

Документ подписан простой электронной подписью
Информация о владельце:
ФИО: Галунин Сергей Александрович
Должность: проректор по учебной работе
Дата подписания: 20.03.2023 16:27:53
Уникальный программный ключ:
08ef34338325bdb0ac5a47baa5472ce36cc3fc3b

Приложение к ОПОП
«Алгоритмическая математика
в вычислениях и моделировании»



СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

МИНОБРНАУКИ РОССИИ

федеральное государственное автономное образовательное учреждение высшего образования
**«Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И.Ульянова (Ленина)»
(СПбГЭТУ «ЛЭТИ»)**

РАБОЧАЯ ПРОГРАММА

дисциплины

«ОБРАБОТКА ЕСТЕСТВЕННЫХ ЯЗЫКОВ»

для подготовки магистров

по направлению

01.04.02 «Прикладная математика и информатика»

по программе

«Алгоритмическая математика в вычислениях и моделировании»

Санкт-Петербург

2022

ЛИСТ СОГЛАСОВАНИЯ

Разработчики:

доцент, к.т.н. Посов И.А.

Рабочая программа рассмотрена и одобрена на заседании кафедры АМ
27.01.2022, протокол № 6

Рабочая программа рассмотрена и одобрена учебно-методической комиссией
ФКТИ, 24.02.2022, протокол № 2

Согласовано в ИС ИОТ

Начальник ОМОЛА Загороднюк О.В.

1 СТРУКТУРА ДИСЦИПЛИНЫ

Обеспечивающий факультет ФКТИ

Обеспечивающая кафедра АМ

Общая трудоемкость (ЗЕТ) 4

Курс 2

Семестр 3

Виды занятий

Лекции (академ. часов) 17

Практические занятия (академ. часов) 17

Иная контактная работа (академ. часов) 1

Все контактные часы (академ. часов) 35

Самостоятельная работа, включая часы на контроль
(академ. часов) 109

Всего (академ. часов) 144

Вид промежуточной аттестации

Дифф. зачет (курс) 2

2 АННОТАЦИЯ ДИСЦИПЛИНЫ

«ОБРАБОТКА ЕСТЕСТВЕННЫХ ЯЗЫКОВ»

В курсе рассматриваются задачи, которые требуют обработки текстов на естественных языках, в первую очередь русском и английском. Список задач включает в себя классификацию текстов, определение тональности, автоматическое реферирование, машинный перевод, многие другие задачи более низкого уровня. Из подходов к решению задач рассматриваются лингвистические подходы, статистические и подходы, использующие глубокое обучение. Курс предполагает решение практических заданий с помощью библиотек и ресурсов для обработки естественных языков для языка программирования python.

SUBJECT SUMMARY

«NATURAL LANGUAGE PROCESSING»

The course is devoted to the approaches to solve problems, that require processing of raw texts in natural languages, primarily Russian and English. The list of problems includes texts classification, sentiment analysis, automatic summarization, machine translation, a number of other low level tasks. The approaches being discussed include purely linguistic approaches, statistical approaches and deep learning approaches. The course supposes practice in coding in python using natural language processing libraries and resources.

3 ОБЩИЕ ПОЛОЖЕНИЯ

3.1 Цели и задачи дисциплины

1. Цель дисциплины -изучение математического аппарата, используемого в основе методов обработки естественных языков, и программных инструментов для обработки естественных языков и приобретение практических навыков в профессиональной деятельности.
2. Задачи дисциплины:
 - изучение прикладных проблем, которые возникают при обработке текстов на естественных языках и подходы к их решению;
 - освоение классификации текстов, определение тональности, автоматическое реферирование, машинный перевод, многие другие задачи более низкого уровня;
 - приобретение навыков решения практических заданий с помощью библиотек и ресурсов для обработки естественных языков для языка программирования python.
3. В результате освоения дисциплины у студента должно быть сформировано знание спектра подходов к решению разных задач обработки естественных языков, методов обработки естественных языков, программных инструментов, используемых при различных подходах и методах.
4. В результате изучения дисциплины студенты должны овладеть умением связать решаемую задачу с задачами из области обработки естественных языков, выбрать подход к решению задачи.
5. В результате изучения дисциплины у студентов должны сформироваться навыки использования математического аппарата и программных инструментов для решения проблематики обработки естественных языков.

