СИЛЛАБУС
Весенний семестр 2023-2024 учебного года
Образовательная программа «7М06101 – Вычислительная лингвистика»

ID и наименование дисциплины	Самостоятельная работа обучающегося (СРО)	Кол-во кредитов			Общее кол-во кредитов	Самостоятельная работа обучающегося под руководством преподавателя (СРОП)
		Лекции (Л)	Практ. занятия (ПЗ)	Лаб. занятия (ЛЗ)		
ID 102269 Методы машинного обучения в обработке естественного языка	4	1,7	0	3,3	5	9

АКАДЕМИЧЕСКАЯ ИНФОРМАЦИЯ О ДИСЦИПЛИНЕ				
Формат обучения	**Цикл, компонент**	**Типы лекций**	**Типы практических занятий**	**Форма и платформа итогового контроля**
Оффлайн	БД, КВ	Проблемно-ориентированный	Изучение концепций обработки естественных языков с помощью моделей машинного обучения	Устный оффлайн
Лектор - (ы)	Карюкин Владислав Игоревич			
e-mail:	vladislav.karyukin@gmail.com vladislav.karyukin@kaznu.kz			
Телефон:	+77019405992			
Ассистент- (ы)	Карюкин Владислав Игоревич			
e-mail:	vladislav.karyukin@gmail.com vladislav.karyukin@kaznu.kz			
Телефон:	+77019405992			

АКАДЕМИЧЕСКАЯ ПРЕЗЕНТАЦИЯ ДИСЦИПЛИНЫ		
Цель дисциплины	**Ожидаемые результаты обучения (РО)***	**Индикаторы достижения РО (ИД)**
Этот курс направлен на изучение концепций обработки естественных языков, базовых принципов и задач NLP, включая обучение с учителем, обучение без учителя, глубокое обучение, включающее архитектуры сверточных нейронных сетей, рекуррентных нейронных сетей, трансформеров и больших языковых моделей, таких как BERT, GPT-3 и т.д.	1. (когнитивный) Теоретические и методологические концепции NLP	1.1 Понимает базовые и расширенные программы по обработке текстов
		1.2 Анализирует особенности методов стемминга и лемматизации текстов
		1.3 Применяет методы разработки приложений по обработке текстов
	2. (функциональный) Работа с библиотеками NumPy, Pandas и Matplotlib	2.1 Использует знания установки библиотек NumPy, Pandas и Matplotlib
		2.2 Применяет данные библиотеки для работы с моделями машинного обучения
		2.3 Формирует навыки работы с данными библиотеками при создании приложений
	3.(функциональный) Разрабатывать программы парсинга текстовых данных из интернет-источников	3.1 Разрабатывает скрипты для парсинга текстов
		3.2 Создает полнофункциональную программу парсинга текстов
		3.3 Разрабатывает скрипты сохранения полученных данных в текстовые файлы
	4. (системный) Создавать веб-краулеры для поиска данных в интернете	4.1 Создает конфигурацию веб-краулера
		4.2 Разрабатывает скрипт выборки источников данных для парсинга
		4.3 Создает полнофункциональный веб-краулер поиска данных в интернете
	5. (системный) Создавать веб-приложения, использующие модели машинного обучения и	5.1 Создавать основной каркас веб-приложения на HTML, CSS и JavaScript

