Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Государственный институт русского языка им. А.С. Пушкина»
Центр дополнительного образования

«УТВЕРЖДАЮ»
Ректор ФГБОУ ВО «Гос. ИРЯ
им. А.С. Пушкина»
_____ Н.С. Трухановская
«29» августа 2023 г.
(печать)

ДОПОЛНИТЕЛЬНАЯ ОБЩЕОБРАЗОВАТЕЛЬНАЯ ОБЩЕРАЗВИВАЮЩАЯ ПРОГРАММА

«Цифровая лингвистика для школьников»
(ознакомительный уровень)

Направленность программы: техническая
Возраст обучающихся: 12–18 лет
Срок реализации программы: 32 часа

Москва
2023

Разработчики:

ФГБОУ ВО «Гос. ИРЯ им. А.С. Пушкина», к.филол.н., заведующая лабораторией когнитивных и лингвистических исследований М.Ю. Лебедева;

ФГБОУ ВО «Гос. ИРЯ им. А.С. Пушкина», к.филол.н., научный сотрудник лаборатории когнитивных и лингвистических исследований А.Н. Лапошина.

Протокол заседания Ученого Совета Гос. ИРЯ им. А.С. Пушкина №_35_ от «_29_» _августа_ 20_23_ г.

Дополнительная общеразвивающая программа составлена в соответствии с действующими законодательными и нормативными правовыми актами Российской Федерации и города Москвы, локальными нормативными актами Гос. ИРЯ им. А.С. Пушкина.

Ученый секретарь _____ Т.Ю. Горелова

Врио директора центра дополнительного образования _____ М.А. Игнатьева

И.о. директора департамента образовательной деятельности _____ И.А. Замилова

1 ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Направленность программы

Дополнительная общеразвивающая программа «**Цифровая лингвистика для школьников**» имеет техническую направленность.

Уровень программы вводный.

Актуальность программы. Цифровые технологии и анализ больших текстовых данных стали неотъемлемой частью современного бизнеса и науки. Курс "Введение в цифровую лингвистику" помогает учащимся понять, как современные технологии меняют способы хранения и обработки текстовой информации, усвоить принципы применения компьютерных алгоритмов лингвистической обработки текста в различных областях экономики, которые лежат в основе высокотехнологичных продуктов: чат-боты, голосовые помощники, автоматические переводчики, базы данных, поисковые системы, корпуса текстов и т.д., а также получить начальные навыки работы с программными инструментами обработки текстовых данных. Кроме того, обзор профессиональных перспектив в этой области может быть полезным для выбора будущей специальности.

Цель программы. Целью изучения программы «Цифровая лингвистика для школьников» является формирование у обучающихся базовых знаний в области цифровой лингвистики и эффективного использования цифровых ресурсов с применением технологий автоматической обработки текста.

Задачи программы

Обучающие:

- обучение истории и теоретическим основам цифровой лингвистики;
- обучение использованию необходимого набора компьютерных программ для автоматической обработки текстов на естественном языке обработки;
- обучение выбору набора данных в зависимости от практической задачи;
- обучение работе с корпусными инструментами для решения учебных задач;
- обучение созданию и анализу собственной коллекции текстов.

Развивающие:

- развитие технических способностей обучающихся в ходе проектирования и использования инструментов автоматической обработки текстов.
- развитие у обучающихся аналитического мышления и способности выявлять закономерности и особенности в языке и текстах;

Воспитательные:

-воспитание ответственности, трудолюбия, целеустремленности и организованности.

- воспитание коммуникативных компетенций; умения работать в команде, владение техниками мозгового штурма и креативной разработки проблемы.

Учащиеся, для которых программы актуальна

Возраст обучающихся по данной программе: 14–18 лет. Учащиеся должны иметь элементарные навыки работы с компьютером и информацией, иметь стабильный доступ в интернет.

Формы и режим занятий

Форма обучения – очная (с применением полностью или частично ЭО и ДОТ), групповая.

Количество обучающихся в группе: 12-16 человек.

Занятия проходят 1 раз в неделю по 4 ак. часа. Предусмотрен перерыв продолжительностью 10 минут в конце каждого учебного часа.

Срок реализации программы

Общее количество учебных часов, запланированных на весь период обучения, – 32 ак. часа.

Планируемые результаты

По итогам обучения по программе обучающиеся будут

знать:

– основные понятия цифровой лингвистики;
– основные библиотеки для автоматической обработки текстовых данных;
– основы работы систем автоматической классификации текстов по заданным параметрам;
– принцип работы больших языковых моделей;
– основы работы с лингвистическими корпусами;

уметь:

– анализировать результаты автоматического морфологического анализа текста;
– анализировать результаты автоматического синтаксического анализа текста;
– создавать эффективные запросы к большим языковым нейросетевым моделям;
– осуществлять поиск в Национальном корпусе русского языка по лексическим, грамматическим и семантическим категориям;
– использовать инструмент анализа текстов AntConc или его аналоги.