3.2 Место дисциплины в структуре ОПОП

Дисциплина изучается на основе ранее освоенных дисциплин учебного плана:

1. «Статистика случайных процессов»
2. «Структуры данных и алгоритмы»
3. «Алгоритмы компьютерной математики»
4. «Математические методы распознавания образов»

и обеспечивает изучение последующих дисциплин:

1. «Производственная практика (преддипломная практика)»

3.3 Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения образовательной программы обучающийся должен достичь следующие результаты обучения по дисциплине:

Код компетенции/ индикатора компетенции	Наименование компетенции/индикатора компетенции
ОПК-2	Способен совершенствовать и реализовывать новые математические методы решения прикладных задач
<i>ОПК-2.1</i>	<i>Знает современные математические методы решения прикладных задач</i>
<i>ОПК-2.2</i>	<i>Умеет обосновывать выбор либо необходимость реализации новых математических методов решения прикладных задач</i>
<i>ОПК-2.3</i>	<i>Знает принципы и основные современные методы решения задач управления в технических системах</i>

4 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Содержание разделов дисциплины

4.1.1 Наименование тем и часы на все виды нагрузки

№ п/п	Наименование темы дисциплины	Лек, ач	Пр, ач	ИКР, ач	СР, ач
1	Введение	2	2		12
2	Задача классификации	3	3		18
3	Работа с последовательностями	3	3		21
4	Теория формальных языков в приложении к естественным	2	2		12
5	Семантика	3	3		18
6	Трансформеры последовательностей	3	3		22
7	Заключение	1	1	1	6
	Итого, ач	17	17	1	109
	Из них ач на контроль	0	0	0	0
	Общая трудоемкость освоения, ач/зе				144/4

4.1.2 Содержание

№ п/п	Наименование темы дисциплины	Содержание
1	Введение	Задачи обработки естественного языка, этапы обработки текста.
2	Задача классификации	Наивный байесовый классификатор и другие традиционные методы классификации. Классификация с помощью нейронных сетей прямого распространения. Лингвистические приложения методов классификации.
3	Работа с последовательностями	Текст как последовательность символов или слов. Н-граммы модели, модели основанные на рекуррентных нейронных сетях. Сглаживание, оценка модели. Классификация элементов последовательности.
4	Теория формальных языков в приложении к естественным	Контекстно-свободные грамматики в описании естественных языков, дерево разбора текста и граф зависимостей.
5	Семантика	Разные подходы к определению понятия смысла языковых единиц: исчисление предикатов, дистрибутивная семантика. Кластеризация. Нейросетевые подходы к векторному представлению слов (word embedding). Разрешение кореферентности.

№ п/п	Наименование темы дисциплины	Содержание
6	Трансформеры последовательностей	Понятие трансформера, энкодера, декодера, механизм внимания, современные нейросетевые модели. Машинный перевод, генерация текста.
7	Заключение	Обзор других современных направлений в обработке естественных языков.

4.2 Перечень лабораторных работ

Лабораторные работы не предусмотрены.

4.3 Перечень практических занятий

Наименование практических занятий	Количество ауд. часов
1. Определение авторства текста	1
2. N-грамм модели	2
3. Определение частей речи в тексте	2
4. Грамматический разбор текста	2
5. Поиск коллокаций	2
6. Word2Vec, кластеризация слов	2
7. BERT для анализа настроения	2
8. Выделение именованных сущностей	2
9. GPT2 для генерации текста	2
Итого	17

4.4 Курсовое проектирование

Курсовая работа (проект) не предусмотрены.

4.5 Реферат

Реферат не предусмотрен.