	нейронные сети	5.2 Создавать подключение моделей машинного обучения для определения угроз
		5.3 Создавать полную конфигурацию работы веб-приложения
Пререквизиты	Формальные грамматики, Информационные технологии для NLP, Языковые ресурсы	
Постреквизиты	Интеллектуальный анализ данных, Технологии машинного перевода, Понимание естественного языка, Глубокое обучение	
Учебные ресурсы	**Литература**: **Основная:** – Python for Everybody: Exploring Data in Python 3 by Dr. Charles Russell Severance, Sue Blumenberg, Elliott Hauser, Aimee Andrion, 2016. – Natural Language Processing with Python and spaCy: A Practical Introduction, Yuli Vasiliev, 2021. – Machine Learning and Deep Learning in Natural Language Processing, Anitha S. Pillai, Roberto Tedesco, 2023. – Natural Language Processing: A Machine Learning Perspective Yue Zhang, Zhiyang Teng, 2021. – Natural Language Processing Projects: Build Next-Generation NLP Applications Using AI Techniques, Akshay Kulkarni, Adarsha Shivananda, Anoosh Kulkarni, 2021. **Дополнительная:** – Learning Scientific Programming with Python, Christian Hill, 2021 – Deep Learning for Natural Language Processing: Creating Neural Networks with Python. Palash Goyal, Sumit Pandey, Karan Jain, 2018 **Профессиональные научные базы данных**: – Лабораторная аудитория 514 – Лабораторная аудитория 323 **Интернет–ресурсы:** – Python Exercises, Practice, Solution – https://www.w3resource.com/python-exercises/ – Natural Language Toolkit – https://www.nltk.org/ – Tensorflow – https://www.tensorflow.org/?hl=ru – Machine learning mastery – https://machinelearningmastery.com/start-here/ **Программное обеспечение:** Python IDE, Anaconda Navigator Python, NLTK, Microsoft Office Word, WinRAR, Power Point, Adobe Reader, Paint.	
Академическая политика дисциплины	Академическая политика дисциплины определяется Академической политикой и Политикой академической честности КазНУ имени аль-Фараби. Документы доступны на главной странице ИС Univer. **Интеграция науки и образования.** Научно-исследовательская работа студентов, магистрантов и докторантов – это углубление учебного процесса. Она организуется непосредственно на кафедрах, в лабораториях, научных и проектных подразделениях университета, в студенческих научно-технических объединениях. Самостоятельная работа обучающихся на всех уровнях образования направлена на развитие исследовательских навыков и компетенций на основе получения нового знания с применением современных научно-исследовательских и информационных технологий. Преподаватель исследовательского университета интегрирует результаты научной деятельности в тематику лекций и семинарских (практических) занятий, лабораторных занятий и в задания СРОП, СРО, которые отражаются в силлабусе и отвечают за актуальность тематик учебных занятий и заданий. **Посещаемость.** Дедлайн каждого задания указан в календаре (графике) реализации содержания дисциплины. Несоблюдение дедлайнов приводит к потере баллов. **Академическая честность.** Практические/лабораторные занятия, СРО развивают у обучающегося самостоятельность, критическое мышление, креативность. Недопустимы плагиат, подлог, использование шпаргалок, списывание на всех этапах выполнения заданий. Соблюдение академической честности в период теоретического обучения и на экзаменах помимо основных политик регламентируют «Правила проведения итогового контроля», «Инструкции для проведения итогового контроля осеннего/весеннего семестра текущего учебного года», «Положение о проверке текстовых документов обучающихся на наличие заимствований». Документы доступны на главной странице ИС Univer.	

	Основные принципы инклюзивного образования. Образовательная среда университета задумана как безопасное место, где всегда присутствуют поддержка и равное отношение со стороны преподавателя ко всем обучающимся и обучающихся друг к другу независимо от гендерной, расовой/ этнической принадлежности, религиозных убеждений, социально-экономического статуса, физического здоровья студента и др. Все люди нуждаются в поддержке и дружбе ровесников и сокурсников. Для всех студентов достижение прогресса скорее в том, что они могут делать, чем в том, что не могут. Разнообразие усиливает все стороны жизни. Все обучающиеся, особенно с ограниченными возможностями, могут получать консультативную помощь по телефону/ e-mail vladislav.karyukin@gmail.com / +77019405992 либо посредством видеосвязи в MS Teams https://kaznukz.sharepoint.com/:u:/r/sites/msteams_011a4b/SitePages/ClassHome.aspx?csf=1&web=1&share=EdS8s-4zbZJJsOQnQpEIDmwBFO-1mV_6Oo5GeRL0ltghHQ&e=iHHZzo

