

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)		
102322 Технологии программирования для NLP	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
Офлайн	БД, КВ	Проблемно-ориентированный	Изучение основных концепций в области программирования для NLP	Устный офлайн
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)
Изучить применение основных концепций языка программирования Python для работы с задачами анализа данных и машинного обучения в области NLP	1. (когнитивный) Теоретические и методологические концепции NLP	1.1 Понимает основные концепции работы с естественным языком
		1.2 Анализирует особенности подходов контекстно-свободной грамматики
		1.3 Применяет методы работы исчисления предикатов
	2. (функциональный) Подходы предобработки текстовых данных	2.1 Использует знания предобработки текстов
		2.2 Применяет библиотеку NLTK для стемминга и лемматизации
		2.3 Формирует навыки векторизации текстов с помощью методов TF-IDF, Word2Vec, FastText и других
	3.(функциональный) Программа классификации текстов	3.1 Применяет библиотеки scikit-learn и tensorflow для обработки и классификации текстов
		3.2 Создает полнофункциональную программу классификации текстов
		3.3 Определяет точность классификации текстов с помощью метрик accuracy, precision, recall и F1-score
	4. (системный) Программы парсера и краулинга	4.1 Создает конфигурацию веб-краулера
		4.2 Разрабатывает скрипт выборки источников данных для парсинга
		4.3 Выполняет парсинг HTML страниц
	5. (системный) Создавать веб-приложения, использующие модели машинного обучения	5.1 Создает основной каркас веб-приложения на HTML, CSS и JavaScript
		5.2 Создает подключение моделей машинного обучения для определения угроз

		5.3 Создает полную конфигурацию работы веб-приложения
Пререквизиты	Формальные грамматики, Информационные технологии для NLP, Языковые ресурсы, Анализ языка	
Постреквизиты	Интеллектуальный анализ данных, Методы поиска и извлечения информации, Понимание естественного языка, Модели и методы нейронных сетей в NLP, Глубокое обучение	
Учебные ресурсы	<p>Литература:</p> <p>Основная:</p> <ul style="list-style-type: none"> – Python for Everybody: Exploring Data in Python 3 by Dr. Charles Russell Severance, Sue Blumenberg, Elliott Hauser, Aimee Andrión, 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. <p>Дополнительная:</p> <ul style="list-style-type: none"> – 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 <p>Профессиональные научные базы данных:</p> <ul style="list-style-type: none"> – Лабораторная аудитория 514 – Лабораторная аудитория 323 <p>Интернет–ресурсы:</p> <ul style="list-style-type: none"> – 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/ <p>Программное обеспечение: Python IDE, Anaconda Navigator Python, NLTK, Microsoft Office Word, WinRAR, Power Point, Adobe Reader, Paint.</p>	
Академическая политика дисциплины	<p>Академическая политика дисциплины определяется <u>Академической политикой и Политикой академической честности КазНУ имени аль-Фараби</u>. Документы доступны на главной странице ИС Univer.</p> <p>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.</p> <p>Посещаемость. Дедлайн каждого задания указан в календаре (графике) реализации содержания дисциплины. Несоблюдение дедлайнов приводит к потере баллов.</p> <p>Академическая честность. Практические/лабораторные занятия, СРО развивают у обучающегося самостоятельность, критическое мышление, креативность. Недопустимы плагиат, подлог, использование шпаргалок, списывание на всех этапах выполнения заданий.</p> <p>Соблюдение академической честности в период теоретического обучения и на экзаменах помимо основных политик регламентируют <u>«Правила проведения итогового контроля»</u>, <u>«Инструкции для проведения итогового контроля осеннего/весеннего семестра текущего учебного года»</u>, <u>«Положение о проверке текстовых документов обучающихся на наличие заимствований»</u>. Документы доступны на главной странице ИС Univer.</p> <p>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 обучающимся друг к другу независимо от</p>	

	<p>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.</p> <p>Все обучающиеся, особенно с ограниченными возможностями, могут получить консультативную помощь по телефону/ e-mail vladislav.karyukin@gmail.com / +77019405992 либо посредством видеосвязи</p> <p style="text-align: center;">В MS Teams https://teams.microsoft.com/l/team/19%3Ail51PNbgOQSGfdP4TYZTqxaUB0UQGBOaqrcLW2k_2Yk1%40thread.tacv2/conversations?groupId=ff412368-4e85-4174-bceb-858360e504b2&tenantId=b0ab71a5-75b1-4d65-81f7-f479b4978d7b</p>
--	--

