

**Федеральное государственное образовательное
бюджетное учреждение высшего образования**
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ
ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)

Департамент анализа данных и машинного обучения
Факультета информационных технологий и анализа больших данных

УТВЕРЖДАЮ

**Проректор по учебной
и методической работе**

_____ **Е.А. Каменева**

25.04.2023 г.

С.В. Макрушин, В.А. Малекова

Обработка текстов на естественных языках

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
09.03.03 - Прикладная информатика,
ОП «Инженерия данных»,
ОП «Прикладные информационные системы в экономике и финансах»

*Рекомендовано Ученым советом
Факультета информационных технологий и анализа больших
данных (протокол №31 от 18.04.2023 г.)*

*Одобрено Советом учебно-научного
Департамента анализа данных и машинного обучения
(протокол №2 от 29.03.2023 г.)*

Москва 2023

СОДЕРЖАНИЕ

| | |
|--|----|
| 1. Наименование дисциплины..... | 2 |
| 2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине..... | 2 |
| 3. Место дисциплины в структуре образовательных программ..... | 4 |
| 4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся (в семестре, в сессию)..... | 4 |
| 5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий..... | 5 |
| 5.1. Содержание дисциплины..... | 5 |
| 5.2. Учебно-тематический план..... | 7 |
| 5.3. Содержание семинаров, практических занятий..... | 9 |
| 6. Учебно-методическое обеспечение для самостоятельной работы обучающихся по дисциплине..... | 11 |
| 6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы..... | 11 |
| 6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю..... | 12 |
| 7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине..... | 13 |
| 8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины..... | 18 |
| 9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины..... | 19 |
| 10. Методические указания для обучающихся по освоению дисциплины..... | 19 |
| 11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем..... | 21 |
| 12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине..... | 21 |

1.Наименование дисциплины

«Обработка текстов на естественных языках».

2.Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине

| Код компетенции | Наименование компетенции | Индикаторы достижения компетенции | Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции |
|-----------------|---|---|---|
| ПКН-4 | Способность проектировать и создавать интеллектуальные информационные системы, выбирать метод обучения в соответствии с анализом задачи | Демонстрирует знание основных понятий машинного обучения и интеллектуального анализа данных, понимание области и границ применимости, основные виды задач. | Знать: Студент должен знать основные понятия машинного обучения и интеллектуального анализа данных, такие как классификация, кластеризация, регрессия, нейронные сети, глубокое обучение и т.д. Он должен понимать область и границы применимости этих методов, а также основные виды задач, которые они могут решать. Уметь: Студент должен уметь проектировать и создавать интеллектуальные информационные системы, выбирать метод обучения в соответствии с анализом задачи. Он должен уметь использовать различные инструменты и библиотеки для реализации этих систем, а также оценивать их эффективность и точность. Кроме того, студент должен уметь анализировать данные, выбирать подходящие методы и модели для их обработки и решения задачи. |
| | | Демонстрирует знание популярных инструментальных средств машинного обучения, собирает датасет, строит модели, проводит их анализ и диагностику, делает содержательные выводы. | Знать: Студент должен знать основные понятия и технологии в области обработки текстов на естественных языках, а также популярные инструменты машинного обучения, используемые в этой области, такие как TensorFlow, Keras, PyTorch, scikit-learn и другие. Он также |

| | | | |
|--|--|---|---|
| | | | <p>должен быть знаком с методами сбора и подготовки данных, анализа и диагностики моделей, а также с методами оценки качества моделей.</p> <p>Уметь: Студент должен уметь проектировать и создавать интеллектуальные информационные системы для обработки текстов на естественных языках, выбирать наиболее подходящий метод обучения для решения конкретной задачи, собирать и подготавливать данные, строить и настраивать модели, проводить анализ и диагностику моделей, а также делать содержательные выводы на основе результатов анализа. Он должен также уметь работать с популярными инструментальными средствами машинного обучения и программирования, такими как Python, TensorFlow, Keras, PyTorch и другими.</p> |
| | | <p>Презентабельно демонстрирует результаты анализа данных и машинного обучения в форме, доступной непрофессионалу, структурирует отчет по проведенному анализу.</p> | <p>Знать: Студент должен знать основы лингвистической обработки естественного языка, методы машинного обучения и анализа данных, а также принципы проектирования информационных систем.</p> <p>Уметь: Студент должен уметь создавать и применять алгоритмы обработки текстов на естественных языках, выбирать и применять соответствующие методы машинного обучения и анализа данных для решения задач в области обработки текстов, проектировать и создавать интеллектуальные информационные системы. Также студент должен уметь структурировать и презентовать результаты анализа данных в доступной форме.</p> |

