Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное
образовательное учреждение высшего образования
«Казанский (Приволжский) федеральный университет»
Институт филологии и межкультурной коммуникации

УТВЕРЖДАЮ
Проректор по образовательной деятельности КФУ
_____ А. Турилова
"___" _____ 20___ г.

Дополнительная профессиональная программа профессиональной переподготовки
Методы лингвистического моделирования с использованием языка программирования Python

Утверждена Ученым советом Института филологии и межкультурной коммуникации КФУ (протокол № 9 от «19» апреля 2023 г.)
Председатель Ученого совета Замалетдинов Радиф Рифкатович _____
(подпись)

Руководитель подразделения,
реализующего ДПП ПП _____ Р.Р. Замалетдинов
(подпись) (инициалы, фамилия)

«____»_____ 20__ г.

Казань 2023

I. Общие положения

1. Дополнительная профессиональная программа (программа профессиональной переподготовки) ИТ-профиля «Методы лингвистического моделирования с использованием языка программирования Python» (далее – Программа) разработана в соответствии с нормами Федерального закона РФ от 29 декабря 2012 года № 273-ФЗ «Об образовании в Российской Федерации», с учетом требований приказа Минобрнауки России от 1 июля 2013 г. № 499 «Об утверждении Порядка организации и осуществления образовательной деятельности по дополнительным профессиональным программам», с изменениями, внесенными приказом Минобрнауки России от 15 ноября 2013 г. № 1244 «О внесении изменений в Порядок организации и осуществления образовательной деятельности по дополнительным профессиональным программам, утвержденный приказом Министерства образования и науки Российской Федерации от 1 июля 2013 г. № 499», приказа Министерства образования и науки РФ от 23 августа 2017 г. N 816 «Об утверждении Порядка применения организациями, осуществляющими образовательную деятельность, электронного обучения, дистанционных образовательных технологий при реализации образовательных программ» (указать при необходимости); паспорта федерального проекта «Развитие кадрового потенциала ИТ-отрасли» национальной программы «Цифровая экономика Российской Федерации»; постановления Правительства Российской Федерации от 13 мая 2021 г. № 729 «О мерах по реализации программы стратегического лидерства «Приоритет-2030» (в редакции постановления Правительства Российской Федерации от 14 марта 2022 г. № 357 «О внесении изменений в постановление Правительства Российской Федерации от 13 мая 2021 г. № 729»); приказа Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации от 28 февраля 2022 г. № 143 «Об утверждении методик расчета показателей федеральных проектов национальной программы «Цифровая экономика Российской Федерации» и признании утратившими силу некоторых приказов Министерства цифрового

развития, связи и массовых коммуникаций Российской Федерации об утверждении методик расчета показателей федеральных проектов национальной программы «Цифровая экономика Российской Федерации» (далее – приказ Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации № 143); федерального государственного образовательного стандарта 45.03.04 Интеллектуальные системы в гуманитарной сфере, утвержденного приказом Минобрнауки России от 24.04.2018№ 324, (далее вместе – ФГОС ВО)), а также профессионального стандарта 06.017 «Руководитель разработки программного обеспечения», утвержденного приказом Министерства труда и социальной защиты РФ от 20 июля 2022 г. № 423н.

2. Профессиональная переподготовка заинтересованных лиц (далее – Слушатели), осуществляемая в соответствии с Программой (далее – Подготовка), имеющей отраслевую направленность[1] Информационно-коммуникационные технологии, проводится в ФГАОУ ВО «Казанский (Приволжский) федеральный университет (далее – Университет) в соответствии с учебным планом в очной форме обучения[2].

3. Разделы, включенные в учебный план Программы, используются для последующей разработки календарного учебного графика, учебно-тематического плана, рабочей программы, оценочных и методических материалов. Перечисленные документы разрабатываются Университетом самостоятельно, с учетом актуальных положений законодательства об образовании, законодательства в области информационных технологий и смежных областей знаний ФГОС ВО и профессионального стандарта «Руководитель разработки программного обеспечения».

4. Программа регламентирует требования к профессиональной переподготовке в области осуществления деятельности по организации и управлению процессами разработки, отладки, проверки работоспособности и

[1] Варианты отраслевой направленности: «Городское хозяйство»; «Финансовые услуги»; «Строительство»; «Добывающая промышленность»; «Обрабатывающая промышленность»; «Транспортная инфраструктура»; «Здравоохранение»; «Энергетическая инфраструктура»; «Образование»; «Сельское хозяйство и агропромышленный комплекс»; «Информационно-коммуникационные технологии»; «Искусство и культура»

[2] При реализации Программы допускается использовать сетевую форму обучения с организациями реального сектора экономики субъекта Российской Федерации

модификации компьютерного программного обеспечения. Срок освоения Программы составляет 9 месяцев, трудоемкость – 252 академических часа.

К освоению Программы в рамках проекта допускаются лица:

- получающие высшее образование по очной (очно-заочной) форме, лица, освоившие основную профессиональную образовательную программу (далее – ОПОП ВО) бакалавриата – в объеме не менее первого курса (бакалавры 2-го курса), ОПОП ВО специалитета – не менее первого и второго курсов (специалисты 3-го курса). Также к освоению ДПП ПП допускаются лица, обучающиеся по программам магистратуры, которые не относятся к ИТ-профилю (согласно приложению к Методике расчета показателя граждан, прошедших обучение по дополнительным образовательным программам) и по программам ординатуры.

5. Область профессиональной деятельности 06 «Связь, информационные и коммуникационные технологии».