ИНФОРМАЦИЯ О ПРЕПОДАВАНИИ, ОБУЧЕНИИ И ОЦЕНИВАНИИ

Балльно-рейтинговая буквенная система оценки учета учебных достижений

Оценка	Цифровой эквивалент баллов	Баллы, % содержание	Оценка по традиционной системе
A	4,0	95-100	Отлично
A-	3,67	90-94	
B+	3,33	85-89	Хорошо
B	3,0	80-84	
B-	2,67	75-79	
C+	2,33	70-74	
C	2,0	65-69	Удовлетворительно
C-	1,67	60-64	
D+	1,33	55-59	
D	1,0	50-54	
FX	0.5	25-49	Неудовлетворительно
F	0	0-24	

Методы оценивания

Критериальное оценивание – процесс соотнесения реально достигнутых результатов обучения с ожидаемыми результатами обучения на основе четко выработанных критериев. Основано на формативном и суммативном оценивании.

Формативное оценивание – вид оценивания, который проводится в ходе повседневной учебной деятельности. Является текущим показателем успеваемости. Обеспечивает оперативную взаимосвязь между обучающимся и преподавателем. Позволяет определить возможности обучающегося, выявить трудности, помочь в достижении наилучших результатов, своевременно корректировать преподавателю образовательный процесс. Оценивается выполнение заданий, активность работы в аудитории во время лекций, семинаров, практических занятий (дискуссии, викторины, дебаты, круглые столы, лабораторные работы и т. д.). Оцениваются приобретенные знания и компетенции.

Суммативное оценивание – вид оценивания, который проводится по завершению изучения раздела в соответствии с программой дисциплины. Проводится 3-4 раза за семестр при выполнении СРО. Это оценивание освоения ожидаемых результатов обучения в соотнесенности с дескрипторами. Позволяет определять и фиксировать уровень освоения дисциплины за определенный период. Оцениваются результаты обучения.

Формативное и суммативное оценивание	**Баллы % содержание**
Активность на лекциях	0
Работа на практических занятиях	25
Самостоятельная работа	25
Проектная и творческая деятельность	10
Итоговый контроль (экзамен)	40
ИТОГО	100

Календарь (график) реализации содержания дисциплины. Методы преподавания и обучения.

Неделя	Название темы	Кол-во часов	Макс. балл
	МОДУЛЬ 1 Основы операции работы с текстовыми данными		
1	**Л 1.** Введение в обработку естественных языков	1	
	ЛЗ 1. Основные операции обработки текстовых данных	2	5
2	**Л 2.** Основные этапы предобработки текстовых данных	1	0
	ЛЗ 2. Создание программы обработки текстовых данных	2	5
	СРОП 1. Консультации по выполнению СРО1 на тему «Реализация проекта с базовыми операциями обработки текстов»		
3	**Л 3.** Выполнение операции стемминга с текстовыми данными	1	
	ЛЗ 3. Создание программы стемминга текстовых данных	2	7
	СРОП 2. Прием СРО 1		20
4	**Л 4.** Выполнение операции лемматизации с текстовыми данными	1	
	ЛЗ 4. Создание программы лемматизации текстовых данных	2	7
	СРОП 3. Проведение коллоквиума по темам 1-3 неделей		5
5	**Л 5.** Выполнение операции векторизации текстовых данных	1	
	ЛЗ 5. Создание программы векторизации текстовых данных	2	7
	СРОП 4. Консультация по выполнению СРО 2 на тему «Создание программы классификации текстовых данных»		
	МОДУЛЬ 2 Обработка текстовых данных моделями машинного обучения		

