МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Российский государственный гуманитарный университет»
(ФГБОУ ВО «РГГУ»)

ИНСТИТУТ ЛИНГВИСТИКИ
Учебно-научный центр компьютерной лингвистики

Б1.В.ДЭ.06.01 Обработка естественного языка на Python

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

45.04.03 Фундаментальная и прикладная лингвистика
Код и наименование направления подготовки/специальности

Фундаментальная и компьютерная лингвистика
Наименование направленности (профиля)/ специализации

Уровень высшего образования: магистратура

Форма обучения: очная

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2023

Обработка естественного языка на Python
Рабочая программа дисциплины

Составитель(и):

Старший преподаватель А.М.Ивойлова

Ответственный редактор:

К.ф.н, доцент Н.А.Коротаев

УТВЕРЖДЕНО
Протокол заседания УНЦ компьютерной лингвистики
№ 6 от 12 апреля 2023 г.

ОГЛАВЛЕНИЕ

1. Пояснительная записка.....
1.1. Цель и задачи дисциплины .....
1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций .....
1.3. Место дисциплины в структуре образовательной программы .....
2. Структура дисциплины.....
3. Содержание дисциплины.....
4. Образовательные технологии .....
5. Оценка планируемых результатов обучения.....
5.1 Система оценивания .....
5.2 Критерии выставления оценки по дисциплине.....
5.3 Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине .....
6. Учебно-методическое и информационное обеспечение дисциплины .....
6.1 Список источников и литературы .....
6.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет». .....
6.3 Профессиональные базы данных и информационно-справочные системы.....
7. Материально-техническое обеспечение дисциплины .....
8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов.....
9. Методические материалы.....
9.1 Планы семинарских/ практических/ лабораторных занятий.....
9.2 Методические рекомендации по подготовке письменных работ .....
9.3 Иные материалы.....

1. Пояснительная записка

1.1. **Цель и задачи дисциплины**

Предметом дисциплины является изучение основных методов машинного обучения для обработки и классификации текстов. В курсе рассматриваются как математические основы методов машинного обучения и статистического анализа данных, так и детали их практического применения, в частности, подробно изучается библиотека scikit-learn, содержащая реализацию основных алгоритмов машинного обучения на языке Python. Особое внимание уделяется использованию методов машинного обучения при классификации текстов, а также в других задачах лингвистического анализа.

Курс направлен на решение следующих задач:

- познакомить обучающихся с основными алгоритмами машинного обучения, применяемыми для решения лингвистических задач, а также с программными модулями языка Python, реализующими данные методы;
- познакомить магистрантов с основными задачами текстовой классификации (жанровая, тематическая, анализ тональности и т. д.) и кластеризации, а также с используемыми в них алгоритмами машинного обучения;
- познакомить магистрантов с математическими методами, лежащими в основе алгоритмов машинного обучения;
- научить магистрантов как предварительно выбирать алгоритм машинного обучения для решения для прикладных лингвистических задач, так и дорабатывать выбранный алгоритм в зависимости от специфики задачи и исходных данных;
- научит магистрантов квалифицированно подбирать признаковое представление данных для алгоритмов машинного обучения, отражающее лингвистическую специфику задачи.
- научить магистрантов анализировать результаты применения статистических алгоритмов к лингвистическим данным;
- у магистрантов знания, позволяющие им квалифицированно читать литературу по специальности, включающую в себя как учебные материалы и научные статьи, так и более специализированные технические материалы, например, программную документацию.

1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Компетенция	Индикаторы компетенций	Результаты обучения
ПК-3 Способен использовать лингвистические технологии для проектирования систем автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистических компонентов интеллектуальных и информационных электронных систем	ПК-3.2 Имеет практический опыт работы с системами автоматической обработки текста и звучащей речи; проектирования модулей таких систем	Знать: – принципы работы лингвистически ориентированных программных продуктов; Уметь: – пользоваться лингвистически ориентированными программными продуктами;

		Владеть: – навыками использования лингвистически ориентированных программных продуктов.