II. Цель

6. Целью подготовки слушателей по Программе является получение компетенции,[3] необходимой для выполнения нового вида профессиональной деятельности в области информационных технологий в лингвистике, обучающимися по специальностям и направлениям подготовки, не отнесенным к ИТ-сфере; приобретение новой квалификации «Руководитель разработки программного обеспечения».

III. Характеристика новой квалификации и связанных с ней видов профессиональной деятельности, трудовых функций и (или) уровней квалификации

7. Виды профессиональной деятельности, трудовая функция, указанные в

[3]Указать целевые группы обучающихся, определенные паспортом Федерального проекта: − обучающиеся по специальностям и направлениям подготовки, не отнесенным к ИТ-сфере, −обучающиеся по специальностям и направлениям подготовки ИТ-сферы (выбрать нужное)

профессиональном стандарте по соответствующей должности «Руководитель группы разработки», представлены в таблице 1:

Таблица 1

Характеристика новой квалификации, связанной с видом профессиональной деятельности и трудовыми функциями в соответствии с профессиональным стандартом «Руководитель разработки программного обеспечения»

Область профессиональной деятельности	Тип задач профессиональной деятельности	Код и наименование профессиональной компетенции	Трудовые действия	Трудовая функция	Обобщенная трудовая функция	Вид профессиональной деятельности
Связь, информационные и коммуникационные технологии	Проектная	**ПК-1** Разработка документации к программным системам и стандартам в области лингвистики	Выявление первоначальных требований заказчика к ИС Информирование заказчика о возможностях типовой ИС и вариантах ее модификации Определение возможности достижения соответствия ИС первоначальным требованиям заказчика Составление протокола переговоров с заказчиком	Руководство разработкой проектной и технической документации на компьютерное программное обеспечение	Руководство процессами разработки компьютерного программного обеспечения	Руководство разработкой компьютерного программного обеспечения

Таблица 2

Характеристика новой и развиваемой цифровой компетенции в ИТ-сфере, связанной с уровнем формирования и развития в результате освоения Программы[4] «Методы лингвистического моделирования с использованием языка программирования Python»

Наименование сферы	**Код и наименование профессиональной компетенции**	**Пример инструментов**	**0 — способность не проявляется/ проявляется в степени, недостаточной для отнесения к 1 уровню сформированности компетенции**	**1 — способность проявляется под внешним контролем / при внешней постановке задачи/ обучающийся пользуется готовыми, рекомендованными продуктами**	**2 — способность проявляется, но обучающийся эпизодически прибегает к экспертной консультации/ самостоятельно подбирает и пользуется готовыми продуктами**	**3 — способность проявляется системно / обучающийся модифицирует способность под определенные задачи / создает новый продукт, обучает других**
Стандарты и методики в ИТ	**ПК-2 (ID-28)** Применяет языки программирования для решения	Python	+	+	+	-

[4] На основании Матрицы компетенций, актуальных для цифровой экономики, указанной в Приложении 1 в Требованиях к ДПП ПП.

	профессиональных задач					
Информационные ресурсы и продукты	**ПК-3 (ID-247)** Оценивает результаты обучения с использованием цифровых ресурсов и продуктов	Moodle, 1С	+	+	+	-

IV. Характеристика новых и развиваемых цифровых компетенций, формирующихся в результате освоения программы

8. В ходе освоения Программы Слушателем приобретаются следующие профессиональные компетенции:

ПК-1 Разработка документации к программным системам и стандартам в области лингвистики.

9. В ходе освоения Программы Слушателем совершенствуются следующие профессиональные компетенции:

ПК-2 Применяет языки программирования для решения профессиональных задач;

ПК-3 Оценивает результаты обучения с использованием цифровых ресурсов и продуктов.

V. Планируемые результаты обучения по ДПП ПП

10. Результатами подготовки слушателей по Программе является получение компетенции, необходимой для выполнения нового вида профессиональной деятельности в области информационных технологий обучающимися по специальностям и направлениям подготовки, не отнесенным к ИТ-сфере; приобретение новой квалификации «Руководитель разработки программного обеспечения».

Наименование компетенции: Разрабатывает документацию к программным системам и стандартам в области лингвистики.

Знать:

- правила редактирования научно-технической документации
- нормативно-технические документы (стандарты и регламенты), определяющие требования к проектной и технической документации
- методы повышения читаемости программного кода
- технологии межличностной и групповой коммуникации в деловом взаимодействии, основы конфликтологии.

Уметь:

- применять нормативно-технические документы (стандарты и регламенты), определяющие требования к проектной и технической документации на компьютерное программное обеспечение
- применять коллективную среду документирования программного обеспечения
- применять методы принятия управленческих решений
- осуществлять коммуникации с заинтересованными сторонами
- работать с источниками информации, необходимой для профессиональной деятельности
- вести деловую переписку.

Иметь навыки:

- межличностной и групповой коммуникации в деловом взаимодействии, основы конфликтологии
- управления взаимоотношениями с клиентами и заказчиками (CRM)
- ведения документооборота в организациях
- культуры речи
- управления содержанием проекта: документирование требований, анализ продукта, модерируемые совещания.

Наименование компетенции: Применяет языки программирования для решения профессиональных задач.

Знать:

- технологии программирования
- синтаксис выбранного языка программирования
- особенности выбранной среды программирования
- средства проверки и отладки программного кода
- интерфейсы взаимодействия с внешней средой
- основы современных систем управления базами данных.

