

**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

СОГЛАСОВАНО

Газпромбанк

Начальник управления алгоритмов МО

_____ Дашковский В.А.

15.12.2022 г.

УТВЕРЖДАЮ

Проректор по учебной и
методической работе

_____ Е.А. Каменева

29.12.2022 г.

С.В. Макрушин, В.А. Малекова

Обработка текстов на естественных языках

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
01.03.02 Прикладная математика и информатика,
ОП «Прикладная математика и информатика»
(Анализ данных и принятие решений в экономике и финансах),
ОП «Прикладное машинное обучение» (Прикладное машинное обучение)

*Рекомендовано Ученым советом
Факультета информационных технологий и анализа больших данных
(протокол №27 от 15.12.2022г.)*

*Одобрено Советом учебно-научного
Департамента анализа данных и машинного обучения
(протокол №6 от 13.12.2022 г.)*

Москва 2022

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.....	8
6. Учебно-методическое обеспечение для самостоятельной работы обучающихся по дисциплине.....	10
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	10
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	17
9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.....	18
10. Методические указания для обучающихся по освоению дисциплины.	18
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем	20
12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.....	20

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
ОП «Прикладная математика и информатика» (Анализ данных и принятие решений в экономике и финансах)			
ПКП-5	Способность применять методы и инструменты анализа данных и машинного обучения при подготовке аналитического обоснования финансово-экономических решений	1.Осуществляет обоснованный выбор методов и инструментов обработки данных для решения экономических и финансовых задач.	Знать: методы обработки текстов на естественных языках. Уметь: выбирать методы обработки текстов на естественных языках.
		2.Владеет навыками решения прикладных задач с использованием методов и инструментов анализа данных и машинного обучения.	Знать: методы для решения прикладных задач методами обработки текстов на естественных языках. Уметь: использовать методы для решения прикладных задач методами обработки текстов на естественных языках.
ОПК-4	Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	1.Использует информационно-коммуникационные технологии при поиске необходимой информации, сбора, визуализации и обработки данных.	Знать: методы подготовки текстов на естественных языках для их обработки математическими методами. Уметь: применять методы подготовки текстов на естественных языках для их обработки математическими методами.
		2.Осуществляет рациональный выбор программного продукта в зависимости от поставленной задачи.	Знать: программные продукты и библиотеки для обработки текстов на естественных языках и знать их преимущества и недостатки. Уметь: выбирать программные продукты и библиотеки для обработки текстов на естественных языках.

		3. Владеет навыками обеспечения информационной безопасности автоматизированных систем.	Знать: методы тестирования алгоритмов для обработки текстов на естественных языках. Уметь: применять методы тестирования алгоритмов для обработки текстов на естественных языках.
ОП «Прикладное машинное обучение» (Прикладное машинное обучение)			
ПКП-1	Способность собирать наборы данных, в том числе больших данных, выполнять их подготовку для анализа в соответствии с решаемой прикладной задачей	1. Владеет навыками поиска внешних и внутренних источников данных для решения прикладной задачи.	Знать: Источники корпусов текстов для решения задач обработки текстов. Уметь: Собирать и готовить к обработке корпуса текстов.
		2. Использует инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников.	Знать: Инструменты для подготовки к обработке корпуса текстов на основе разнородных источников. Уметь: Использовать инструменты для подготовки к обработке корпуса текстов на основе разнородных источников.
		3. Владеет навыками обработки больших данных для решения прикладных задач машинного обучения.	Знать: Инструменты для обработки больших массивов текстовых данных. Уметь: Использовать инструменты для обработки больших массивов текстовых данных.
ПКП-5	Способность строить, обучать и оценивать качество моделей глубокого обучения в прикладных задачах	1. Демонстрирует знание принципов построения искусственных нейронных сетей, их обучения и оценки качества.	Знать: Принципы построения искусственных нейронных сетей. Уметь: Обучать искусственные нейронные сети.
		2. Демонстрирует знание современных архитектур искусственных нейронных сетей.	Знать: Современные архитектуры искусственных нейронных сетей. Уметь: Использовать современные архитектуры искусственных нейронных сетей.
		3. Использует современные инструментальные средства для построения	Знать: Современные архитектуры искусственных нейронных сетей для обработки текстов на естественном языке.

