МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОСВЕЩЕНИЯ»
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОСВЕЩЕНИЯ)

Лингвистический факультет
Кафедра теории языка, англистики и прикладной лингвистики

Согласовано
деканом лингвистического факультета
«14» марта 2024 г.
_____
/Вековищева С.Н./

Рабочая программа дисциплины

Введение в NLP

Направление подготовки
45.03.02 Лингвистика

Профиль:
Цифровая лингвистика (английский язык + китайский или корейский языки)

Квалификация
Бакалавр

Форма обучения
Очная

Согласовано учебно-методической комиссией лингвистического факультета
Протокол «14» марта 2024 г. № 5
Председатель УМКом_____
/Горбачева О.А./

Рекомендовано кафедрой теории языка, англистики и прикладной лингвистики
Протокол от «26» февраля 2024 г. № 8
Зав. кафедрой_____
/Холстинина Т.В./

Мытищи
2024

Автор-составитель:
Иванов Владимир Андреевич, доцент , кандидат филологических наук

Рабочая программа дисциплины «ВВЕДЕНИЕ В NLP» составлена в соответствии с требованиями федерального государственного образовательного стандарта высшего образования по направлению подготовки 45.03.02 Лингвистика, утвержденного приказом МИНОБРНАУКИ РОССИИ от 12.08.2020 № 969.

Дисциплина входит в часть, формируемую участниками образовательных отношений, Блока 1 «Дисциплины (модули)» и является элективной дисциплиной.

Год начала подготовки (по учебному плану)2024

СОДЕРЖАНИЕ

1. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

1.1. Цель и задачи дисциплины

Цель освоения дисциплины «Введение в NLP» предполагает повышение уровня культуры и образования студентов путем приобщения их к общенаучному знанию, пониманию основных идей и методов компьютерной обработки естественного языка.

Практическая цель состоит в формировании у студентов компетенций, необходимых для использования методов NLP (Natural Language Processing, компьютерная обработка естественного языка), готовности применения этих компетенций в научно-исследовательской и научно-практической деятельности.

Задачи дисциплины:

- актуализация и развитие знаний в области NLP;
- формирование навыков построения формальных и математических моделей языка.

1.2. Планируемые результаты обучения

В результате освоения данной дисциплины у обучающихся формируются следующие компетенции:

СПК-3. Владеет основными математико-статистическими методами обработки лингвистической информации с учетом элементов программирования и автоматической обработки лингвистических данных.

СПК-4. Способен применять основные современные методы научного исследования, в том числе и в смежных областях, в самостоятельных исследованиях.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина входит в часть, формируемую участниками образовательных отношений, Блока 1 «Дисциплины (модули)» и является элективной дисциплиной.

Дисциплина знакомит студентов с базовыми понятиями, идеями и методами NLP (Natural Language Processing, компьютерная обработка естественного языка).

Дисциплина опирается на знания, полученные студентами в рамках школьного образования, а также в результате освоения таких дисциплин, как «Введение в информационные технологии», «Понятийный аппарат математики», «Компьютерная лингвистика», «Инструменты искусственного интеллекта для анализа и обработки текста», «Обучающие лингвистические системы», «Введение в анализ больших данных», «Корпусная лингвистика», «Информационно-поисковые системы», «База данных», «Квантитативная лингвистика» и др.

3. ОБЪЕМ И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

3.1. Объем дисциплины

Показатель объема дисциплины	Форма обучения
	Очная
Объем дисциплины в зачетных единицах	3
Объем дисциплины в часах	108
Контактная работа:	48.2
Лекции	12
Практические занятия	36
Контактные часы на промежуточную аттестацию:	0.2
Зачет с оценкой	0,2
Самостоятельная работа	52
Контроль	7,8

Форма промежуточной аттестации: зачет с оценкой в 7 семестре.

3.2. Содержание дисциплины

Наименование разделов (тем) дисциплины с кратким содержанием	Количество часов	
	Лекции	Практические занятия
Тема 1. NLP как научная и практическая область знания Место NLP среди дисциплин, связанных с автоматической обработкой естественного языка. NLP и компьютерная лингвистика. Задачи и методы NLP. Подходы к решению задач: правила, машинное обучение, нейронные сети. Показатели качества: точность, полнота, F-мера	2	6
Тема 2. Языковые модели Языковые модели на основе n-грамм, перплексия, методы сглаживания, линейная интерполяция. Нейронные языковые модели. Применение языковых моделей.	2	6
Тема 3. Разметка текста и извлечение информации Частеречная разметка. Извлечение именованных сущностей. Подходы: скрытые марковские модели, машинное обучение, рекуррентные нейронные сети.	2	6
Тема 4. Классификация текстов и анализ тональности Задачи классификации. Наивный байесовский классификатор. Проблемы классификации текстов. Анализ тональности.	2	6
Тема 5. Информационный поиск Векторные модели текстов. Матричное представление. Индекс. Ранжированный информационный поиск. Коэффициент Жаккара. TF-IDF. Методы оценки качества поиска.	2	6
Тема 6. Вычислительная семантика Семантические ресурсы (WordNet). Измерение семантической близости. Дистрибутивные семантические модели. Контекстуализированные векторные представления. Word2Vec.	2	6
Итого	**12**	**36**

4. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

Темы для самостоятельного изучения	Изучаемые вопросы	Количество часов	Формы самостоятельной работы	Методическое обеспечение	Формы отчетности
Тема 1. NLP как научная и практическая область знания	Место NLP среди дисциплин, связанных с автоматической обработкой естественного языка. NLP и компьютерная лингвистика. Задачи и методы NLP. Подходы к решению задач: правила, машинное обучение, нейронные сети. Показатели качества: точность, полнота, F-мера	7	Изучение источников, чтение литературы. Решение задач.	Основная и дополнительная литература, интернет-источники	Обсуждение и анализ источников. Проверка домашних заданий.
Тема 2. Языковые модели	Языковые модели на основе n-грамм, перплексия, методы сглаживания, линейная интерполяция. Нейронные языковые модели. Применение языковых моделей.	9	Изучение источников, чтение литературы. Решение задач.	Основная и дополнительная литература, интернет-источники	Обсуждение и анализ источников. Проверка домашних заданий.
Тема 3.	Частеречная разметка.	9	Изучение	Основная и	Обсуждение

Разметка текста и извлечение информации	Извлечение именованных сущностей. Подходы: скрытые марковские модели, машинное обучение, рекуррентные нейронные сети.		источников, чтение литературы. Решение задач.	дополнительная литература, интернет-источники	и анализ источников. Проверка домашних заданий.
Тема 4. Классификация текстов и анализ тональности	Задачи классификации. Наивный байесовский классификатор. Проблемы классификации текстов. Анализ тональности.	9	Изучение источников, чтение литературы. Решение задач.	Основная и дополнительная литература, интернет-источники	Обсуждение и анализ источников. Проверка домашних заданий.
Тема 5. Информационный поиск	Векторные модели текстов. Матричное представление. Индекс. Ранжированный информационный поиск. Коэффициент Жаккара. TF-IDF. Методы оценки качества поиска.	9	Изучение источников, чтение литературы. Решение задач.	Основная и дополнительная литература, интернет-источники	Обсуждение и анализ источников. Проверка домашних заданий.
Тема 6. Вычислительная семантика	Семантические ресурсы (WordNet). Измерение семантической близости. Дистрибутивные семантические модели. Контекстуализированные векторные представления. Word2Vec.	9	Изучение источников, чтение литературы. Решение задач.	Основная и дополнительная литература, интернет-источники	Обсуждение и анализ источников. Проверка домашних заданий.
Итого		**52**			

5. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕЙ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

5.1. Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы

Код и наименование компетенции	Этапы формирования
СПК-3. Владеет основными математико-статистическими методами обработки лингвистической информации с учетом элементов программирования и автоматической обработки лингвистических данных.	1. Работа на учебных занятиях 2. Самостоятельная работа.
СПК-4. Способен применять основные современные методы научного исследования, в том числе и в смежных областях, в самостоятельных исследованиях.	1. Работа на учебных занятиях 2. Самостоятельная работа.

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

Оцениваемые компетенции	Уровень сформированности	Этап формирования	Описание показателей	Критерий оценивания	Шкала оценивания
СПК-3	Пороговый	1. Работа на учебных занятиях 2. Самостоятельная	**Знать:** математико-статистические методы обработки	Устный опрос, выполнение практическог	Шкала оценивания устного

		работа.	лингвистической информации, основы программирования, принципы автоматической обработки корпусов текстов **Уметь:** применять полученные знания для анализа и обработки нового лингвистического материала на изучаемых языках	о задания	опроса Шкала оценивания практического задания
	Продвинутый	1. Работа на учебных занятиях 2. Самостоятельная работа.	**Знать:** математико-статистические методы обработки лингвистической информации, основы программирования, принципы автоматической обработки корпусов текстов **Уметь:** применять полученные знания для анализа и обработки нового лингвистического материала на изучаемых языках **Владеть:** способами представления полученных результатов, методикой изложения, принятой в соответствующей области лингвистического знания	Устный опрос, выполнение практического задания	Шкала оценивания устного опроса Шкала оценивания практического задания
СПК-4	Пороговый	1. Работа на учебных занятиях 2.	**Знать:** общенаучные методы и	Устный опрос, выполнение	Шкала оценивания