Уметь:

- писать программный код на выбранном языке программирования

- анализировать значения полученных характеристик программного обеспечения
- выбирать средства реализации требований к программному обеспечению.

Иметь навыки:[5]

- использовать выбранную среду программирования
- применять инструментарий для создания и актуализации исходных текстов программ.

Наименование компетенции: Оценивает результаты обучения с использованием цифровых ресурсов и продуктов.

Знать:

- возможности типовой ИС
- предметную область автоматизации
- современные стандарты информационного взаимодействия систем
- современные подходы и стандарты автоматизации организации (например, CRM, MRP, ERP..., ITIL, ITSM)
- способы оценки образовательных результатов
- цифровые технологии для оценки результатов обучения.

Уметь:

- создавать проверочные задания с использованием цифровых технологий
- проводить корректировку обучения на основе анализа данных

Иметь навыки:

- использования рекомендованных цифровых продуктов для оценки результатов обучения

VI. Организационно-педагогические условия реализации ДПП

[5] планируемые результаты по компетенциям, указанным в Таблице 1 и 2, прописываются по отдельности в разрезе каждой компетенции.

12. Реализация Программы должна обеспечить получение компетенции, необходимой для выполнения нового вида профессиональной деятельности в области информационных технологий, обучающимися по специальностям и направлениям подготовки, не отнесенным к ИТ-сфере; приобретение новой квалификации «Руководитель разработки программного обеспечения».

13. Учебный процесс организуется с применением[6] электронного обучения и дистанционных образовательных технологий, инновационных технологий и методик обучения, способных обеспечить получение слушателями знаний, умений и навыков в области[7] разработки программного обеспечения.

14. Реализация Программы обеспечивается научно-педагогическими кадрами Университета, допустимо привлечение к образовательному процессу высококвалифицированных специалистов ИТ-сферы и/или дополнительного профессионального образования в части, касающейся профессиональных компетенций в области создания алгоритмов и программ, пригодных для практического применения, с обязательным участием представителей профильных организаций-работодателей. Возможно привлечение региональных руководителей цифровой трансформации (отраслевых ведомственных и/или корпоративных) к проведению итоговой аттестации, привлечение работников организаций реального сектора экономики субъектов Российской Федерации. Не менее 50% общего объема аудиторных часов в рамках ДПП ПП реализуются научно-педагогическими работниками, отвечающими следующим критериям:

− наличие высшего профильного образования в ИТ-сфере и/или дополнительного профессионального образования в части, касающейся профессиональных компетенций в области создания алгоритмов и программ, пригодных для практического применения;

− наличие стажа педагогической работы в образовательных организациях

[6] При необходимости указать нужное — электронного обучения, дистанционных образовательных технологий
[7] Разрабатывается на основе ФГОС ВО (3++), соответствует разделу 1.11 ФГОС ВО и конкретному профстандарту

высшего образования Российской Федерации и/или стажа практической работы в профильной организации ИТ-отрасли не менее 3 лет.

Не менее 20% от общего объема аудиторных часов в рамках ДПП ПП реализуются лицами, имеющими подтвержденный стаж в профессии в ИТ-сфере или в отрасли цифровой экономики не менее двух лет, полученный не более четырех лет назад.

VII. Учебный план ДПП

15. Объем Программы составляет 9 месяцев, трудоемкость – 252 часа.

16. Учебный план Программы определяет перечень, последовательность, общую трудоемкость разделов и формы контроля знаний.

Учебный план программы профессиональной переподготовки «Методы лингвистического моделирования с использованием языка программирования Python»

№ п/п	Наименование раздела (модуля)	Общая трудоемкость (252 часа)	Форма контроля
1.	ИТ-Модуль (лекции + практические занятия)	54	Домашняя работа
2.	Лингвистический модуль (лекции + практические занятия)	54	Домашняя работа
3.	Самостоятельная работа	54	Домашняя работа
4.	Промежуточная аттестация	**6**	Контрольные работы и тесты
5.	Проектная практика	54	Отчет по практике
6.	Итоговая аттестация	**30**	Защита проекта
	Итого:	252	

VIII. Календарный учебный график

18. Календарный учебный график представляет собой график учебного процесса, устанавливающий последовательность и продолжительность обучения и итоговой аттестации по учебным дням.

Календарный учебный график программы профессиональной переподготовки «Методы лингвистического моделирования с использованием языка программирования Python»

		Учебные недели																																					
№	Наименование раздела (модуля)	1	2	3	4	5	6	7	8	9	10	11	12	13.	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1.	ИТ-Модуль: Обработка текстов на естественном языке, конвейер обработки текста, токенизация, лемматизация, частеречная разметка, распознавание именованных сущностей	+	+	+	+	+	+	+	+	+	+	+	+	+																							
2.	Квантитативная лингвистика, математические методы в лингвистике, компьютерные технологии в лингвистике, современные цифровые технологии текстовой аналитики														+	+	+	+	+	+	+	+	+	+	+	+											
3.	Проектная практика																									+	+	+	+	+	+	+	+	+	+		
4.	Промежуточная аттестация													+												+											
5.	Итоговая аттестация																													+	+	+	+	+	+	+	+

IX. Рабочая программа учебных предметов, курсов, дисциплин (модулей)

19. Рабочая программа содержит перечень разделов и тем, а также рассматриваемых в них вопросов с учетом их трудоемкости.

Рабочая программа разрабатывается Университетом с учетом профессионального стандарта «Руководитель разработки программного обеспечения».