4.6 Индивидуальное домашнее задание

В процессе обучения по дисциплине «Статистика случайных процессов» студент обязан выполнить индивидуальные домашние задания (ИДЗ). Задачи ИДЗ охватывают весь спектр тематики, рассматриваемой в семестре, и содержит за-

дания на:

- **ИДЗ №1** Определение частей речи в тексте;
- **ИДЗ №2** Грамматический разбор текста;
- **ИДЗ №3** Word2Vec, кластеризация слов.

Требования по оформлению ИДЗ:

- Формат оформления: произвольный печатный формат . При оформлении ИДЗ следует использовать текстовые редакторы, электронные таблицы, результаты расчетов в математическом пакете следует вставлять в отчет в виде копии экрана.
- При оформлении ИДЗ рекомендуется использовать стандартные шрифты редакторов (например, Times New Roman, Calibri, Arial); размер шрифта 12-14 пунктов, межстрочный интервал 1,15-1,5 пунктов. Каждую задачу следует оформлять на новом листе.
- Таблицы и рисунки следует оформлять, придерживаясь сквозного просмотра. Т.е. если в задаче предусмотрена таблица или рисунок, то они должны быть приведены внутри или в конце решаемой задачи. Общее приложения для рисунков и таблиц не предусматривается.
- Объем ИДЗ зависит только от количества задач и/или заданий. Каждая задача должна содержать исходные данные, решение и ответ.
- Количество используемых источников не ограничено, решение должно производиться в одном из математических пакетов.
- Каждое ИДЗ состоит из: титульного листа (название дисциплины, ФИО, звание преподавателя, номер группы, ФИО студента, номер варианта, дата сдачи работы) списка решенных задач и/или заданий, списка используемых источников.
- Формат сдачи работы зависит от общих требований Университета (при очном обучении - ИДЗ сдается преподавателю в письменном виде или печатном виде; при дистанционном обучении - в печатном или электрон-

ном виде работы размещается в Moodle или отправляются преподавателю на электронную почту).

ИДЗ должны быть решены и представлены на проверку в установленное преподавателем время.

4.7 Доклад

Доклад не предусмотрен.

4.8 Кейс

Кейс не предусмотрен.

4.9 Организация и учебно-методическое обеспечение самостоятельной работы

Изучение дисциплины сопровождается самостоятельной работой студентов с рекомендованными преподавателем литературными источниками и информационными ресурсами сети Интернет.

Планирование времени для изучения дисциплины осуществляется на весь период обучения, предусматривая при этом регулярное повторение пройденного материала. Обучающимся, в рамках внеаудиторной самостоятельной работы, необходимо регулярно дополнять сведениями из литературных источников материал, законспектированный на лекциях. При этом на основе изучения рекомендованной литературы целесообразно составить конспект основных положений, терминов и определений, необходимых для освоения разделов учебной дисциплины.

Особое место уделяется консультированию, как одной из форм обучения и контроля самостоятельной работы. Консультирование предполагает особым образом организованное взаимодействие между преподавателем и студентами,

при этом предполагается, что консультант либо знает готовое решение, которое он может предписать консультируемому, либо он владеет способами деятельности, которые указывают путь решения проблемы.

Текущая СРС	Примерная трудоемкость, ач
Работа с лекционным материалом, с учебной литературой	34
Опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	0
Самостоятельное изучение разделов дисциплины	0
Выполнение домашних заданий, домашних контрольных работ	20
Подготовка к лабораторным работам, к практическим и семинарским занятиям	20
Подготовка к контрольным работам, коллоквиумам	0
Выполнение расчетно-графических работ	0
Выполнение курсового проекта или курсовой работы	0
Поиск, изучение и презентация информации по заданной проблеме, анализ научных публикаций по заданной теме	0
Работа над междисциплинарным проектом	0
Анализ данных по заданной теме, выполнение расчетов, составление схем и моделей, на основе собранных данных	0
Подготовка к зачету, дифференцированному зачету, экзамену	35
ИТОГО СРС	109