		искусственных нейронных сетей для решения прикладных задач.	Уметь: Использовать современные архитектуры искусственных нейронных сетей для обработки текстов на естественном языке.
--	--	---	---

3. Место дисциплины в структуре образовательных программ

Дисциплина «Обработка текстов на естественных языках» является дисциплиной Профиля «Анализ данных и принятие решений в экономике и финансах» по направлению подготовки 01.03.02 Прикладная математика и информатика, ОП «Прикладная математика и информатика.

Дисциплина «Обработка текстов на естественных языках» является дисциплиной Профиля «Прикладное машинное обучение» по направлению подготовки 01.03.02 Прикладная математика и информатика, ОП «Прикладное машинное обучение».

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)

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Семестр 5 / 6*
Общая трудоемкость дисциплины	4/144	144
Контактная работа – Аудиторные занятия	68	68
<i>Лекции</i>	16	16
<i>Семинары, практические занятия</i>	52	52
Самостоятельная работа	76	76
Вид текущего контроля	Контрольная работа	Контрольная работа
Вид промежуточной аттестации	Экзамен	Экзамен

* для ОП «Прикладное машинное обучение» (Прикладное машинное обучение)

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 CKY;

Применение парсинга в различных задачах.

Тема 8. Машинный перевод

Классические подходы: пословный, трансферный, интерлингвальный;

Статистический машинный перевод;

Модели IBM;

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.

Тема10. Автоматическое реферирование

Экстрактивное и абстрактное реферирование;

Реферирование нескольких документов;

Реферирование, основанное на запросе;

Обучение с учителем и без учителя в контексте автоматического реферирования;

Оценка систем реферирования; ROUGE.

5.2. Учебно-тематический план

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваемости
		Всего	Контактная работа - Аудиторная работа			Самостоя тельная работа	
			Общая, в т.ч.:	Лекции	Семинары, практическ ие занятия		
1.	Введение в автоматическую обработку естественного языка	9	5	1	4	4	Самостоятельн ые работы. Участие в решении задач на
2.	Базовая обработка текста и дистанция редактирования	11	5	1	4	6	
3.	Языковые модели	11	5	1	4	6	
4.	Задачи разметки текста и скрытые марковские модели	11	5	1	4	6	

5.	Классификация текстов и анализ тональности	17	8	2	6	9	практических занятиях. Собеседования по домашним заданиям.
6.	Информационный поиск	17	8	2	6	9	
7.	Парсинг	17	8	2	6	9	
8.	Машинный перевод	17	8	2	6	9	
9.	Компьютерная семантика	17	8	2	6	9	
10.	Автоматическое реферирование	17	8	2	6	9	
	В целом по дисциплине	144	68	16	52	76	Согласно учебному плану: контрольная работа
	Итого в %		47	24	76	53	

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
ОП «Прикладная математика и информатика» (Анализ данных и принятие решений в экономике)			
ПКП-5 Способность применять методы и инструменты анализа данных и машинного обучения при подготовке аналитического обоснования	1. Осуществляет обоснованный выбор методов и инструментов обработки данных для решения экономических и финансовых задач.	Знать: методы обработки текстов на естественных языках. Уметь: выбирать методы обработки текстов на естественных языках.	Решите задачу токенизации текста и удаления стоп слов.