№ п/п	Наименование и краткое содержание раздела(модуля)	Объем, часов
1.	ИТ-Модуль *Основные темы:* Обработка текстов на естественном языке, Конвейер обработки текста, Токенизация, Лемматизация, Частеречная разметка, Распознавание именованных сущностей *Краткое содержание:* Как компьютеры понимают естественный язык Применение машинного обучения для обработки естественного языка Что такое статистическая модель в NLP Настройка рабочей среды Установка статистических моделей для библиотеки spaCy Базовые операции NLP в библиотеке spaCy Использование лемматизации для распознавания смысла Поиск соответствующих глаголов с помощью тегов частей речи Важность контекста Синтаксические отношения	54
2.	Лингвистический Модуль *Основные темы:* Квантитативная лингвистика, Математические методы в лингвистике, Компьютерные технологии в лингвистике, Современные цифровые технологии текстовой аналитики *Краткое содержание:* Ключевые понятия квантитативной лингвистики Метод статистического анализа текста Лингвистические аспекты разработок в области искусственного интеллекта Частотный словарь как структурно-вероятностная модель языка и речи Методика, сущность, этапы подготовки и проведения контент-анализа Лингвистический анализ текста Жанрово-стилевая организация текста Базовые категории и свойства текста	54

3.	Самостоятельная работа *Краткое содержание:* Обучающиеся самостоятельно выполняют домашнюю работу по каждой теме.	54
4.	Промежуточная аттестация *Краткое содержание:* Контрольные работы и тесты по пройдённым темам	**6**
5.	Проектная практика *Краткое содержание:* Практика проводится на базе организаций различных организационно-правовых форм и форм собственности или их основных структурных подразделений, осуществляющих деятельность, соответствующую виду (видам) деятельности, к которому (которым) готовится обучающийся.	54
6.	Итоговая аттестация *Краткое содержание:* В ходе итоговой аттестации обучающиеся обеспечивают презентацию (защиту) разработанного цифрового решения (проекта), а также перечня решаемых им проблем. Оценивается использование технологий, изученных в курсе, самостоятельность выполняемого решения, степень участия каждого члена команды в разработке.	**30**

20. Учебно-тематический план Программы определяет тематическое содержание, последовательность разделов и (или) тем и их трудоемкость.

№ п/п	Наименование раздела(модуля)	Количество часов		
		аудиторных		самостоятельной работы (выполнение домашних заданий)
		Лекции	Практика	
1.	ИТ-Модуль	10	44	54
2.	Лингвистический Модуль	10	44	54
3.	Самостоятельная работа (домашнее задание)	54		
4.	Промежуточная аттестация	**6**		
5.	Проектная практика	54		

6.	Итоговая аттестация	**30**

X. Формы аттестации

21. Слушатели, успешно выполнившие все элементы учебного плана, допускаются к итоговой аттестации.

Итоговая аттестация по Программе проводится в форме демонстрационного экзамена.

22. Лицам, успешно освоившим Программу (в области создания алгоритмов и программ, пригодных для практического применения, или навыков использования и освоения цифровых технологий, необходимых для выполнения нового вида профессиональной деятельности) и прошедшим итоговую аттестацию в рамках проекта «Цифровые кафедры», выдается документ о квалификации: диплом о профессиональной переподготовке.

При освоении ДПП ПП параллельно с получением высшего образования диплом о профессиональной переподготовке выдается не ранее получения соответствующего документа об образовании и о квалификации (за исключением лиц, имеющих среднее профессиональное или высшее образование).

23. Лицам, не прошедшим итоговую аттестацию или получившим на итоговой аттестации неудовлетворительные результаты, а также лицам, освоившим часть Программы и (или) отчисленным из Университета, выдается справка об обучении или о периоде обучения по образцу, самостоятельно устанавливаемому Университетом.

XI. Оценочные материалы

24. Контроль знаний, полученных слушателями при освоении разделов (модулей) Программы, осуществляется в следующих формах:

- текущий контроль успеваемости – обеспечивает оценивание хода освоения разделов Программы, проводится в форме проверки домашнего

задания;

- промежуточная аттестация – завершает изучение отдельного модуля Программы, проводится в форме контрольных работ и тестирования;

- итоговая аттестация – завершает изучение всей программы.

25. В ходе освоения Программы каждый слушатель выполняет следующие отчетные работы:

№ п/п	Наименование раздела (модуля)	Задание	Критерии оценки
1.	ИТ-Модуль	Домашнее задание	50% процентов выполненных требований к заданию
2.	Лингвистический Модуль	Домашнее задание	50% процентов выполненных требований к заданию
3.	Промежуточная аттестация	Контрольная работа или тестирование после каждого раздела	60% правильных ответов на контрольные работы и тесты
4.	Проектная практика	Прохождение практики	Отчет по практике
5.	Итоговая аттестация	Выполнение проекта	Оценка аттестационной комиссии на основе выполнения требований, указанных в описании проекта

26. Текущий контроль. Перечень примерных заданий
26.1. ИТ-Модуль

1. Изучить программное обеспечение для машинного перевода, которое использует обработку естественного языка для преобразования текста или речи с одного языка на другой с сохранением контекстуальной точности. Сервис AWS, поддерживающий машинный перевод, – Amazon Translate.

1) Выполнить предобработку текстов коллекции

2) Построить статистическую модель

3) Обучить нейронную сеть

4) Провести классификацию текстов

Жизненный цикл обычной системы машинного обучения включает три этапа: обучение модели, контроль и выполнение предсказаний.