ИНФОРМАЦИЯ О ПРЕПОДАВАНИИ, ОБУЧЕНИИ И ОЦЕНИВАНИИ

Балльно-рейтинговая буквенная система оценки учета учебных достижений			Методы оценивания															
Оценка	Цифровой эквивалент баллов	Баллы, % содержание	Оценка по традиционной системе	<p>Критериальное оценивание – процесс соотнесения реально достигнутых результатов обучения с ожидаемыми результатами обучения на основе четко выработанных критериев. Основано на формативном и суммативном оценивании.</p> <p>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.</p> <p>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.</p> <table border="1" style="width: 100%; margin-top: 10px;"> <thead> <tr> <th>Формативное и суммативное оценивание</th> <th>Баллы % содержание</th> </tr> </thead> <tbody> <tr> <td>Активность на лекциях</td> <td>0</td> </tr> <tr> <td>Работа на практических занятиях</td> <td>25</td> </tr> <tr> <td>Самостоятельная работа</td> <td>25</td> </tr> <tr> <td>Проектная и творческая деятельность</td> <td>10</td> </tr> <tr> <td>Итоговый контроль (экзамен)</td> <td>40</td> </tr> <tr> <td>ИТОГО</td> <td>100</td> </tr> </tbody> </table>	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
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																	
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																
C-	1,67	60-64																
D+	1,33	55-59																
D	1,0	50-54	Удовлетворительно															
FX	0,5	25-49																
F	0	0-24																
			Неудовлетворительно															

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

Неделя	Название темы	Кол-во часов	Макс. балл
МОДУЛЬ 1 Основные концепции NLP			
1	Л 1. Введение в NLP	1	
	ЛЗ 1. Основные операции в NLP	2	5
2	Л 2. Основные концепции контекстно-свободной грамматики	1	
	ЛЗ 2. Создание программы обработки данных контекстно-свободной грамматики	2	5
	СРОП 1. Консультации по выполнению СРО1 на тему «Реализация проекта с базовыми операциями NLP»		
3	Л 3. Обзор методов исчисления предикатов	1	
	ЛЗ 3. Создание программы исчисления предикатов	2	7
	СРОП 2. Прием СРО 1		20
4	Л 4. Обзор операций предобработки текстов	1	
	ЛЗ 4. Создание программы предобработки текстов	2	7
	СРОП 3. Проведение коллоквиума по темам 1-3 недель		5
5	Л 5. Выполнение операции векторизации текстовых данных	1	
	ЛЗ 5. Создание программы векторизации с помощью tf-idf, Word2Vec	2	7
	СРОП 4. Консультация по выполнению СРО 2 на тему «Создание программы обработки текстовых данных»		
МОДУЛЬ 2 Обработка текстовых данных моделями машинного обучения			
6	Л 6. Классификация текстов	1	
	ЛЗ 6. Создание программы классификации текстов с применением библиотек scikit-learn и tensorflow	2	7

7	Л 7. Модели машинного обучения Наивный Байеса, Логистическая регрессия, Дерево решений, Случайный лес и т.д.	1	
	ЛЗ 7. Создание программы классификации моделями машинного обучения	2	12
	СРОП 5. Прием СРО 2		25
Рубежный контроль 1			100
8	Л 8. Нейронные сети Deep neural network, Convolutional neural network и Long short term memory neural network	1	
	ЛЗ 8. Создание программы классификации текстов нейронными сетями	2	5
	СРОП 6. Консультация по выполнению СРО 3 на тему «Разработка программы анализа тональности текстов»		
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 NLP (20% от 100% РК1)

Критерий	«Отлично» 16-20%	«Хорошо» 11-15%	«Удовлетворительно» 6-10%	«Неудовлетворительно» 0-5%
Знание и понимание основных концепций обработки текстовых данных операциями NLP	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 операциями NLP	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 программном 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 обработки текстовых данных (25% от 100% РК1)

Критерий	«Отлично» 21-25%	«Хорошо» 11-20%	«Удовлетворительно» 6-10%	«Неудовлетворительно» 0-5%
Работа программы обработки текстовых данных	Понимание степени соответствия, актуальности и достоверности работы с данными в приложении. Знание и понимание всех основных 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 операций обработки текстовых данных	Поверхностное понимание/ отсутствие понимания степени соответствия, актуальности и достоверности работы с базами данных. Отсутствие знания операций обработки текстовых данных
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 программном 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 (25% от 100% РК2)

Критерий	«Отлично» 21-25%	«Хорошо» 11-20%	«Удовлетворительно» 6-10%	«Неудовлетворительно» 0-5%
Работа с моделями анализа тональности текстов	Понимание степени соответствия, актуальности и достоверности работы с данными в приложении. Знание и понимание всех основных операций анализа тональности текстов	Понимание степени соответствия, актуальности и достоверности найденных данных. Знание больше части операций анализа тональности текстов	Ограниченное понимание степени соответствия, актуальности и достоверности операций по анализу тональности текстов	Поверхностное понимание/ отсутствие понимания степени соответствия, актуальности и достоверности работы с базами данных. Отсутствие знания операций анализа тональности текстов
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 программном 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 программном 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.