3. Место дисциплины в структуре образовательных программ

Дисциплина «Обработка текстов на естественных языках» относится к Циклу профиля (элективный) по направлению подготовки 09.03.03 – Прикладная информатика, ОП «Инженерия данных», ОП «Прикладные информационные системы в экономике и финансах».

4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся (в семестре, в сессию)

Очная форма обучения / очно-заочная форма обучения

| Вид учебной работы по дисциплине | Всего (в з/е и часах) | Семестр 7 / 8 (в часах) |
|---|--------------------------|----------------------------|
| Общая трудоемкость дисциплины | 3/108 | 108 |
| Контактная работа – Аудиторные занятия | 34 | 34 |
| <i>Лекции</i> | <i>16</i> | <i>16</i> |
| <i>Семинары, практические занятия</i> | <i>18</i> | <i>18</i> |
| Самостоятельная работа | 74 | 74 |
| Вид текущего контроля | Контрольная работа | Контрольная работа |
| Вид промежуточной аттестации | экзамен*, зачет | экзамен*, зачет |

* для ОП «Инженерия данных»

Институт онлайн-образования, заочная форма обучения

| Вид учебной работы по дисциплине | Всего (в з/е и часах) | Семестр 8 (в часах) |
|---|--------------------------|------------------------|
| Общая трудоемкость дисциплины | 3/108 | 108 |
| Контактная работа – Аудиторные занятия | 12 | 12 |
| <i>Лекции</i> | <i>4</i> | <i>4</i> |
| <i>Семинары, практические занятия</i> | <i>8</i> | <i>8</i> |
| Самостоятельная работа | 96 | 96 |
| Вид текущего контроля | Контрольная работа | Контрольная работа |
| Вид промежуточной аттестации | зачет | зачет |

5.Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий

5.1. Содержание дисциплины

Тема1. Введение в автоматическую обработку естественного языка

Структурные особенности текстов на естественном языке;

Неоднозначность на всех уровнях языка;

Основные задачи автоматического анализа текстов;

Основные подходы к решению задач: правила, написанные вручную и машинное обучение;

Показатели качества: точность, полнота, F-мера; state-of-the-art.

Тема 2. Базовая обработка текста и дистанция редактирования

Предобработка текста: токенизация и сегментация;

Нормализация слов: стеммеры, лемматизаторы, морфологические анализаторы;

Регулярные выражения; дистанция редактирования.

Тема 3. Языковые модели

N-граммы;

Перплексия;

Методы сглаживания;

Линейная интерполяция;

Применение языковых моделей: предсказание ввода, исправление ошибок правописания, распознавание речи, порождение текста.

Тема 4. Задачи разметки текста и скрытые марковские модели

Разметка по частям речи;

Извлечение именованных сущностей как задача разметки;

Скрытые марковские модели, их достоинства и недостатки;

Модификации скрытых марковских моделей.

Тема 5. Классификация текстов и анализ тональности

Задачи классификации; наивный байесовский классификатор;

Проблемы классификации текстов;

Анализ тональности;

Извлечение аспектов.

Тема 6. Информационный поиск

Векторные модели текстов;

Матричное представление;

Обратный индекс;

Фразовые запросы;

Ранжированный информационный поиск;

Коэффициент Жаккара;

Tf-idf;

Методы оценки поисковых машин.