Обработка текстов на естественном языке (Natural Language Processing, NLP) — направление искусственного интеллекта, нацеленное на обработку и анализ данных на естественном языке и обучение машин взаимодействию с людьми на естественном языке (языке, сформировавшемся естественным путем на протяжении истории).

2. Используя предложенные варианты, попробуйте улучшить качество определения частей речи:

1. Экспериментировать с параметрами и архитектурой - количеством каналов (размерностью эмбеддинга), глубиной нейросети, силой Dropout, добавить BatchNorm или другую нормализацию.

2. Подключить прореженные (dilated) свёртки, чтобы увеличить рецептивное поле без увеличения числа параметров.

3. Добавить взвешивание классов.

4. Использовать в качестве обозначения начала и конца слова не 0, а какой-нибудь другой токен (для 0 nn.Embedding всегда выдаёт нулевой вектор, а в этом случае для начала а конца слова будут учиться специальные вектора).

3. Ознакомиться с инструментом NLP для Python, который называется Natural Language Toolkit (NLTK). Выполнить следующие шаги:

1) Импортирование NLTK

2) Загрузка данных и разметчика NLTK

3) Токенизация

4) Присвоение тегов

5) Подсчёт тегов

6) Запуск сценария NLP

26.2. Лингвистический Модуль

1. Программа «Wordstat» предназначена для статистического анализа текстов. Обработать можно любой текст, предварительно сохранив его в формате txt или html. В результате работы программы пользователь получает список слов из заданного текста с указанием частоты их употребления в тексте.

На основе программы «Wordstat» определите частоту слов в данном тексте.

Дом, который построил Джек

Вот дом,
Который построил Джек.
А это пшеница,
Которая в тёмном чулане хранится
В доме,
Который построил Джек.
А это весёлая птица-синица,
Которая часто ворует пшеницу,
Которая в тёмном чулане хранится
В доме,
Который построил Джек.
Вот кот,
Который пугает и ловит синицу,
Которая часто ворует пшеницу,
Которая в тёмном чулане хранится
В доме,
Который построил Джек.
и т.д.

Для решения подобных задач можно использовать следующий алгоритм. Для начала создайте файл html в формате с текстом одного автора (откройте «Блокнот»; загрузите нужный текст; в меню «Файл» выберите «Сохранить как» и назовите файл text1.html) и сохраните файл в одной папке сс текстом. Затем откройте программу wordstat.exe и скопируйте туда текст. Если вам необходимо обработать несколько текстов одного автора, обработайте все файлы по очереди (следите, чтобы была включена опция «накапливать сумму результатов). Автоматически откроется файл (по умолчанию) под названием wordstat.txt. В нем вы обнаружите результаты.

2. Задана КС-грамматика G = <V, W, S, R>, где V – терминальный алфавит, W – нетерминальный алфавит, S – аксиома, R – множество правил. Обозначим через L – язык, порождаемый грамматикой G.

Какие из следующих утверждений являются верными?

1) Если в правой части (то есть после стрелки) любого правила из R находятся только символы из W+, то грамматика G порождает непустой язык.

2) Если в грамматике G возможен хотя бы один полный вывод, то грамматика G порождает пустой язык.

3) Если в грамматике G возможен хотя бы один полный вывод, то грамматика G порождает непустой язык.

Решения:

1) Неверное утверждение, так как такое множество правил не позволит вывести терминальную цепочку. Следовательно, язык, порождаемый грамматикой G, пустой.

2) Неверное утверждение, так как полный вывод является выводом терминальной цепочки. Следовательно, язык L(G) содержит, по крайней мере, одну цепочку и не является пустым.

3) Верное утверждение (см. предыдущий пункт задания).. Компьютерные технологии в лингвистике

3. Для разработки морфологического словаря создайте таблицу, где в первой колонке будет записана словоформа, во второй – нормальная форма, а в третьей – набор параметров.

Для морфологического анализа в таком словаре необходимо просто найти все соответствующие словоформы и выдать найденные результаты. Для синтеза требуется найти заданную нормальную форму с требуемым набором параметров и выдать словоформу, находящуюся в той же строке.

27. Промежуточная аттестация. Перечень примерных заданий

27. 1. ИТ-Модуль

1. Контрольная работа.

Создать классификатор текстовых документов, используя среду разработки Jupyter notebook на платформе Google Colab.
В работе понадобятся следующие библиотеки:
1. scikit-learn – свободно распространяемая библиотека на python, содержащая реализации различных методов машинного обучения;
2. nltk – пакет библиотек для символьной и статистической обработки естественного языка;
3. matplotlib – библиотека, содержащая набор инструментов для визуализации данных, — понадобится для отображения «облака слов».

Выполнить токенизацию текста с использованием RegEx (регулярных выражений) в Python.
Текст:
Joseph Arthur was a young businessman. He was one of the shareholders at Ryan Cloud's Start-Up with James Foster and George Wilson. The Start-Up took its flight in the mid-90s and became one of the biggest firms in the United States of America. The business was expanded in all major sectors of livelihood, starting from Personal Care to Transportation by the end of 2000. Joseph was used to be a good friend of Ryan.