финансово-экономических решений	2. Владеет навыками решения прикладных задач с использованием методов и инструментов анализа данных и машинного обучения.	<u>Знать:</u> методы для решения прикладных задач методами обработки текстов на естественных языках. <u>Уметь:</u> использовать методы для решения прикладных задач методами обработки текстов на естественных языках.	Решите задачу классификации коротких текстов по тональности (положительная / нейтральная / отрицательная).
ОПК-4 Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	1. Использует информационно-коммуникационные технологии при поиске необходимой информации, сбора, визуализации и обработки данных.	<u>Знать:</u> методы подготовки текстов на естественных языках для их обработки математическими методами. <u>Уметь:</u> применять методы подготовки текстов на естественных языках для их обработки математическими методами.	Проведите разбиение текста на предложения, токенизацию и лемматизацию текста.
	2. Осуществляет рациональный выбор программного продукта в зависимости от поставленной задачи.	<u>Знать:</u> программные продукты и библиотеки для обработки текстов на естественных языках и знать их преимущества и недостатки. <u>Уметь:</u> выбирать программные продукты и библиотеки для обработки текстов на естественных языках.	Предложите библиотеку для лемматизации русскоязычного текста, обоснуйте свой выбор.
	3. Владеет навыками обеспечения информационной безопасности автоматизированных систем.	<u>Знать:</u> методы тестирования алгоритмов для обработки текстов на естественных языках. <u>Уметь:</u> применять методы тестирования алгоритмов для обработки текстов на естественных языках.	Проведите тестирования алгоритма машинного обучения на переобучение.

ОП «Прикладное машинное обучение» (Прикладное машинное обучение)			
ПКП-1 Способность собирать наборы данных, в том числе больших данных, выполнять их подготовку для анализа в соответствии с решаемой прикладной задачей	1. Владеет навыками поиска внешних и внутренних источников данных для решения прикладной задачи.	Знать: Источники корпусов текстов для решения задач обработки текстов. Уметь: Собирать и готовить к обработке корпуса текстов.	Собрать корпус русскоязычных текстов объемом более 50 мегабайт.
	2. Использует инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников.	Знать: Инструменты для подготовки к обработке корпуса текстов на основе разнородных источников. Уметь: Использовать инструменты для подготовки к обработке корпуса текстов на основе разнородных источников.	Преобразовать корпус текстов на русском языке к файлу в виде последовательности слов в нормальной форме.
	3. Владеет навыками обработки больших данных для решения прикладных задач машинного обучения.	Знать: Инструменты для обработки больших массивов текстовых данных. Уметь: Использовать инструменты для обработки больших массивов текстовых данных.	Преобразовать крупный корпус текстов на русском языке к файлу в виде последовательности слов в нормальной форме.
ПКП-5 Способность строить, обучать и оценивать качество моделей глубокого обучения в прикладных задачах	1. Демонстрирует знание принципов построения искусственных нейронных сетей, их обучения и оценки качества.	Знать: Принципы построения искусственных нейронных сетей. Уметь: Обучать искусственные нейронные сети.	Постройте и обучите многослойную нейронную сеть различающую язык твита написанного в кодировке Unicode.
	2. Демонстрирует знание современных архитектур искусственных нейронных сетей.	Знать: Современные архитектуры искусственных нейронных сетей. Уметь: Использовать современные архитектуры искусственных нейронных сетей.	Постройте и обучите нейронную сеть на архитектуре LSTM искать аномалии в последовательности.
	3. Использует современные инструментальные средства для построения	Знать: Современные архитектуры искусственных нейронных сетей для обработки текстов на	Постройте и обучите нейронную сеть на архитектуре LSTM искать именованные сущности в тексте.

	искусственных нейронных сетей для решения прикладных задач.	естественном языке. Уметь: Использовать современные архитектуры искусственных нейронных сетей для обработки текстов на естественном языке.	
--	---	--	--

Примерные вопросы для подготовки к экзамену

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. Диструбитивная 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: **01.03.02 Прикладная математик и информатика**

Профиль: _____

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: 20.01.2023). — 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: 20.01.2023).

— 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: 20.01.2023). — 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.x в среде Windows.

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

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».