Тема 7. Парсинг

Синтаксис составляющих и синтаксис зависимостей;

Контекстнезависимые грамматики;

Вероятностный подход к парсингу;

Лексикализованные вероятностные грамматики;

Алгоритм SKY;

Применение парсинга в различных задачах.

Тема 8. Машинный перевод

Классические подходы: пословный, трансферный, интерлингвальный;

Статистический машинный перевод;

Выравнивание текстов;

Оценка параметров в моделях IBM;

Фразовые модели;

Извлечение фразовых лексиконов; алгоритм декодирования.

Тема 9. Компьютерная семантика

Значение и смысл;

WordNet и аналогичные лексические базы данных;

Измерение семантической близости;

Тезаурусные методы;

Дистрибуционные (корпусные) методы.

Тема 10. Автоматическое реферирование

Экстрактивное и абстрактное реферирование;

Реферирование нескольких документов;

Реферирование, основанное на запросе;

Обучение с учителем и без учителя в контексте автоматического реферирования.

5.2. Учебно-тематический план

Очная форма обучения, очно-заочная форма обучения

| № п/п | Наименование тем (разделов) дисциплины | Трудоемкость в часах | | | | | Формы текущего контроля успеваемости |
|----------|---|----------------------|--|--------|---------------------------------------|-------------------------------|---|
| | | Всего | Контактная работа - Аудиторная работа | | | Самостояте льная работа | |
| | | | Общая, в т.ч.: | Лекции | Семинары, практическ ие занятия | | |
| 1 | Введение в автоматическую обработку естественного языка | 10 | 2 | 1 | 1 | 8 | Самостоятельн ые работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям. |
| 2 | Базовая обработка текста и дистанция редактирования | 10 | 3 | 1 | 2 | 7 | |
| 3 | Языковые модели | 12 | 2 | 1 | 1 | 10 | |
| 4 | Задачи разметки текста и скрытые марковские модели | 10 | 3 | 1 | 2 | 7 | |
| 5 | Классификация текстов и анализ тональности | 11 | 4 | 2 | 2 | 7 | |
| 6 | Информационный поиск | 11 | 4 | 2 | 2 | 7 | |

| | | | | | | | |
|-----|------------------------------|-----|----|----|----|----|--|
| 7 | Парсинг | 11 | 4 | 2 | 2 | 7 | |
| 8 | Машинный перевод | 11 | 4 | 2 | 2 | 7 | |
| 9. | Компьютерная семантика | 11 | 4 | 2 | 2 | 7 | |
| 10. | Автоматическое реферирование | 11 | 4 | 2 | 2 | 7 | |
| | В целом по дисциплине | 108 | 34 | 16 | 18 | 74 | Согласно учебному плану: контрольная работа |
| | Итого в % | | 31 | 47 | 53 | 69 | |

Институт онлайн-образования, заочная форма обучения

| № п/п | Наименование тем (разделов) дисциплины | Трудоемкость в часах | | | | | Формы текущего контроля успеваемости |
|----------|---|----------------------|--|--------|---------------------------------------|-------------------------------|---|
| | | Всего | Контактная работа - Аудиторная работа | | | Самостояте льная работа | |
| | | | Общ ая, в т.ч.: | Лекции | Семинары, практическ ие занятия | | |
| 1. | Введение в автоматическую обработку естественного языка | 15 | 1 | 1 | - | 14 | Самостоятельн ые работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям. |
| 2. | Базовая обработка текста и дистанция редактирования | 8 | 1 | 1 | - | 7 | |
| 3. | Языковые модели | 16 | 2 | 1 | 1 | 14 | |
| 4. | Задачи разметки текста и скрытые марковские модели | 10 | 2 | 1 | 1 | 8 | |
| 5. | Классификация текстов и анализ тональности | 10 | 1 | - | 1 | 9 | |
| 6. | Информационный поиск | 10 | 1 | - | 1 | 9 | |
| 7. | Парсинг | 10 | 1 | - | 1 | 9 | |
| 8. | Машинный перевод | 10 | 1 | - | 1 | 9 | |
| 9. | Компьютерная семантика | 9 | 1 | - | 1 | 8 | |
| 10. | Автоматическое реферирование | 10 | 1 | - | 1 | 9 | |
| | В целом по дисциплине | 108 | 12 | 4 | 8 | 96 | Согласно учебному |

| | | | | | | | |
|--|-----------|--|----|----|----|----|---------------------------------|
| | | | | | | | плану: контрольная работа |
| | Итого в % | | 11 | 33 | 67 | 89 | |