		Самостоятельная работа.	конкретные методики изучения данных в соответствующей области лингвистики; принципы работы с библиографическими источниками **Уметь:** использовать основные информационно-поисковые и экспертные системы, системы представления знаний в данной предметной области, принципы научно-доказательного изложения материала.	практического задания	устного опроса Шкала оценивания практического задания
	Продвинутый	1. Работа на учебных занятиях 2. Самостоятельная работа.	**Знать:** общенаучные методы и конкретные методики изучения данных в соответствующей области лингвистики; принципы работы с библиографическими источниками **Уметь:** использовать основные информационно-поисковые и экспертные системы, системы представления знаний в данной предметной области, принципы научно-доказательного изложения материала.	Устный опрос, выполнение практического задания	Шкала оценивания устного опроса Шкала оценивания практического задания

			Владеть: проблематикой смежных с лингвистикой областей и возможными подходами к их решению с позиций комплексного подхода		

Описание шкал оценивания

Шкала оценивания устного опроса

Критерии оценивания	**Баллы**
Выполнено правильно как минимум 80% заданий	26 баллов/отлично
Выполнено правильно как минимум 60% заданий	23 балла/хорошо
Выполнено правильно как минимум 40% заданий	16 баллов/удовлетворительно
Выполнено правильно менее 40% заданий	12 баллов/неудовлетворительно

Шкала оценивания практического задания

Критерии оценивания	**Баллы**
Выполнено правильно как минимум 80% предложенного задания	26 баллов/отлично
Выполнено правильно как минимум 60% предложенного задания	22 балла/хорошо
Выполнено правильно как минимум 40% предложенного задания	18 баллов/удовлетворительно
Выполнено правильно менее 40% предложенного задания	12 баллов/неудовлетворительно

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

Примерные вопросы к устному опросу

1) NLP with Python.
2) Библиотека NumPy.
3) Библиотека PyQt.

Примерные образцы практических заданий

1. Оценить качество с точки зрения полноты и точности. Вычислить F-меру.
2. Написать простую реализацию n-граммной модели на языке Python.
3. Сравнить качество частеречной разметки нескольких POS-tagger-ов.
4. Построить векторное представление текста.
5. Посчитать семантическую близость на основе WordNet.

Примерные вопросы к зачету с оценкой

1. Место NLP среди дисциплин, связанных с автоматической обработкой

естественного языка. NLP и компьютерная лингвистика. Задачи и методы NLP.
2. Подходы к решению задач: правила, машинное обучение, нейронные сети.
3. Показатели качества: точность, полнота, F-мера
4. Языковые модели на основе n-грамм, перплексия, методы сглаживания, линейная интерполяция.
5. Нейронные языковые модели. Применение языковых моделей.
6. Частеречная разметка. Подходы.
7. Извлечение именованных сущностей. Подходы: скрытые марковские модели, машинное обучение, рекуррентные нейронные сети.
8. Задачи классификации. Наивный байесовский классификатор. Проблемы классификации текстов.
9. Анализ тональности.
10. Векторные модели текстов. Матричное представление.
11. Индекс. Ранжированный информационный поиск.
12. Коэффициент Жаккара. TF-IDF.
13. Методы оценки качества поиска.
14. Семантические ресурсы (WordNet). Измерение семантической близости.
15. Дистрибутивные семантические модели. Контекстуализированные векторные представления.
16. Word2Vec.

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

В рамках освоения дисциплины предусмотрены следующие формы текущего контроля: выполнение практического задания, домашнего задания, устного опроса, решение задач.

Общее максимальное количество баллов по дисциплине — 100 баллов.

Максимальное количество баллов, которое может набрать студент в течение семестра за текущий контроль, равняется 70 баллам.

Максимальная сумма баллов, которые студент может набрать на зачете с оценкой , равняется 30 баллам.

При оценке знаний на зачете учитываются:

1. Понимание и степень усвоения теории курса.
2. Уровень знания фактического материала в объеме программы.
3. Правильность формулировки основных понятий и закономерностей.
4. Логика, структура и грамотность изложения вопроса.
5. Использование примеров.
6. Умение связать теорию с практическим применением.
7. Умение сделать обобщение, выводы.
8. Умение ответить на дополнительные вопросы.
9. Умение выделять главное, существенное.