Ознакомиться с пакетом инструментов rupostagger для Python. Выполнить следующие шаги:
1) Импортирование rupostagger
2) Загрузка данных и разметчика rupostagger
3) Токенизация
4) Присвоение тегов
5) Подсчёт тегов
6) Запуск сценария rupostagger

2. Тест

1. Какой из следующих методов можно использовать для нормализации ключевых слов в NLP, процесс преобразования ключевого слова в его базовую форму
+а) Лемматизация
б) Soundex
с) Косинус сходства
д) N-грамм
2. N-граммы определяются как комбинация N ключевых слов вместе. Сколько биграмм можно составить из данного предложения: “Analytics Vidhya is a great source to learn data science”

а)7
б)8
+с)9
д)10
3. Какой из следующих методов можно использовать для вычисления расстояния между двумя векторами слов в NLP?
а) Лемматизация
+б) Евклидово расстояние
+с) Косинус сходства
д) N-грамм
4. Каковы возможные особенности корпуса текстов в NLP?
а) Количество слов в документе
б) Векторное обозначение слова
с) Часть речевого тега
д) Базовая грамматика зависимостей
е) Все вышеперечисленное
4. Вы создали матрицу терминов документа на основе входных данных 20 000 документов для модели машинного обучения. Что из перечисленного можно использовать для уменьшения размерности данных?
1 Нормализация ключевых слов.
2 Скрытое семантическое индексирование.
3 Скрытое распределение Дирихле.
А) только 1
Б) 2, 3
В) 1, 3
+Г) 1, 2, 3
5. Какой из методов синтаксического анализа текста можно использовать для обнаружения именной фразы, глагольной фразы, обнаружения субъекта и обнаружения объекта в NLP.
а. Тегирование части речи
б. Пропустить извлечение граммов и N-граммов
в. Непрерывный мешок слов
+г. Анализ зависимостей и анализ групп
6. Различие между словами, выраженное с помощью косинусного сходства, будет иметь значения, значительно превышающие 0,5.
а. Истинный
б. Ложь
7. Что из нижеперечисленного относится к методам нормализации ключевых слов в NLP?
+а. Стемминг
б. Часть речи
в. Распознавание именованных объектов
+г. Лемматизация
8. Что из нижеперечисленного относится к вариантам использования NLP?
а. Обнаружение объектов на изображении
б. Распознавание лиц
в. Речь биометрическая

+г. Обобщение текста
9. В корпусе из N документов один случайно выбранный документ содержит в общей сложности T терминов, а термин «привет» встречается K раз. Каково правильное значение произведения TF (частота терминов) и IDF (обратная частота документа), если термин «привет» встречается примерно в одной трети всех документов?
а. KT * Log(3)
б. T * Log(3) / K
+в. K * Log(3) / T
г. Log(3) / KT
10. В NLP алгоритм уменьшает вес часто используемых слов и увеличивает вес слов, которые редко используются в наборе документов.
а. Термин Частота (TF)
+б. Обратная частота документа (IDF)
в. Word2Vec
г. Скрытое распределение Дирихле (LDA)
11. В NLP процесс удаления таких слов, как «и», «является», «а», «ан», «то» из предложения называется
а. Стемминг
б. лемматизация
+в. Стоп-слово
г. Все вышеперечисленное
12. В NLP процесс преобразования предложения или абзаца в токены называется стеммингом.
а. Истинный
+б. Ложь
13. В NLP токены преобразуются в числа перед передачей в любую нейронную сеть.
+а. Истинный
б. Ложь
14. определить лишнее
а. nltk
б. scikit learn
в. SpaCy
+г. BERT
15. TF-IDF помогает вам установить?
а. наиболее часто встречающееся слово в документе
+б. самое важное слово в документе
16. В NLP процесс идентификации людей, организации из заданного абзаца предложения называется
а. Стемминг
б. Лемматизация
в. Удаление стоп-слов
+д. Распознавание именованных объектов
17. Что из нижеперечисленного не относится к методам предварительной обработки в NLP?
а. Стемминг и лемматизация

б. Преобразование в нижний регистр
в. Удаление знаков препинания
г. Удаление стоп-слов
+д. Анализ настроений
18. При анализе текста преобразование текста в токены, а затем преобразование их в целочисленные векторы или векторы с плавающей запятой можно выполнить с помощью
+а. CountVectorizer
б. TF-IDF
в. Мешок слов
г. NER
19. В NLP слова, представленные в виде векторов, называются нейронными вложениями слов.
+а. Истинный
б. Ложь
20. В NLP поддерживается контекстное моделирование, с помощью которого одно из следующих вложений слов
а. Word2Vec
б. GloVe
+в. BERT
г. Все вышеперечисленное

27.2. Лингвистический Модуль
1. Контрольная работа

1) Какому термину соответствует определение: процесс разделения текста на составляющие.

2) Какому термину соответствует определение: процесс приведения словоформы к лемме — её нормальной (словарной) форме.

3) Какому термину соответствует определение: процесс нахождения основы слова для заданного исходного слова.

4) Приведите формулу для расчета TF-IDF представлений.

5) Напишите программу для получения векторного представления “мешок слов” по заданному тексту.

6) Перечислите типы классификационной задачи?

7) Какие метрики качества используются для оценки классификационных моделей?

8) На каком предположении строятся векторные представления слов word2vec?

9) Можно ли получить векторное представление токена не присутствовавшего в обучающей выборке для модели word2vec?

10) В чем заключается отличие CBOW и Skip-Gram модели?

2. Тест

1. Что такое корпус?
a) Коллекция текстов, хранящихся на компьютере.
b) Электронная база данных, похожая на словарь.
c) Любая большая коллекция слов, такая как коллекция книг, газет и журналов.

2. Почему исследователи-лингвисты используют корпусы?
a) Потому что другие методы анализа языка не являются надежными.
b) Потому что компьютеры могут подтвердить наши интуитивные представления о языке.
c) Потому что компьютеры могут помочь нам обнаружить интересные закономерности в языке, которые было бы трудно обнаружить в ином случае.
d) Потому что с помощью корпусов можно ответить на все вопросы о языке.