5.3. Содержание семинаров, практических занятий

| Наименование тем (разделов) дисциплины | Перечень вопросов для обсуждения на семинарских, практических занятиях, рекомендуемые источники из разделов 8,9 (указывается раздел и порядковый номер источника) | Формы проведения занятий |
|---|--|--|
| Введение в автоматическую обработку естественного языка | Структурные особенности текстов на естественном языке; Неоднозначность на всех уровнях языка; Основные задачи автоматического анализа текстов; Основные подходы к решению задач: правила, написанные вручную и машинное обучение; <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |
| Базовая обработка текста и дистанция редактирования | Предобработка текста: токенизация и сегментация; Регулярные выражения; дистанция редактирования <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |
| Языковые модели | N-граммы; Перплексия; Методы сглаживания; Линейная интерполяция; <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |
| Задачи разметки текста и скрытые марковские модели | Разметка по частям речи; Извлечение именованных сущностей как задача разметки; Скрытые марковские модели, их достоинства и недостатки <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |

| | | |
|--|--|---|
| Классификация текстов и анализ тональности | Задачи классификации; наивный байесовский классификатор; Проблемы классификации текстов; Анализ тональности <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |
| Информационный поиск | Векторные модели текстов; Матричное представление; Обратный индекс; Фразовые запросы; Ранжированный информационный поиск <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |
| Парсинг | Синтаксис составляющих и синтаксис зависимостей; Контекстнезависимые грамматики; Вероятностный подход к парсингу; Лексикализованные вероятностные грамматики <i>Рекомендуемые источники: п.8, [1]- [3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |
| Машинный перевод | Классические подходы: пословный, трансферный, интерлингвальный; Статистический машинный перевод; Модели IBM; Выравнивание текстов; <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |
| Компьютерная семантика | Значение и смысл; WordNet и аналогичные лексические базы данных; Измерение семантической близости <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |
| Автоматическое реферирование | Экстрактивное и абстрактное реферирование; Реферирование нескольких документов; Реферирование, основанное на запросе; Обучение с учителем и без учителя в контексте автоматического реферирования <i>Рекомендуемые источники: п.8, [1]-[3]</i> | Интерактивная форма, Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений |

6. Учебно-методическое обеспечение для самостоятельной работы обучающихся по дисциплине

6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы

| Наименование разделов, тем входящих в дисциплину | Перечень вопросов, отводимых на самостоятельное освоение | Формы внеаудиторной самостоятельной работы |
|---|---|--|
| Введение в автоматическую обработку естественного языка | Показатели качества: точность, полнота, F-мера; state-of-the-art. | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |
| Базовая обработка текста и дистанция редактирования | Нормализация слов: стеммеры, лемматизаторы, морфологические анализаторы; | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |
| Языковые модели | Применение языковых моделей: предсказание ввода, исправление ошибок правописания, распознавание речи, порождение текста | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |
| Задачи разметки текста и скрытые марковские модели | Модификации скрытых марковских моделей | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |
| Классификация текстов и анализ тональности | Извлечение аспектов | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |
| Информационный поиск | Коэффициент Жаккара; Tf-idf; Методы оценки поисковых машин. | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |

| | | |
|------------------------------|--|--|
| Парсинг | Алгоритм SKY; Применение парсинга в различных задачах. | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |
| Машинный перевод | Оценка параметров в моделях IBM; Фразовые модели; Извлечение фразовых лексиконов; алгоритм декодирования. | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |
| Компьютерная семантика | Тезаурусные методы; Дистрибуционные (корпусные) методы | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |
| Автоматическое реферирование | Оценка систем реферирования; ROUGE | Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию. |