1.3. Место дисциплины в структуре образовательной программы

Дисциплина «Обработка естественного языка на Python» является элективной дисциплиной и относится к части, формируемой участниками образовательных отношений.

Для освоения дисциплины необходимы знания, умения и владения, сформированные в ходе изучения следующих дисциплин и прохождения практик: Основы языка программирования Python, Инструменты лингвистического анализа в Python.

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 3 з.е., 108 академических часов.

Структура дисциплины для очной формы обучения

Объем дисциплины в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Семестр	Тип учебных занятий	Количество часов
3	Практические занятия	30
Всего:		30

Объем дисциплины (модуля) в форме самостоятельной работы обучающихся составляет 78 академических часов.

3. Содержание дисциплины

№ п/п	Наименование раздела дисциплины	Содержание
1.	Введение. Основные задачи машинного обучения, их применение в компьютерной лингвистике.	• Общая постановка задачи машинного обучения, обучение с учителем и без учителя; • Примеры задач машинного обучения; • Основные задачи компьютерной лингвистики. Их интерпретация как задач машинного обучения.
2.	Задачи классификации в обработке естественного языка и компьютерной лингвистике. Признаковое описание объектов, типы признаков	• Общая постановка задачи классификации. Типы признаков: двоичные, категориальные, вещественные, порядковые; • Двоичная и многоклассовая классификация; • Задачи компьютерной лингвистики как задачи классификации; Типы признаков в различных задачах, их извлечение из текста.
3.	Векторная модель классификации,	• Векторное представление текста в задачах

	матрица объекты-признаки, линейные классификаторы.	компьютерной лингвистики, модель мешка слов; • Матрица объекты-признаки; • Сведение задачи классификации к подбору оптимальной разделяющей плоскости; Свойства задач текстовой классификации (большое количество признаков, разреженность).
4.	Библиотека scikit-learn для машинного обучения. Основные объекты и типы данных.	• Решение задач машинного обучение на языке Python; • Представление данных, модуль Numpy, массивы и матрицы; • Операции с матрицами и векторами в модуле Numpy; • Библиотека scikit-learn, основные классы и методы; • Разреженные матрицы и модуль scipy.
5.	Наивный байесовский классификатор, вероятностная модель. Его недостатки и достоинства.	• Вероятностная модель классификации, наивный байесовский классификатор. Его связь с моделью мешка слов; • Достоинства и недостатки наивного байесовского классификатора; • Проблема нулевых вероятностей, сглаживание признаков; • Реализация наивного байесовского классификатора в библиотеке scikit-learn.
6.	Линейные классификаторы в библиотеке scikit-learn, основные алгоритмы.	• Понятие линейного классификатора, случай двух и более классов; • Решающее правило классификатора, вектор весов, разделяющая гиперплоскость; • Основные алгоритмы нахождения вектора весов: метод опорных векторов, логистическая регрессия.
7.	Извлечение признаков из текста, меры качества признаков. Оценка качества классификации.	• Стандартные признаки в задачах классификации: подсчёт количество слов, символьные энграммы; • Зависимость качества классификации от используемых признаков. Меры оценки качества (точность, полнота, F-мера); • Обучающая и контрольная выборка, скользящий контроль, переобучение; • Отбор признаков, основные методы оценки качества признака (вес признака, вероятность класса).
8.	Автоматическая классификация текстов, модель мешка слов.	• Стандартные приёмы в задачах текстовой классификации
9.	Обучение без учителя, кластеризация, стандартные алгоритмы: иерархическая кластеризация и метод средних.	• Понятие обучения без учителя, постановка задачи кластеризации; • Методы измерения расстояния между объектами, их зависимость от задачи; • Иерархическая кластеризация, вычисление

