

**ГОУ ВПО Российско-Армянский (Славянский)
университет**

Утверждено
Директор Института Математики и Информатики
Арамян Р.Г.



«21» марта 2025, протокол № 9.1

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС ДИСЦИПЛИНЫ

Наименование дисциплины: NLP

Автор: Варданян Эдгар Араевич

Направление подготовки: 01.04.02 Прикладная математика и информатика

Наименование образовательной программы: Машинное обучение и наука о данных

1. АННОТАЦИЯ

1.1. Краткое описание содержания данной дисциплины

Целями освоения дисциплины NLP (Natural Language Processing) являются знакомство с основными проблемами в области компьютерной лингвистики, базовыми алгоритмами, математическими методами моделирования языковых феноменов, основными инструментами и технологиями в области автоматической обработки естественного языка, умение представлять в алгоритмическом виде процессы анализа и синтеза текста.

1.2. Трудоемкость в академических кредитах и часах, формы итогового контроля

Трудоемкость курса составляет 4 академических кредита, что эквивалентно **144** часам (1 академический кредит равен 36 часам). Итоговый контроль будет проведен в форме **зачета**.

1.3. Взаимосвязь дисциплины с другими дисциплинами учебного плана специальности

При изучении дисциплины «NLP» используются понятия и методы Python, Big Data, Mathematica

1.4. Результаты освоения программы дисциплины:

Код компетенции	Наименование компетенции	Код индикатора достижения компетенций	Наименование индикатора достижений компетенций
ПК-7	Способностью разрабатывать и оптимизировать бизнес-планы научно-прикладных проектов	ПК-7.1	Знать основы разработки бизнес-планов для научно-прикладных проектов
		ПК-7.2	Уметь создавать и оптимизировать бизнес-планы, оценивать их эффективность
		ПК-7.3	Владеть методами финансового и стратегического планирования в контексте

			научных проектов
ОПК-3	Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности	ОПК-3.1	Знать принципы и методы математического моделирования и анализа
		ОПК-3.2	Уметь создавать точные и эффективные математические модели для конкретных прикладных задач
		ОПК-3.3	Владеть глубокими знаниями в анализе моделей, умением оценивать их адекватность и точность, а также способностью предсказывать их поведение в различных условиях

2. УЧЕБНАЯ ПРОГРАММА

2.1. Цели и задачи дисциплины

Цели дисциплины:

Обучить студентов основам и современным методам обработки естественного языка.

Развить у студентов навыки применения алгоритмов NLP для анализа текстовых данных.

Подготовить студентов к решению практических задач в области NLP.

Задачи дисциплины:

Изучение теоретических основ и практического применения методов морфологического, синтаксического и семантического анализа текстов.

Освоение методов машинного обучения для задач классификации текста, автоматического резюмирования, и машинного перевода.

Развитие умения работать с большими объемами текстовых данных и использовать библиотеки и инструменты NLP.

Применение полученных знаний для создания и оценки NLP-моделей в реальных проектах.

2.2. Трудоемкость дисциплины и виды учебной работы

Виды учебной работы	Всего, в акад. часах	Распределение по семестрам					
		— сем	2 — сем	— сем	— сем.	— сем	— сем.
1	2	3	4	5	6	7	8
1.Общая трудоемкость изучения дисциплины по семестрам, в т. ч.:	32		32				
1.1. Аудиторные занятия, в т. ч.:	32		32				
1.1.1. Лекции							
1.1.2. Практические занятия, в т. ч.	32		32				
Итоговый контроль			Зачет				

2.3. Содержание дисциплины

2.3.1. Тематический план и трудоемкость аудиторных занятий (модули, разделы дисциплины и виды занятий) по рабочему учебному плану

Разделы и темы дисциплины	Всего ак. часов	Лекции, ак. часов	Практ. занятия, ак. часов	Семина- ры, ак. часов	Лабор, ак. часов
1	3=4+5+6 +7+8	4	5	6	7
Модуль 1.	32		32		
Введение в NLP	1		1		
Раздел 1.	10		10		
Тема 1.1. Проблемы обработки естественного языка	1		1		
Тема 1.2. Обсуждение проблем и приложений НЛП	1		1		
Тема 1.3. Формальные языки, синтаксис и семантика	2		2		
Тема 1.4. Нормализация текста, токенизация	2		2		
Тема 1.5. Лемматизация, Поверхностные формы, POS-теги.	2		2		
Тема 1.6. Расстояние Левенштейна, Пространство	2		2		
Раздел 2. Простые модели и вложения слов	21		21		
Тема 2.1. N-граммное языковое моделирование	1		1		