6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю

Примерные вопросы контрольной работы

1. Структурные особенности текстов на естественном языке.
2. Извлечение именованных сущностей как задача разметки
3. Ранжированный информационный поиск
4. Вероятностный подход к парсингу
5. Синтаксис составляющих и синтаксис зависимостей
6. В чем разница между NLU и NLG?
7. Как НЛП связано с ИИ?
8. Каковы два основных типа анализа тональности?
9. Найти и аргументировать ошибку экспертной разметки 5 предложенных текстов (тексты заданы)
10. Составить схему разметки тональностей сообщений (по названиям

заголовков)

Примерные задания контрольной работы

1. Объясните и приведите пример, как регулярные выражения используются в NLP.
2. [Перечислите](#) 3 распространенных типа нормализации текста и о каждом подробно расскажите.
3. Приведите 5 примеров акустической неоднозначности.
4. Приведите пример простой токенизации в Python (1 строка, без знаков препинания).
5. Назовите 5-7 задач NLP.
6. Приведите 2-3 примера значимых функций для сегментации предложений
7. Для чего используются языковые модели?
8. Как вычислить вероятность предложения по триграмме языковой модели?
9. Приведите 3-4 конкретных примера задач поиска информации.
10. В сети интернет найти 5 сообщений о перспективах запуска «Северного потока - 2» и представить выбранные новости в структуре «Название» - «Аннотация» - «Содержание» - «Оценка тональности». Обосновать выбранную меру оценки тональности.

Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических рекомендациях Департамента анализа данных и машинного обучения Факультета информационных технологий и анализа больших данных.

7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы содержится в разделе **2. «Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине».**

**Типовые контрольные задания или иные материалы, необходимые
для оценки индикаторов достижения компетенций, умений и знаний**

| Наименование компетенции | Наименование индикаторов достижения компетенции | Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции | Типовые контрольные задания |
|---|--|--|--|
| ПКН-4 Способность проектировать и создавать интеллектуальные информационные системы, выбирать метод обучения в соответствии с анализом задачи | Демонстрирует знание основных понятий машинного обучения и интеллектуального анализа данных, понимание области и границ применимости, основные виды задач. | <p><u>Знать:</u> Студент должен знать основные понятия машинного обучения и интеллектуального анализа данных, такие как классификация, кластеризация, регрессия, нейронные сети, глубокое обучение и т.д. Он должен понимать область и границы применимости этих методов, а также основные виды задач, которые они могут решать.</p> <p><u>Уметь:</u> Студент должен уметь проектировать и создавать интеллектуальные информационные системы, выбирать метод обучения в соответствии с анализом задачи. Он должен уметь использовать различные инструменты и библиотеки для реализации этих систем, а также оценивать их эффективность и точность. Кроме того,</p> | Реализуйте алгоритм поиска расстояния редактирования между двумя строками. |

| | | | |
|--|---|--|---|
| | | студент должен уметь анализировать данные, выбирать подходящие методы и модели для их обработки и решения задачи. | |
| | Демонстрирует знание популярных инструментальных средств машинного обучения, собирает датасет, строит модели, проводит их анализ и диагностику, делает содержательные выводы. | <p><u>Знать:</u> Студент должен знать основные понятия и технологии в области обработки текстов на естественных языках, а также популярные инструменты машинного обучения, используемые в этой области, такие как TensorFlow, Keras, PyTorch, scikit-learn и другие. Он также должен быть знаком с методами сбора и подготовки данных, анализа и диагностики моделей, а также с методами оценки качества моделей.</p> <p><u>Уметь:</u> Студент должен уметь проектировать и создавать интеллектуальные информационные системы для обработки текстов на естественных языках, выбирать наиболее подходящий метод обучения для решения конкретной задачи, собирать и подготавливать данные, строить и настраивать модели, проводить анализ и диагностику моделей, а также делать содержательные</p> | Реализуйте на языке программирования Python набор правил для токенизации текста на слова. |