		межкластерного расстояния; • Метод k-средних, его сравнение с иерархической кластеризацией; • Применения кластеризации.
10.	Дистрибутивная семантика, семантические вектора, их применение в различных задачах.	• Дистрибутивная семантика, основные подходы; • Матрица совместных вхождений; • Понижение размерности семантических векторов.
11.	Дистрибутивная семантика на основе нейронных сетей. Обзор её применений.	• Общее знакомство с нейронными сетями; • Методы skipgram и CBOW для получения дистрибутивных векторов, модель Т. Миколова; • Свойства дистрибутивных векторов; Применение дистрибутивных векторов в задачах компьютерной лингвистики.
12.	Обзор применений машинного обучения в различных задачах компьютерной лингвистики (морфология, синтаксис и т.д.)	• Применение методов машинного обучения в различных задачах компьютерной лингвистики; • Автоматический морфологический анализ, определение грамматической категории слова на основе признаков; • Автоматический синтаксический анализ, использование машинного обучения для снятия неоднозначности; Переранжирование гипотез.

4. Образовательные технологии

Для проведения учебных занятий по дисциплине используются различные образовательные технологии. Для организации учебного процесса может быть использовано электронное обучение и (или) дистанционные образовательные технологии.

5. Оценка планируемых результатов обучения

5.1 Система оценивания

Форма контроля	**Макс. количество баллов**	
	За одну работу	**Всего**
Текущий контроль:		
- домашние задания	*5 баллов*	*30 баллов*
- выполнение заданий на семинаре	*5 баллов*	*10 баллов*
- участие в соревновании	*20 баллов*	*20 баллов*
Промежуточная аттестация – зачет		40 баллов
Итого за семестр		100 баллов

Полученный совокупный результат конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

100-балльная шкала	Традиционная шкала		Шкала ECTS
95 – 100	отлично	зачтено	A
83 – 94			B
68 – 82	хорошо		C
56 – 67	удовлетворительно		D
50 – 55			E
20 – 49	неудовлетворительно	не зачтено	FX
0 – 19			F

5.2 Критерии выставления оценки по дисциплине

Баллы/ Шкала ECTS	**Оценка по дисциплине**	**Критерии оценки результатов обучения по дисциплине**
100-83/ A,B	отлично/ зачтено	Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения. Свободно ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «высокий».
82-68/ C	хорошо/ зачтено	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».
67-50/ D,E	удовлетво-рительно/ зачтено	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «достаточный».
49-0/ F,FX	неудовлет-ворительно/ не зачтено	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами.

Баллы/ Шкала ECTS	Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
		Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

5.3 Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

В качестве домашних заданий предлагаются задания следующих типов

ДЗ1. Подбор признакового описания для выбранной задачи компьютерной лингвистики.
ДЗ2. Знакомство с библиотекой scikit-learn.
ДЗ3. Тестирование наивного байесовского классификатора на модельной задаче.
ДЗ4. Исследовательский проект (часть 1): автоматическое определение языка коротких текстов.
ДЗ5. Исследовательский проект (часть 2): автоматическое определение языка коротких текстов.
ДЗ6. Классификация текстов по тематическим категориям.
Проверка и обсуждение исследовательского проекта и ДЗ6.
ДЗ7. Кластеризация «вручную» набора данных различными методами.
ДЗ8. Применение семантических векторов для автоматической классификации.
ДЗ9. Применение машинного обучения в автоматическом морфологическом анализе.

Экзамен ориентирован на следующие контрольные вопросы

Основные типы задач машинного обучения.
Интерпретация задач компьютерной лингвистики в терминах машинного обучения.
Признаковое описание, типы признаков, их зависимость от задачи.
Наивный байесовский классификатор.
Линейные алгоритмы классификации.
Оценка качества классификации, основные меры.
Стандартное признаковое описание в задачах текстовой классификации.
Постановка задачи кластеризации, основные алгоритмы.
Дистрибутивная семантика, применение нейронных сетей.
Машинное обучение в задачах компьютерной морфологии и синтаксиса.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1 Список источников и литературы