3. Что такое аннотации к корпусу?
a) Дополнительная информация к тексту, добавляемая с целью обеспечения более сложных поисков.
b) Разделение текста на предложения.
c) Добавление критических замечаний к тексту.

4 Что такое «специализированный корпус»?
a) Корпус, который используется для исторических исследований в области языка.
b) Корпус, который состоит из текстов разных жанров.
c) Корпус, который используется узкими специалистами в области языкознания.
d) Корпус, который фокусируется на узкой области, например, текстах одного жанра, одного периода, одного места создания.

5. Что из этого НЕ является типом корпуса?
a) Многоязычный корпус.
b) Учебный корпус.
c) Диахронический корпус.
d) Обзорный корпус.

6. Что такое BNC

a) Большой корпус британского английского языка
b) Корпус разных жанров письменного английского языка
c) Большой корпус британского разговорного языка
d) Специализируется корпус, представляющий язык английских газет

7. Какое из этих утверждений не является правдой о мониторинговом корпусе?
a) Он часто обновляется
b) Bank of English является примером мониторингового корпуса.
c) BNC является примером мониторингового корпуса.
d) Он используется для контроля быстрых изменений языка.

8. Что такое конкорданс?
a) Информация о частотах слов, нормированных на миллион сов.
b) Перечень примеров слова, представленных в корпусе, с некоторым контекстом справа и некоторым контекстом слева.
c) Алфавитный список слов, которые появляются в тексте.
d) Список слов и их частот, которые могут быть использованы для идентификации важных слов в тексте.

9. Что такое колоквиализм?
a) Тенденция ораторов говорить одновременно.
b) Тенденция слов взаимодействовать друг сдругом.
c) Тенденция слов появляться каждый раз вуникальных, различных контекстах.
d) Тенденция предложений порождать смысл.

10. Что такое распределение частот в корпус?
a) Информация о том, как часто встречается слово в корпусе.
b) Информация об использовании термина в целом ряде разлчных текстов, секциях, корпусах и т.п.
c) Информация о том, какова частотность слова на миллион слов.
d) Социолингвистическая информация о поле говорящих, которые представлены в корпус.

Ответы:
1 a
2 c
3 a
4 d
5 b
6 a
7 c
8 b
9 a
10 c

28. Итоговая аттестация. Перечень примерных заданий

Итоговая аттестация представляет собой выполнение и демонстрацию и защиту технологического проекта со следующими требованиями:
Обучающиеся должны показать навыки разработки документации к программным системам и стандартам в области лингвистики. Как пример можно написать Техническое задание для разработки программы обработки текста на языке Python.

Техническое задание (ТЗ) должно быть составлено так, чтобы программист, получивший этот документ, разработал именно то, что заказчик хотел получить. Как инструмент коммуникации в связке общения заказчик-исполнитель, техническое задание позволяет: – заказчику осознать, что именно ему нужно и требовать от исполнителя соответствия продукта всем условиям, оговоренным в ТЗ; – исполнителю понять суть задачи, показать заказчику «технический облик» программного изделия или автоматизированной системы; спланировать выполнение проекта и работать по намеченному плану; отказаться от выполнения работ, не указанных в ТЗ.

Далее проектные группы пишут пробную версию программы на языке Python, которая может обрабатывать текст, задавать параметр поиск определенных частей речи и выводить данные в таблицу Excel.

Данная программа поможет проводить анализ текста и может быть полезна как прототип цифровой технологии для оценки результатов обучения.

XII. Материально-техническое и учебно-методическое обеспечение Программы

Электронная информационно-образовательная среда КФУ (ЭИОС) представляет собой совокупность электронных информационных ресурсов, электронных образовательных ресурсов, информационных технологий, телекоммуникационных технологий, соответствующих технологических средств, обеспечивающих освоение обучающимися образовательных программ или их частей, а также взаимодействие между всеми участниками образовательного процесса независимо от места их нахождения;

ЭИОС обеспечивает:

—доступ к учебным планам, рабочим программам дисциплин (модулей), практик, к изданиям электронных библиотечных систем и электронным образовательным ресурсам;

—фиксацию хода образовательного процесса, результатов промежуточной аттестации и результатов освоения основной образовательной программы;

—проведение всех видов занятий, процедур оценки результатов обучения, реализация которых предусмотрена с применением электронного обучения, дистанционных образовательных технологий; https://edu.kpfu.ru/

—формирование электронного портфолио обучающегося, в том числе сохранение работ обучающегося, рецензий и оценок на эти работы со стороны любых участников образовательного процесса;

—взаимодействие между участниками образовательного процесса, в том числе синхронное и (или) асинхронное взаимодействие посредством сети «Интернет».

Система «Антиплагиат.ВУЗ» и другие ресурсы позволяющие обеспечивать освоение обучающимися образовательных программ в полном объеме независимо от места нахождения обучающихся.