Шкала оценивания зачета с оценкой

Критерии оценивания	**Баллы**
Выставляется за ответ, который демонстрирует прекрасное знание предмета, умение соединять знания из различных разделов курса, легко и безошибочно иллюстрировать теоретические положения примерами, как взятыми из учебника, так и своими собственными; владение терминологией из различных разделов курса, безошибочное выполнение практического задания	30-21 балл

Выставляется за ответ, который демонстрирует хорошее знание и понимание изученного материала, подкреплён примерами, взятыми из лекций или учебника; допускаются единичные ошибки, которые экзаменуемый исправляет самостоятельно после замечаний преподавателя.	20-16 баллов
Выставляется за ответ, который обнаруживает самое общее понимание теории, однако, плохо подкрепляемое практическими примерами. При таком ответе студент проявляет неуверенность, не всегда даёт исчерпывающие аргументированные ответы на заданные вопросы	15-11 баллов
Выставляется за ответ, который обнаруживает непонимание сути вопроса, являясь механическим повторением курса лекций или учебника; незнание терминологии, искажение смысла понятий; неумение соотнести теорию с практикой.	10-0 баллов

Итоговая шкала выставления оценки по дисциплине

Итоговая оценка складывается из оценки за выполнения всех предусмотренных в программе дисциплины форм отчетности в рамках текущего контроля, а также оценки на промежуточной аттестации.

Баллы, полученные в течение освоения дисциплины	**Оценка по дисциплине**
81-100	отлично
61-80	хорошо
41-60	удовлетворительно
0-40	не удовлетворительно

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И РЕСУРСНОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

6.1. Основная литература

1. Николаев, И. С. Прикладная и компьютерная лингвистика. Изд.2. / Николаев И. С., Митренина О. В., Ландо Т. М. (Ред.). — М.: URSS, 2017. — 320 с. — ISBN 978-5-9710-4633-2.
2. Большакова, Е. И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика [Электронный ресурс]: учеб. пособие / Большакова Е. И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с — ISBN 978–5–94506–294–8. — URL: http://www.hse.ru/data/2012/04/05/1251263483/пособие%20школа%20по%20компьютерной%20лингвистике%20-%20копия.pdf.

6.2. Дополнительная литература

1. Боярский, К. К. Введение в компьютерную лингвистику : учебное пособие / К. К. Боярский. — Санкт-Петербург : НИУ ИТМО, 2013. — 72 с. — Текст : электронный. — URL: https://e.lanbook.com/book/70822 (дата обращения: 23.06.2024).
2. Щипицина, Л. Ю. Информационные технологии в лингвистике : учебное пособие / Л. Ю. Щипицина. — 3-е изд., стер. — Москва : ФЛИНТА, 2017. — 126 с. — ISBN 978-5-9765-1431-7. — Текст : электронный. — URL: https://e.lanbook.com/book/119463 (дата обращения: 23.06.2024).

6.3. Ресурсы информационно-телекоммуникационной сети Интернет

1. ПостНаука [Электронный ресурс]. — URL: https://postnauka.ru/
2. НаукаPRO: просветительский проект [Электронный ресурс]. — URL: https://nauka-pro.ru/
3. «Элементы большой науки»: популярный сайт о фундаментальной науке [Электронный ресурс]. — URL: https://elementy.ru/
4. N+1: научные статьи, новости, открытия [Электронный ресурс]. — URL: https://nplus1.ru/
5. Энциклопедия Кругосвет: Универсальная научно-популярная энциклопедия [Электронный ресурс]. — URL: https://www.krugosvet.ru/
6. Электронные ресурсы библиотеки Государственного университета прсвещения.

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

1.Методические рекомендации по организации самостоятельной работы студентов

8.ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ

Лицензионное программное обеспечение:

Microsoft Windows
Microsoft Office
Kaspersky Endpoint Security

Информационные справочные системы:

Система ГАРАНТ
Система «КонсультантПлюс»

Профессиональные базы данных:

fgosvo.ru – Портал Федеральных государственных образовательных стандартов высшего образования

pravo.gov.ru - Официальный интернет-портал правовой информации

www.edu.ru – Федеральный портал Российское образование

Свободно распространяемое программное обеспечение, в том числе отечественного производства:

ОМС Плеер (для воспроизведения Электронных Учебных Модулей)

7-zip

Google Chrome

8. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Материально-техническое обеспечение дисциплины включает в себя:

- учебные аудитории для проведения занятий лекционного и семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, укомплектованные учебной мебелью, доской, демонстрационным оборудованием;
- помещения для самостоятельной работы, укомплектованные учебной мебелью, персональными компьютерами с подключением к сети Интернет и обеспечением доступа к электронным библиотекам и в электронную информационно-образовательную среду