Тема 2.2. TF-IDF, Пакет слов, наивные байесовские классификаторы	4		4		
Тема 2.3. Задача WordNet.	4		4		
Тема 2.4. Реализация Word2Vec, Glove	2		2		
Тема 2.5. Смысл сходства, арифметика встраивания слов	2		2		
Тема 2.6. Информация о подслове. Встраивание символов FastText и N-грамм.	2		2		
Тема 2.7. Информация о подслове. Модели подслов, кодирование Byte-Pair, модели WordPiece и Unigram.	2		2		
Тема 2.8. Структурированный прогноз. Скрытые марковские модели.	2		2		
Тема 2.9. Структурированный прогноз. Условные случайные поля.	2		2		
ИТОГО	32		32		

2.3.2. Краткое содержание разделов дисциплины в виде тематического плана

- **Тема 1.1. Проблемы обработки естественного языка**

Обзор основных вызовов, связанных с анализом и пониманием естественных языков, включая многообразие языков и неоднозначность текстов.

- **Тема 1.2. Обсуждение проблем и приложений НЛП**

Рассмотрение реальных примеров применения NLP в различных областях, таких как машинный перевод и анализ эмоций.

- **Тема 1.3. Формальные языки, синтаксис и семантика**

Изучение основ формальных языков, а также синтаксической и семантической структуры языка.

- **Тема 1.4. Нормализация текста, токенизация**

Обзор методов предварительной обработки текста, включая нормализацию и разбиение текста на токены.

- **Тема 1.5. Лемматизация, Поверхностные формы, POS-теги**

Описание процессов приведения слов к их базовой форме, определения частей речи и анализа морфологической структуры слов.

- **Тема 1.6. Расстояние Левенштейна, Пространство**

Изучение метрик для измерения различий между текстовыми строками, включая расстояние Левенштейна.

Раздел 2: Простые модели и вложения слов

- **Тема 2.1. N-граммное языковое моделирование**

Введение в создание и использование n-граммных моделей для предсказания следующего слова на основе предыдущих.

- **Тема 2.2. TF-IDF, Пакет слов, наивные байесовские классификаторы**

Обзор методов векторизации текста и классификации на основе статистических подходов.

- **Тема 2.3. Задача WordNet**

Рассмотрение WordNet как инструмента для семантического анализа и семантических отношений между словами.

- **Тема 2.4. Реализация Word2Vec, Glove**

Изучение моделей для получения векторных представлений слов, обеспечивающих захват контекстуальных значений.

- **Тема 2.5. Смысл сходства, арифметика встраивания слов**

Анализ способов измерения семантической близости слов и применения арифметики вложений для решения NLP задач.

- **Тема 2.6. Информация о подслове. Встраивание символов FastText и N-грамм**

Обзор подходов к встраиванию, учитывающих подсловную информацию для улучшения качества анализа.

- **Тема 2.7. Информация о подслове. Модели подслов, кодирование Byte-Pair, модели WordPiece и Unigram**

Исследование методов сегментации слов на подсловные единицы и их использование в моделях обучения.

- **Тема 2.8. Структурированный прогноз. Скрытые марковские модели**

Введение в использование скрытых марковских моделей для задач структурированного прогнозирования в NLP.

- **Тема 2.9. Структурированный прогноз. Условные случайные поля**

Обзор использования условных случайных полей для более сложных задач структурированного прогнозирования.

2.3.3. Краткое содержание семинарских/практических занятий/лабораторного практикума

Лабораторные работы: Студенты выполняют практические задания на компьютерах с использованием программных инструментов и библиотек, таких как Python, NLTK, spaCy, TensorFlow и PyTorch для обработки текстовых данных, создания и тестирования моделей NLP.

Проектная работа: Разработка собственных проектов, где студенты применяют алгоритмы NLP к конкретным примерам, таким как анализ тональности, машинный перевод или автоматическое резюмирование текстов.

Работа с данными: Обработка и анализ больших объемов текстовых данных, обучение статистических моделей, оценка их эффективности на практических примерах.

2.3.4. Материально-техническое обеспечение дисциплины

Компьютерный класс: Оборудован современными компьютерами, на которых установлено необходимое программное обеспечение для разработки и тестирования NLP-проектов, включая Python, Jupyter Notebook, а также библиотеки, такие как NLTK, spaCy, TensorFlow и PyTorch.

