.

Задание 8: Сравнение алгоритмов различия текстов

Напишите программу, которая будет сравнивать два текста и определять различия между ними. Используйте различные алгоритмы, такие как метод Левенштейна (для оценки редактирования) и метод Jaccard для оценки сходства между текстами. Проанализируйте время выполнения и точность каждого из методов.

Задание 9: Классификация плагиата

Создайте систему для обнаружения плагиата в текстах. Используйте алгоритмы сравнения текстов и извлечение признаков на основе TF-IDF для определения степени схожести между текстами. Представьте примеры, когда система справилась успешно и когда были сложности.

Задание 10: Фреймворк обработки естественного языка

Изучите и разработайте проект с использованием одного из современных фреймворков NLU, таких как Hugging Face Transformers или SpaCy. Реализуйте проект, который может выполнять несколько задач NLU, таких как Named Entity Recognition, анализ тональности и раскладка частей речи.

Критерии оценивания задания:

40-50 б. – задание выполнено верно;

21-39 б. – при выполнении задания были допущены неточности, не влияющие на результат;

11-20 б. – при выполнении задания были допущены ошибки;

0-10 б. – при выполнении задания были допущены существенные ошибки.

Максимальное количество баллов за индивидуальное задание– 50.

3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Процедуры оценивания включают в себя текущий контроль и промежуточную аттестацию.

Текущий контроль успеваемости проводится с использованием оценочных средств, представленных в п. 2 данного приложения. Результаты текущего контроля доводятся до сведения студентов до промежуточной аттестации.

Промежуточная аттестация проводится в форме зачета.

Зачет проводится по расписанию промежуточной аттестации. Количество вопросов в задании – 2. Объявление результатов производится в день зачета. Результаты аттестации заносятся в ведомость и зачетную книжку студента. Студенты, не прошедшие промежуточную аттестацию по графику промежуточной аттестации, должны ликвидировать задолженность в установленном порядке.

Приложение 2

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Учебным планом предусмотрены следующие виды занятий:

- лекции;
- лабораторные занятия.

В ходе лекционных занятий рассматриваются основные теоретические вопросы, даются рекомендации для самостоятельной работы и подготовке к лабораторным занятиям.

В ходе лабораторных углубляются и закрепляются знания студентов по ряду рассмотренных на лекциях вопросов, развиваются навыки практической работы.

При подготовке к лабораторным каждый студент должен:

– изучить рекомендованную учебную литературу;

– изучить конспекты лекций;

– подготовить ответы на все вопросы по изучаемой теме.

В ходе выполнения индивидуального задания применяются знания студентов, полученные на лекциях и при выполнении лабораторных работ, развиваются навыки практической работы.

При выполнении индивидуального задания каждый студент должен:

– изучить рекомендованную учебную литературу;

– изучить конспекты лекций;

- материалы лабораторных заданий;

– проанализировать и выполнить индивидуальное задание, подготовить отчет в соответствии с требованиями к оформлению.

В процессе подготовки к лабораторным занятиям студенты могут воспользоваться консультациями преподавателя.

Вопросы, не рассмотренные на лекциях и лабораторных занятиях, должны быть изучены студентами в ходе самостоятельной работы. Контроль самостоятельной работы студентов над учебной программой курса осуществляется в ходе защиты индивидуального задания, выполнения лабораторных заданий. В ходе самостоятельной работы каждый студент обязан прочитать основную и по возможности дополнительную литературу по изучаемой теме, дополнить конспекты лекций недостающим материалом, выписками из рекомендованных первоисточников, выделить непонятные термины, найти их значение в энциклопедических словарях.

Студент должен готовиться к предстоящему лабораторному занятию по всем обозначенным в рабочей программе дисциплины вопросам.

Для подготовки к занятиям, текущему контролю и промежуточной аттестации студенты могут воспользоваться электронно-библиотечными системами. Также обучающиеся могут взять на дом необходимую литературу на абонементе университетской библиотеки или воспользоваться читальными залами.