Основная литература

1. Бахвалов Ю.Н., Малыгин Л.Л., Черкас П.С. Метод машинного обучения на основе алгоритма многомерной интерполяции и аппроксимации случайных функций. // Вестник Череповецкого государственного университета, 2012. – 4 стр.
2. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. – О
3. Reilly Media, 2017 – 340 с.
4. Sebastani F. machine Learning in Automated Text Categorization. 2001 – 77 p.
5. Scikit-learn, библиотека алгоритмов машинного обучения. http://scikit-learn.org/stable/
6. Ресурс по распознаванию данных и машинному обучению. http://machinelearning.ru

Рекомендованная литература

1. К. В. Воронцов. Лекции по машинному обучению. http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_(%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2)
2. Прикладная и компьютерная лингвистика, Под. ред. А. В. Митрениной. М., УРСС, 2016.
3. Manning C. D. et al. Foundations of statistical natural language processing. – Cambridge : MIT press, 1999
4. Bishop C. M. Pattern recognition //Machine Learning. – 2006

6.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

№ п/п	Наименование
1	Международные реферативные наукометрические БД, доступные в рамках национальной подписки в 2020 г. Web of Science Scopus
2	Профессиональные полнотекстовые БД, доступные в рамках национальной подписки в 2020 г. Журналы Cambridge University Press ProQuest Dissertation & Theses Global SAGE Journals Журналы Taylor and Francis
3	Профессиональные полнотекстовые БД JSTOR Издания по общественным и гуманитарным наукам Электронная библиотека Grebennikon.ru

6.3 Профессиональные базы данных и информационно-справочные системы

Доступ к профессиональным базам данных: https://liber.rsuh.ru/ru/bases
Scikit-learn, библиотека алгоритмов машинного обучения. http://scikit-learn.org/stable/
Ресурс по распознаванию данных и машинному обучению. http://machinelearning.ru

7. Материально-техническое обеспечение дисциплины

№п/п	Наименование ПО	Производитель	Способ распространения (*лицензионное или свободно распространяемое*)
1	Microsoft Share Point 2010	Microsoft	лицензионное
2	Windows 10 Pro	Microsoft	лицензионное
3	Kaspersky Endpoint Security	Kaspersky	лицензионное
4	Microsoft Office 2016	Microsoft	лицензионное
5	Zoom	Zoom	лицензионное

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

- для слепых и слабовидящих: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением или могут быть заменены устным ответом; обеспечивается индивидуальное равномерное освещение не менее 300 люкс; для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств; письменные задания оформляются увеличенным шрифтом; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.
- для глухих и слабослышащих: лекции оформляются в виде электронного документа, либо предоставляется звукоусиливающая аппаратура индивидуального пользования; письменные задания выполняются на компьютере в письменной форме; экзамен и зачёт проводятся в письменной форме на компьютере; возможно проведение в форме тестирования.
- для лиц с нарушениями опорно-двигательного аппарата: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

При проведении процедуры оценивания результатов обучения предусматривается использование технических средств, необходимых в связи с индивидуальными особенностями обучающихся. Эти средства могут быть предоставлены университетом, или могут использоваться собственные технические средства.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

- для слепых и слабовидящих: в печатной форме увеличенным шрифтом, в форме электронного документа, в форме аудиофайла.
- для глухих и слабослышащих: в печатной форме, в форме электронного документа.
- для обучающихся с нарушениями опорно-двигательного аппарата: в печатной форме, в форме электронного документа, в форме аудиофайла.

Учебные аудитории для всех видов контактной и самостоятельной работы, научная библиотека и иные помещения для обучения оснащены специальным оборудованием и учебными местами с техническими средствами обучения:

- для слепых и слабовидящих: устройством для сканирования и чтения с камерой SARA CE; дисплеем Брайля PAC Mate 20; принтером Брайля EmBraille ViewPlus;
- для глухих и слабослышащих: автоматизированным рабочим местом для людей с нарушением слуха и слабослышащих; акустический усилитель и колонки;
- для обучающихся с нарушениями опорно-двигательного аппарата: передвижными, регулируемыми эргономическими партами СИ-1; компьютерной техникой со специальным программным обеспечением.