| | | | |
|--|---|---|--|
| | | <p>выводы на основе результатов анализа. Он должен также уметь работать с популярными инструментальными средствами машинного обучения и программирования, такими как Python, TensorFlow, Keras, PyTorch и другими.</p> | |
| | <p>Презентабельно демонстрирует результаты анализа данных и машинного обучения в форме, доступной непрофессионалу, структурирует отчет по проведенному анализу.</p> | <p><u>Знать:</u> Студент должен знать основы лингвистической обработки естественного языка, методы машинного обучения и анализа данных, а также принципы проектирования информационных систем.</p> <p><u>Уметь:</u> Студент должен уметь создавать и применять алгоритмы обработки текстов на естественных языках, выбирать и применять соответствующие методы машинного обучения и анализа данных для решения задач в области обработки текстов, проектировать и создавать интеллектуальные информационные системы. Также студент должен уметь структурировать и презентовать результаты анализа данных в доступной форме.</p> | <p>Решите задачу автоматического формирования списка стоп-слов по корпусу текстов.</p> |

Примерные вопросы для подготовки к экзамену / зачету

1. NLP как одна из ведущих областей искусственного интеллекта.
2. Естественный язык как объект автоматической обработки.
3. Популярные задачи NLP и общие подходы к их решению.
4. Предварительная обработка текста. Регулярные выражения.
5. Стеммеры, лемматизаторы, морфологические анализаторы.
6. N-граммы. Дистрибутивная гипотеза. Матрица совместной встречаемости.
7. Применение языковых моделей: предсказание ввода, исправление ошибок правописания.
8. Проблемы с языковыми моделями и их решения.
9. Проблемы с тегами; полезность автоматических аннотаций.
10. Скрытые марковские модели, их плюсы и минусы.
11. Классификация текстов: постановка задачи и методы.
12. Наивный байесовский классификатор. Проблемы с классификацией текста.
13. Анализ тональности, извлечение аспектов
14. Меры оценки системы NLP.
15. Поиск информации. Бинарный поиск.
16. Поиск информации. Фразовые запросы.
17. TF-IDF.
18. Лексические базы данных. WordNet – организация, специфика, применение.
19. Семантическое сходство. Недистрибутивные методы.
20. Семантическое сходство. Фразовые запросы. Косинусное расстояние / подобие.
21. Python как язык программирования и инструмент для написания проектов NLP.
22. Векторная модель word2vec.
23. Векторная модель BERT.
24. Машинное обучение в НЛП.
25. Обзор RapidMiner и Orange.

Пример экзаменационного билета

Федеральное государственное образовательное бюджетное учреждение
высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»**
(Финансовый университет)

Департамент анализа данных и машинного обучения

Дисциплина **Обработка текстов на естественных языках**

Факультет информационных технологий и анализа больших данных

Форма обучения: **очная**

Направление подготовки: **09.03.03 - Прикладная информатика**

Профиль: **Инженерия данных**

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ №

1. Скрытые марковские модели, их плюсы и минусы. **(20 баллов)**
2. Реализуйте функцию сегментации предложений. **(20 баллов)**
3. Решите задачу автоматического формирования списка стоп-слов по корпусу текстов. **(20 баллов)**

8.Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

[1] Иванищева, О. Н. Прикладная лингвистика : учебное пособие / О. Н. Иванищева. — Москва : Русайнс, 2021. — 235 с. — ЭБС BOOK.ru. — URL: <https://book.ru/book/942005> (дата обращения: 22.03.2023). — Текст : электронный.

[2] Влавацкая, М. В. Введение в языкознание : учебное пособие / М. В. Влавацкая. — Новосибирск : НГТУ, 2019. — 416 с. — ЭБС Лань. — URL: <https://e.lanbook.com/book/152389>; ЭБС Университетская библиотека online. - URL: <https://biblioclub.ru/index.php?page=book&id=575297> (дата обращения: 22.03.2023). — Текст : электронный.

[3] Махлина, С. Т. Лингвистика и семиотика : учебник и практикум для вузов / С. Т. Махлина. — Москва : Юрайт, 2023. — 260 с. — (Высшее образование). — ЭБС Юрайт. — URL: <https://urait.ru/bcode/519973> (дата обращения: 22.03.2023). — Текст :

электронный.