Проекторы: Используются для демонстрации учебных материалов, проведения лекций и презентации студенческих проектов и кейсов на интерактивных занятиях.

2.4. Модульная структура дисциплины с распределением весов по формам контролей

Формы контролей	Вес формы (форм) текущего контроля в результирующей оценке текущего контроля (по модулям)		Вес формы промежуточного контроля в итоговой оценке промежуточного контроля		Вес итоговой оценки промежуточного контроля в результирующей оценке промежуточных контролей		Вес итоговой оценки промежуточного контроля в результирующей оценке промежуточных контролей (семестровой оценке)		Веса результирующей оценки промежуточных контролей и оценки итогового контроля в результирующей оценке итогового контроля	
	М1	М2	М1	М2	М1	М2				
Вид учебной работы/контроля										
Контрольная работа <i>(при наличии)</i>				0,7						
Устный опрос <i>(при наличии)</i>										
Тест <i>(при наличии)</i>										
Лабораторные работы <i>(при наличии)</i>										
Письменные домашние задания <i>(при наличии)</i>				0,3						
Реферат <i>(при наличии)</i>										
Эссе <i>(при наличии)</i>										
Проект <i>(при наличии)</i>										
<i>Другие формы (при наличии)</i>										
Веса результирующих оценок текущих контролей в итоговых оценках промежуточных контролей						0,3				
Веса оценок промежуточных контролей в итоговых оценках промежуточных контролей						0,7				
Вес итоговой оценки 1-го промежуточного контроля в результирующей оценке промежуточных контролей										
Вес итоговой оценки 2-го промежуточного контроля в результирующей оценке промежуточных контролей							1			

Вес результирующей оценки промежуточных контролей в результирующей оценке итогового контроля								1
Вес итогового контроля (Экзамен/зачет) в результирующей оценке итогового контроля								
	$\Sigma = 1$	$\Sigma = 1$	$\Sigma = 1$	$\Sigma = 1$	$\Sigma = 1$	$\Sigma = 1$	$\Sigma = 1$	$\Sigma = 1$

3. Теоретический блок

3.1. Материалы по теоретической части курса

3.1.1. Учебники

1. Stanford CS224N Natural Language Processing course,
2. Speech and Language Processing textbook by Dan Jurafsky,
3. LxMLS Lisbon Machine Learning / NLP Summer School course,
4. Oxford Deep NLP course

4. Фонды оценочных средств

Контрольные работы

Проектные работы

Домашние задания

Устные опросы

4.1. Планы практических и семинарских занятий

Практические занятия включают выполнение заданий на обработку текстов, классификацию данных, машинный перевод и другие задачи NLP, ориентированы на обсуждение теоретических аспектов, анализ статей, кейс-стади и разбор реальных проектов из области NLP.

5. Методический блок

5.1. Методика преподавания

Методика преподавания курса NLP ориентирована на сочетание теоретических лекций и активных практических занятий. Она направлена на формирование глубоких знаний и практических навыков в области NLP.

5.1.1. Методические рекомендации для студентов по подготовке к семинарским, практическим или лабораторным занятиям, по организации самостоятельной работы студентов при изучении конкретной дисциплины.

Подготовка к семинарским и лабораторным занятиям:

- **Предварительное изучение материала:** До занятий студентам рекомендуется изучить теоретические материалы и примеры, которые будут обсуждаться. Это поможет активнее участвовать в дискуссиях и практических заданиях.
- **Активное участие:** Важно задавать вопросы и делиться мнениями во время занятий. Активное участие помогает углубить понимание материала и развивает критическое мышление.
- **Заметки:** Ведение заметок во время лекций и практических занятий помогает лучше усваивать информацию и облегчает подготовку к экзаменам.

Организация самостоятельной работы:

- **Регулярное повторение:** Студенты должны регулярно пересматривать учебные материалы, чтобы улучшить понимание и запоминание информации.
- **Решение задач:** Практика является ключевым элементом изучения NLP. Решение задач и выполнение проектов позволяют применить теоретические знания на практике.
- **Использование ресурсов:** Рекомендуется использовать дополнительные ресурсы, такие как онлайн-курсы, вебинары, научные статьи и учебники для расширения знаний.
- **Взаимодействие с однокурсниками:** Обсуждение учебных вопросов и совместная работа над проектами помогают улучшить понимание сложных концепций и способствуют развитию коммуникативных навыков.