5 Учебно-методическое обеспечение дисциплины

5.1 Перечень основной и дополнительной литературы, необходимой для освоения дисциплины

№ п/п	Название, библиографическое описание	К-во экз. в библ.
Основная литература		
1	Гольдберг Й. Нейросетевые методы в обработке естественного языка [Электронный ресурс] : руководство / Й. Гольдберг, 2019. -282 с.	неогр.
2	Лейн Хобсон Обработка естественного языка в действии [Электронный ресурс] / Хобсон Лейн, Ханнес Хапке, Коул Ховард, 2021. -576 с.	неогр.
Дополнительная литература		
1	Ганегедара Т. Обработка естественного языка с TensorFlow [Электронный ресурс] : руководство / Т. Ганегедара, 2020. -382 с.	неогр.
2	Риз Р. Обработка естественного языка на Java [Электронный ресурс], 2016. -264 с.	неогр.

5.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет», используемых при освоении дисциплины

№ п/п	Электронный адрес
1	Каталог ресурсов для обработки естественного языка https://nlpub.ru/
2	Открытый корпус http://opencorpora.org/
3	Фреймворк обработки естественных языков для языка программирования Python https://www.nltk.org/
4	Конспект курса по обработке естественных языков Якова Эйзенштейна https://raw.githubusercontent.com/jacobeisenstein/gt-nlp-class/master/notes/eisenstein-nlp-notes.pdf
5	Mining of Massive Datasets http://mmds.org/

5.3 Адрес сайта курса

Адрес сайта курса: <https://vec.etu.ru/moodle/course/view.php?id=7330>

6 Критерии оценивания и оценочные материалы

6.1 Критерии оценивания

Для дисциплины «Обработка естественных языков» предусмотрены следующие формы промежуточной аттестации: зачет с оценкой.

Зачет с оценкой

Оценка	Описание
Неудовлетворительно	Курс не освоен. Студент испытывает серьезные трудности при ответе на ключевые вопросы дисциплины
Удовлетворительно	Студент в целом овладел курсом, но некоторые разделы освоены на уровне определений и формулировок теорем
Хорошо	Студент овладел курсом, но в отдельных вопросах испытывает затруднения. Умеет решать задачи
Отлично	Студент демонстрирует полное овладение курсом, способен применять полученные знания при решении конкретных задач.

Особенности допуска

Допуск к дифференцированному зачету студент получает по результатам текущего контроля, который включает в себя контроль посещаемости (не менее 80 % занятий), выполнение 3-х ИДЗ и сдаче 5 коллоквиумов. Оценка дифференцированного зачета полностью базируется на результатах текущего контроля. Процент выполнения практических работ приравнивается к количеству баллов. Зачет с оценкой проводится в форме собеседования по вопросам п. 6.2.

6.2 Оценочные материалы для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине

Вопросы к дифф.зачету

№ п/п	Описание
1	Задачи обработки естественного языка, этапы обработки текста
2	Задачи классификации
3	Наивный байесовский классификатор и другие традиционные методы классификации
4	Классификация с помощью нейронных сетей прямого распространения
5	Лингвистические приложения методов классификации
6	Основные принципы работы с последовательностями
7	Текст как последовательность символов или слов
8	N-грамм модели, модели основанные на рекуррентных нейронных сетях
9	Сглаживание, оценка модели
10	Классификация элементов последовательности
11	Определение и задачи теории формальных языков в приложении к естественным языкам
12	Контекстно-свободные грамматики в описании естественных языков, дерево разбора текста и граф зависимостей
13	Понятие и сущность семантики
14	Разные подходы к определению понятия смысла языковых единиц: исчисление предикатов, дистрибутивная семантика
15	Кластеризация
16	Нейросетевые подходы к векторному представлению слов (word embedding)
17	Разрешение кореферентности
18	Понятие и виды трансформеров последовательностей
19	Понятие трансформера, энкодера, декодера, механизм внимания, современные нейросетевые модели
20	Машинный перевод, генерация текста
21	Современные направления в обработке естественных языков