9.Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины:

1. Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/>
2. Электронно-библиотечная система BOOK.RU <http://www.book.ru>
3. Электронно-библиотечная система «Университетская библиотека ОНЛАЙН» <http://biblioclub.ru/>
4. Электронно-библиотечная система Znanium <http://www.znanium.com>
5. Электронно-библиотечная система издательства «ЮРАЙТ» <https://urait.ru/>
6. Электронно-библиотечная система издательства Проспект <http://ebs.prospekt.org/books>
7. Электронно-библиотечная система издательства Лань <https://e.lanbook.com/>
8. Деловая онлайн-библиотека Alpina Digital <http://lib.alpinadigital.ru/>
9. Электронная библиотека Издательского дома «Гребенников» <https://grebennikon.ru/>
10. Научная электронная библиотека eLibrary.ru <http://elibrary.ru>
11. Национальная электронная библиотека <http://нэб.рф/>
12. Финансовая справочная система «Финансовый директор» <http://www.1fd.ru/>

10.Методические указания для обучающихся по освоению дисциплины.

Основные этапы работы студента по дисциплине **Обработка текстов на естественных языках**

1. Предварительная ориентировка в подлежащем изучению учебном материале по программе.
2. Ознакомление с рекомендованной учебной литературой.
3. Слушание и конспектирование лекций, а также выполнение других видов учебной работы.
4. Планирование самостоятельной работы.

5. Обобщение и систематизация информации, почерпнутой из лекций и прочитанной литературы.
6. Выполнение контрольной работы.

Рекомендации по работе с учебным материалом:

1. Осознавайте наличный уровень полученных вами знаний.
2. В ситуации непонимания нужно выявить тот первичный уровень и факторы непонимания, которые стали препятствием понимания последующего.
3. Задавайте сами себе вопросы и пытайтесь ответить на них.

Рекомендации по работе на лекции и с лекционным материалом:

1. Основная задача на лекции – осмысление излагаемого в ней материала. Для этого необходимо слушать лекцию с самого начала, не упуская общих, ориентирующих в материале рассуждений и установок лектора.
2. Ведение записей на лекции важно и полезно для лучшего осмысливания материала, для сохранения информации, с целью ее дальнейшего использования.
3. Для облегчения записи рекомендуется применять сокращения повторяющихся терминов или хорошо известных понятий.

Рекомендации по работе с литературой:

1. Если возникли затруднения при разыскивании материала, по какому-либо конкретному вопросу, следует обратиться к предметному указателю, напечатанному, как правило, в конце каждого литературного источника.
2. Предметный указатель – это алфавитный список основных научных понятий (терминов), содержание которых раскрыто в книге, рядом с термином стоят числа, обозначающие номера страниц, на которых изложен материал, относящийся к данному понятию.

Рекомендации по выполнению контрольной работы:

1. Перед выполнением контрольной работы студент должен изучить соответствующие разделы учебной литературы.
2. Контрольную работу студент должен выполнять самостоятельно, используя те навыки и умения, которые получил на лекциях и практических занятиях.

3. При затруднениях, возникших при выполнении контрольной работы, студент может получить консультацию преподавателя.

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем

11.1. Комплект лицензионного программного обеспечения

1. Пакет офисных программ,
2. Антивирус Kaspersky.

11.2. Современные профессиональные базы данных и информационные справочные системы

1. Информационно-правовая система «Гарант»,
2. Информационно-правовая система «Консультант Плюс»,
3. Электронная энциклопедия: <http://ru.wikipedia.org/wiki/Wiki>,
4. Система комплексного раскрытия информации «СКРИН» - <http://www.skrin.ru/>

11.3. Сертифицированные программные и аппаратные средства защиты информации:

- не предусмотрены.

11.4. Язык программирования Python 3.8 (или старше).

11.5. Платформа для научных исследований, основанная на языке программирования Python, Anaconda, библиотека PyTorch.

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Наличие аудитории, оснащенной компьютерной техникой и проектором, с возможностью подключения к сети «Интернет».