Основными элементами ЭИОС КФУ являются:

а) электронные информационные ресурсы:

— официальный сайт КФУ (https://kpfu.ru/);

— личные кабинеты участников образовательного процесса, обеспечивающие доступ к

— компонентам ЭИОС КФУ;

— корпоративная электронная почта;

— сайт Научной библиотеки им. Н.И. Лобачевского;

— информационно-аналитическая система управления образовательным процессом;

— система автоматического поиска текстовых заимствований;

— другие базы данных и файловые системы, используемые в образовательном процессе;

б) электронные образовательные ресурсы:

— система управления обучением Moodle;

— сайт дистанционного обучения (https://edu.kpfu.ru/), содержащий более 3500 цифровых образовательных ресурсов;

— площадка для создания и тестирования курсов (https://do.kpfu.ru);

в) электронные библиотечные системы:

— внутренняя электронная библиотечная система КФУ, обеспечивающая доступ к информационным ресурсам, включающая печатные и электронные документы на русском и иностранных языках;

— внешние электронные библиотечные системы и электронные библиотеки, доступ к которым осуществляется на договорной основе;

г) средства вычислительной техники:

— серверное оборудование КФУ;

— компьютеры, эксплуатируемые в КФУ;

— ноутбуки, планшеты, смартфоны и другие портативные, мобильные персональные компьютеры;

— средства организационной и множительной техники;

— мультимедийное оборудование.

Система электронного (дистанционного) обучения (далее – СДО) – электронная информационно-образовательная среда в виде системно-организованной совокупности информационно-коммуникационных средств и технологий, процессов программно-аппаратного и организационно-методического обеспечения, деятельности научно-педагогического, педагогического, учебно-вспомогательного и инженерного персонала (работников), ориентированная на реализацию системы сопровождения учебного процесса с целью удовлетворения образовательных потребностей

обучающихся независимо от места их нахождения

Доступ в СДО обеспечивается непрерывно (в круглосуточном режиме с коэффициентом доступности всех компонентов среды не ниже 99,5 %) и из любой точки подключения к сети Интернет с заданными характеристиками канала связи.

Доступ ко всем сервисам СДО является персонализированным (под единой учетной записью).

Освоение ДПП ПП предполагает использование следующего программного обеспечения и информационно-справочных систем:

— Язык Python версии 3, среда выполнения языка
— Блокноты разработки – Notepad++, Sublime

XIII. Список литературы

Васильев Юлий Обработка естественного языка. Python и spaCy на практике. – СПб.: Питер, 2021. – 256 с.: ил. – (Серия «Библиотека программиста»).

Златопольский Д.М. Основы программирования на языке Python.
// М.: ДМК Пресс, 2017. – 284 с.

Мартишин С.А. Проектирование и реализация баз данных в СУБД MySQL с использованием MySQL Workbench: Методы и средства проектирования информационных систем и технолог // М.: Форум, 2017. – 62 с.

Стивенс Род. Алгоритмы и практическое применение Москва Издательство «Э» 2016 год 544с.

К. Дж. Дейт Введение в системы баз данных 2005 1328 с. (8-ое издание)

Озкарахан Э. Машины баз данных Издательство Мир, год выпуска 1989 ISBN 5-03-000482-3695 с.

Т.И. Сергеева М.Ю. Сергеев Распределенная обработка информации ФГБОУ ВПО «Воронежский государственный университет». – 2014. – 96 с.

Зубов А.В. Информационные технологии в лингвистике [Текст] : учеб. пособие для вузов / А.В. Зубов, И.И. Зубова. М.: Академия, 2004. 208 с. (Высшее профессиональное образование). Рек. УМО. В пер. Библиогр.: с. 192

Баранов А.Н. Введение в прикладную лингвистику [Текст]: учебное пособие /

А.Н. Баранов; МГУ им. М.В. Ломоносова.

Богданова С.Ю. Квантитативная лингвистика и НИТ: учебное пособие. Иркутск: 'Аспринт', 2017. 87 с. (28 экз.)

Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов направления Лингвистика. 2-е изд., перераб. и дополн. СПб.: СПбГУ, Филологический факультет, 2013. 148 с. (4 экз.).

Гребенщикова А.В. Основы квантитативной лингвистики и новых информационных технологий: учеб. пособие. Издание: 2-е изд., стер. Москва: ФЛИНТА : Наука, 2015. 152 с.

Щипицина, Л. Ю. Информационные технологии в лингвистике [Электронный ресурс] : учеб. пособие / Л. Ю. Щипицина. - М. : ФЛИНТА, 2013. - 128 с. - ISBN 978-5-9765-1431-7

Сулейманов Д.Ш., Хадиев Р.М., Якушев Р.С. Компьютерные информационные технологии. - Казань: КГУ, 2004. - 191 с. (12 экз.).

Термины информатики и информационных технологий: Англо-татарско-русский толковый словарь. - Казань: Магариф, 2006. - 383 с. (2 экз.).

Гладкий А.В., Мельчук И.А. Элементы математической лингвистики. - М.: Наука, 1969. - 192 с. (3 экз.).

Чернявская, В. Е. Лингвистика текста. Лингвистика дискурса : учеб. пособие / В. Е. Чернявская. М. : Флинта :
Наука, 2013. - 208 с.

Буянова, Л. Ю. Терминологическая деривация в языке науки: когнитивность, семиотичность, функциональность [Электронный ресурс]: монография / Л. Ю. Буянова. – 2-е изд., стереотип. – М.: Флинта, 2011. – 389 с.

Математические методы в приложениях. Математическое программирование. Тензорная алгебра, Журбенко, Лариса Никитична; Зайцева, О. Н.;Нуриев, А. Н., 2011г.

Математические методы в историко-экономических и историко-культурных исследованиях, Ковальченко, И. Д., 2008г.

Горобец Е.А. Основы веб-технологий: учебно-методическое пособие для студентов-филологов / сост. Е.А.Горобец. – Казань, 2011. – 76 с.

Балашова С.А. Математика и информатика: Учебное пособие. – М.: РУДН, 2009. – 193 с.