9. Методические материалы

9.1 Планы семинарских/ практических/ лабораторных занятий

1. Основные задачи машинного обучения, их применение в компьютерной лингвистике.

Общая постановка задачи машинного обучения, обучение с учителем и без учителя. Примеры задач машинного обучения. Основные задачи компьютерной лингвистики. Их интерпретация как задач машинного обучения.

2. Задачи классификации в обработке естественного языка и компьютерной лингвистике. Признаковое описание объектов, типы признаков

Общая постановка задачи классификации. Типы признаков: двоичные, категориальные, вещественные, порядковые. Двоичная и многоклассовая классификация. Задачи компьютерной лингвистики как задачи классификации. Типы признаков в различных задачах, их извлечение из текста.

3. Векторная модель классификации, матрица объекты-признаки, линейные классификаторы.

Векторное представление текста в задачах компьютерной лингвистики, модель мешка слов. Матрица объекты-признаки. Сведение задачи классификации к подбору оптимальной разделяющей плоскости. Свойства задач текстовой классификации (большое количество признаков, разреженность).

4. Библиотека scikit-learn для машинного обучения. Основные объекты и типы данных.

Решение задач машинного обучение на языке Python. Представление данных, модуль Numpy, массивы и матрицы. Операции с матрицами и векторами в модуле Numpy. Библиотека scikit-learn, основные классы и методы. Разреженные матрицы и модуль scipy.

5. Наивный байесовский классификатор, вероятностная модель.

Вероятностная модель классификации, наивный байесовский классификатор. Его связь с моделью мешка слов. Достоинства и недостатки наивного байесовского классификатора. Проблема нулевых вероятностей, сглаживание признаков. Реализация наивного байесовского классификатора в библиотеке scikit-learn.

6. Линейные классификаторы в библиотеке scikit-learn, основные алгоритмы.

Понятие линейного классификатора, случай двух и более классов. Решающее правило классификатора, вектор весов, разделяющая гиперплоскость. Основные алгоритмы нахождения вектора весов: метод опорных векторов, логистическая регрессия.

7. Извлечение признаков из текста, меры качества признаков. Оценка качества классификации.

Стандартные признаки в задачах классификации: подсчёт количество слов, символьные энграммы. Зависимость качества классификации от используемых признаков. Меры оценки качества (точность, полнота, F-мера). Обучающая и контрольная выборка, скользящий контроль, переобучение. Отбор признаков, основные методы оценки качества признака (вес признака, вероятность класса).

8. Автоматическая классификация текстов, модель мешка слов. Стандартные приёмы в задачах текстовой классификации.

9. Обучение без учителя, кластеризация, стандартные алгоритмы: иерархическая кластеризация и метод средних.

Понятие обучения без учителя, постановка задачи кластеризации. Методы измерения расстояния между объектами, их зависимость от задачи. Иерархическая кластеризация, вычисление межкластерного расстояния. Метод k-средних, его сравнение с иерархической кластеризацией. Применения кластеризации.

10. Дистрибутивная семантика, семантические вектора, их применение в различных задачах.

Дистрибутивная семантика, основные подходы. Матрица совместных вхождений. Понижение размерности семантических векторов

11. Дистрибутивная семантика на основе нейронных сетей. Обзор её применений

Общее знакомство с нейронными сетями. Методы skipgram и CBOW для получения дистрибутивных векторов, модель Т. Миколова. Свойства дистрибутивных векторов. Применение дистрибутивных векторов в задачах компьютерной лингвистики.

12. Обзор применений машинного обучения в различных задачах компьютерной лингвистики (морфология, синтаксис и т.д.).

9.2 Иные материалы

Все необходимые для обучения материалы публикуются по адресу https://github.com/rsuh-python/ в соответствующих репозиториях.