Весь комплект контрольно-измерительных материалов для проверки сформированности компетенции (индикатора компетенции) размещен в закрытой части по адресу, указанному в п. 5.3

6.3 График текущего контроля успеваемости

Неделя	Темы занятий	Вид контроля
1	Работа с последовательностями	
2		ИДЗ / ИДРГЗ / ИДРЗ
3	Задача классификации	
4		Коллоквиум
5	Теория формальных языков в приложении к естественным	
6		ИДЗ / ИДРГЗ / ИДРЗ
7	Работа с последовательностями	
8		Коллоквиум
9	Теория формальных языков в приложении к естественным	
10		Коллоквиум
11	Семантика	
12		ИДЗ / ИДРГЗ / ИДРЗ
13	Семантика	
14		Коллоквиум
15	Трансформеры последовательностей	
16		Коллоквиум

6.4 Методика текущего контроля

на лекционных занятиях текущий контроль включает в себя:

- контроль посещаемости (не более 20% (баллов) от общего объема оценивания текущей аттестации);

на практических занятиях текущий контроль включает в себя:

- контроль посещаемости (не более 20% (баллов) от общего объема оценивания текущей аттестации);
- контроль активности студентов. В ходе проведения практических занятий происходит привлечение студентов к активному участию в дискуссиях, решении задач, обсуждениях и т. д. При этом активность студентов учитывается преподавателем, как один из параметров текущего контроля на практических занятиях (не более 5% (баллов) от общего объема оценивания текущей аттестации);
- распределенный коллоквиум - 5 коллоквиумов по тематике дисциплины

(не более 50% (баллов) от общего объема оценивания текущей аттестации). Для допуска к дифф. зачету студенту необходимо решить задачи, выданные в течение семестра. Из каждой темы курса должна быть решена хотя бы одна задача. Список задач формируется на основе актуальной проблематики, связанной с обработкой естественных языков, и обновляется каждый семестр. Коллоквиум проводится на основе вопросов к дифф. зачету, изученных до момента проведения коллоквиума.

Критерии оценивания ответов:

- ответ дан без ошибок, обоснован теоретически и проиллюстрирован примерами - оценка "отлично";
- ответ дан без ошибок, проиллюстрирован примерами, но обоснования не всегда полны - оценка "хорошо";
- ответ дан без ошибок, проиллюстрирован примерами, но не все обоснования приведены корректно - оценка "удовлетворительно";
- в ответе есть ошибки, либо студент не видит связи между приводимыми формулами и утверждениями, не понимает их смысла оценка "неудовлетворительно".

самостоятельной работы студентов

Контроль самостоятельной работы студентов осуществляется на лекционных и практических занятиях студентов по методикам, описанным выше.

По результатам текущего контроля (выполнения всех параметров **более чем на 60 %** (баллов)) студент получает допуск на дифф. зачет.

7 Описание информационных технологий и материально-технической базы

Тип занятий	Тип помещения	Требования к помещению	Требования к программному обеспечению
Лекция	Лекционная аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, ПК и проектор, экран, меловая или маркерная доска.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Практические занятия	Аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, ПК и проектор, экран, меловая или маркерная доска.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Самостоятельная работа	Помещение для самостоятельной работы	Оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.	1) Windows 7 и выше или дистрибутив Linux, основанный на Ubuntu или Fedora; 2) Microsoft Office 2007 и выше или Libre Office 6.0 и выше.

8 Адаптация рабочей программы для лиц с ОВЗ

Адаптированная программа разрабатывается при наличии заявления со стороны обучающегося (родителей, законных представителей) и медицинских показаний (рекомендациями психолого-медико-педагогической комиссии). Для инвалидов адаптированная образовательная программа разрабатывается в соответствии с индивидуальной программой реабилитации.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Дата	Изменение	Дата и номер протокола заседания УМК	Автор	Начальник ОМОЛА