6	**Л 6.** Подготовка текстовых данных для классификации моделями машинного обучения	1	
	ЛЗ 6. Создание программы подготовки текстовых данных для классификации моделями машинного обучения	2	7
7	**Л 7.** Классификация текстовых данных моделями машинного обучения	1	
	ЛЗ 7. Создание программы классификации текстов моделями Наивного Байеса, Логистической регрессии, Дерева решений, Случайного леса и т.д.	2	12
	СРОП 5. Прием СРО 2		25
Рубежный контроль 1			**100**
8	**Л 8.** Классификация текстовых данных нейронными сетями	1	
	ЛЗ 8. Создание программы классификации текстов моделями Deep neural network, Convolutional neural network и Long short term memory neural network	2	5
	СРОП 6. Консультация по выполнению СРО 3 на тему «Разработка программы анализа тональности текстов с помощью BERT»		
9	**Л 9.** Большие языковые модели BERT, GPT	1	
	ЛЗ 9. Создание программы обработки текстовых данных моделью BERT	2	10
10	**Л 10.** Анализ и обработка текстов с помощью запросов ChatGPT	1	
	ЛЗ 10. Создание программы обработки текстов с API ChatGPT	2	
	СРОП 7. Прием СРО 3		25
МОДУЛЬ 3 Работа с парсингом текстовых данных			
11	**Л 11.** Основные методы анализа текстовых документов в HTML формате	1	
	ЛЗ 11. Создание программы парсинга текстов библиотекой BeautifulSoup	2	5
	СРОП 8. Консультация по выполнению СРО 4 на тему «Создание приложения веб-краулера»		
12	**Л12.** Работа с парсингом HTML страниц с помощью BeautifulSoup и Scrapy	1	
	ЛЗ 12. Создание программы парсинга текста библиотекой Scrapy	2	5
13	**Л 13.** Работа с веб-краулерами	1	
	ЛЗ 13. Создание программы парсинга текста с помощью Selenium web driver	2	5
	СРОП 9. Прием СРО 4		25
14	**Л 14.** Добавление функций поиска данных в социальных сетях	1	
	ЛЗ 14. Разработка многофункционального веб-краулера	2	10
15	**Л 15.** Основные этапы создания веб-приложения с моделями машинного обучения	1	
	ЛЗ 15. Разработка веб-приложения на Django	2	10
Рубежный контроль 2			**100**
Итоговый контроль (экзамен)			**100**
ИТОГО за дисциплину			**100**

РУБРИКАТОР СУММАТИВНОГО ОЦЕНИВАНИЯ

КРИТЕРИИ ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ

СРО 1. Реализация проекта с базовыми операциями обработки текстов (25% от 100% РК1)

Критерий	**«Отлично»** 21-25%	**«Хорошо»** 11-20%	**«Удовлетворительно»** 6-10%	**«Неудовлетворительно»** 0-5%
Знание и понимание основных концепций обработки текстовых данных	Понимание степени соответствия, актуальности и достоверности найденных данных. Знание и понимание всех основных элементов и операций обработки текстовых данных	Понимание степени соответствия, актуальности и достоверности найденных данных. Знание больше части операций обработки текстовых данных	Ограниченное понимание степени соответствия, актуальности и достоверности элементов и операций обработки текстовых данных	Поверхностное понимание/ отсутствие понимания степени соответствия, актуальности и достоверности найденных данных. Отсутствие знания элементов и операций обработки текстовых данных
Навыки написания программного кода обработки текстовых данных	Четкое и ясное представление программного кода, отсутствие в коде синтаксических ошибок	В программном коде имеются небольшие логические ошибки	Большое количество логических и синтаксических ошибок в прогрммном коде, что делают его практически неработоспособным	Отсутствие программного кода или наличие нескольких строк кода
Написание отчета	Письмо демонстрирует ясность, лаконичность и правильность.	Письмо демонстрирует ясность, лаконичность и корректность. В основном отсутствуют ошибки.	В письме есть некоторые ключевые ошибки, и ясность нуждается в улучшении.	Написанное неясно, трудно следовать за содержанием. Много ошибок в тексте

СРО2. Создание программы классификации текстовых данных (25% от 100% РК1)

Критерий	**«Отлично»** 21-25%	**«Хорошо»** 11-20%	**«Удовлетворительно»** 6-10%	**«Неудовлетворительно»** 0-5%
Работа с моделями машинног обучения классификации текстовых данных	Понимание степени соответствия, актуальности и достоверности работы с данными в приложении. Знание и понимание всех основных операций классификации текстовых данных моделями машинного обучения	Понимание степени соответствия, актуальности и достоверности найденных данных. Знание больше части операций создания моделей машинного обучения	Ограниченное понимание степени соответствия, актуальности и достоверности операций по созданию моделей машинного обучения	Поверхностное понимание/ отсутствие понимания степени соответствия, актуальности и достоверности работы с базами данных. Отсутствие знания операций создания моделей машинного обучения
Навыки написания программного кода	Четкое и ясное представление программного кода, отсутствие в коде синтаксических ошибок	В программном коде имеются небольшие логические ошибки	Большое количество логических и синтаксических ошибок в прогрммном коде, что делают его практически неработоспособным	Отсутствие программного кода или наличие нескольких строк кода
Написание отчета	Письмо демонстрирует ясность, лаконичность и правильность.	Письмо демонстрирует ясность, лаконичность и корректность. В основном отсутствуют ошибки.	В письме есть некоторые ключевые ошибки, и ясность нуждается в улучшении.	Написанное неясно, трудно следовать за содержанием. Много ошибок в тексте

СРО 3. Разработка программы анализа тональности текстов с помощью BERT (25% от 100% РК2)

Критерий	**«Отлично»** 21-25%	**«Хорошо»** 11-20%	**«Удовлетворительно»** 6-10%	**«Неудовлетворительно»** 0-5%
Работа с моделями машинного обучения классификации текстовых данных с помощью BERT	Понимание степени соответствия, актуальности и достоверности работы с данными в приложении. Знание и понимание всех основных операций классификации текстовых данных с помощью BERT	Понимание степени соответствия, актуальности и достоверности найденных данных. Знание больше части операций создания моделей BERT	Ограниченное понимание степени соответствия, актуальности и достоверности операций по созданию моделей BERT	Поверхностное понимание/ отсутствие понимания степени соответствия, актуальности и достоверности работы с базами данных. Отсутствие знания операций создания моделей моделей BERT
Навыки написания программного кода	Четкое и ясное представление программного кода, отсутствие в коде синтаксических ошибок	В программном коде имеются небольшие логические ошибки	Большое количество логических и синтаксических ошибок в прогрммном коде, что делают его практически неработоспособным	Отсутствие программного кода или наличие нескольких строк кода
Написание отчета	Письмо демонстрирует ясность, лаконичность и правильность.	Письмо демонстрирует ясность, лаконичность и корректность. В основном отсутствуют ошибки.	В письме есть некоторые ключевые ошибки, и ясность нуждается в улучшении.	Написанное неясно, трудно следовать за содержанием. Много ошибок в тексте

СРО 4. Создание приложения веб-краулера (25% от 100% РК2)

Критерий	**«Отлично»** 21-25%	**«Хорошо»** 11-20%	**«Удовлетворительно»** 6-10%	**«Неудовлетворительно»** 0-5%
Знание и понимание библиотек создания веб-краулера	Понимание степени соответствия, актуальности и достоверности работы с веб-краулером	Понимание степени соответствия, актуальности и достоверности работы с веб-краулером	Ограниченное понимание работы с веб-краулером	Поверхностное понимание/ отсутствие понимания основных операций работы с веб-краулером
Навыки написания программного кода	Четкое и ясное представление программного кода, отсутствие в коде синтаксических ошибок	В программном коде имеются небольшие логические ошибки	Большое количество логических и синтаксических ошибок в прогрммном коде, что делают его практически неработоспособным	Отсутствие программного кода или наличие нескольких строк кода
Написание отчета	Письмо демонстрирует ясность, лаконичность и правильность.	Письмо демонстрирует ясность, лаконичность и корректность. В основном отсутствуют ошибки.	В письме есть некоторые ключевые ошибки, и ясность нуждается в улучшении.	Написанное неясно, трудно следовать за содержанием. Много ошибок в тексте

И.о. декана _____ Тұрар О.Н.

Заведующий кафедрой _____ Мусиралиева Ш.Ж.

Лектор _____ Карюкин В.И.