Личностные результаты

– развитие целеустремленности, ответственности и трудолюбия, воспитание организованности и аккуратности в рабочем процессе;
–развитие навыков коммуникации в коллективе, умения формулировать свои мысли и ставить задачи в рабочем процессе.

Метапредметные результаты

- развитие навыков организации рабочего процесса;
- развитие способности критиковать собственную работу;

2 ФОРМЫ АТТЕСТАЦИИ И ОЦЕНОЧНЫЕ МАТЕРИАЛЫ

Формы контроля

Реализация программы «**Цифровая лингвистика для школьников**» предусматривает входную диагностику, текущий контроль, промежуточную и итоговую аттестацию обучающихся.

Входная диагностика проводится в форме опроса.

Текущий контроль включает следующие формы: тест, практическое задание.

Промежуточная аттестация не предусмотрена.

Итоговая аттестация проводится в форме практической самостоятельной работы по техническому заданию.

Основным механизмом выявления результатов обучения является формирующее оценивание.

Публичная презентация образовательных результатов программы осуществляется в форме презентации самостоятельной практической работы, выполненной в соответствии с полученным техническим заданием.

Обучающимся, успешно освоившим программу и прошедшим аттестацию в форме, предусмотренной программой, выдается документ, установленного образца, подтверждающий освоение программы.

Средства контроля

Контроль освоения обучающимися программы осуществляется путем оценивания степени соответствия самостоятельной практической работы техническому заданию.

Результативность обучения дифференцируется по трем уровням: низкий, средний, высокий.

При низком уровне освоения программы обучающимся самостоятельная практическая работа на 59-77% соответствует техническому заданию;

При среднем уровне освоения программы обучающимся самостоятельная практическая работа на 78-89% соответствует техническому заданию;

При высоком уровне освоения программы обучающийся самостоятельная практическая работа на 90-100% соответствует техническому заданию.

Позиции педагогического наблюдения:

- уровень овладения учебным материалом;
- коммуникативные компетенции и овладение навыком работы в коллективе.

3 СОДЕРЖАНИЕ ПРОГРАММЫ

Учебно-тематический план

№ п/п	Название раздела, темы	Количество часов			Формы аттестации (контроля) по разделам
		Всего	Теорети-ческих	Практи-ческих	
1.	**Введение**	**4**	**4**	-	опрос
1.1.	Основные цели и задачи программы. Принципы практической работы на курсе (сервис Colab)	1	1	-	
1.2.	Основные понятия цифровой лингвистики	1	1	-	
1.3.	Практические задачи цифровой лингвистики	2	2	-	
2.	**Основы автоматической обработки текста**	**8**	**5**	**3**	
2.1	Общие понятия и краткая история автоматической обработки текстов	1	1	-	опрос
2.2	Уровни автоматического анализа языка	1	1	-	опрос
2.3.	Автоматический морфологический анализ текста	2	1	1	Выполнение практических заданий
2.4.	Автоматический синтаксический анализ текста	2	1	1	Выполнение практических заданий
2.5.	Дополнительные возможности библиотек автоматической обработки текстов	2	1	1	Выполнение практических заданий
3.	**Практика автоматической классификации текстов**	**4**	**1**	**3**	
3.1	Автоматическая классификация текстов.	1	1	-	опрос

3.2	Практика классификации текстов по эмоциональной окраске (тональности)	3	-	3	Выполнение практических заданий
4.	**Большие языковые модели**	**4**	**2**	**2**	
4.1.	Принципы работы больших генеративных текстовых моделей	2	2	-	
4.2.	Применение и оценки качества результатов работы больших генеративных моделей	2	-	2	Выполнение практических заданий
5.	**Корпус как особый вид языковых данных**	**2**	**2**	**2**	
5.1.	Введение в корпусную лингвистику.	1	1	-	
5.2.	Эффективный поиск в корпусе	3	1	2	Выполнение практических заданий
6.	**Создание и анализ собственного корпуса**	**4**	**3**	**1**	
6.1.	Методы проектирования собственного корпуса текстов	**2**	2	0	опрос
6.2.	Анализ собственной коллекции текстов с помощью корпусного менеджера	**2**	1	1	Выполнение практических заданий
7.	**Итоговое занятие**	**4**	**-**	**4**	**Практическая самостоятельная работа**
	Итого	**32**	**17**	**15**	

Содержание учебно-тематического плана

1. Раздел «Введение»

1.1 Тема

Теоретическая часть

Основные цели и задачи программы. Знакомство со слушателями. Обзор главных целей и задач курса. Техника практической работы на курсе (обзор сервиса Colab).

1.2. Тема

Основные понятия цифровой лингвистики. Основные понятия цифровой и компьютерной лингвистики, краткая история научной области. Навыки цифрового лингвиста, его профессия.

1.3. Тема

Практические задачи цифровой лингвистики. Разнообразие областей практического применения методов цифровой лингвистики (на материале конкретных примеров): языкознание, литературоведение, обучение языкам, судебная лингвистика. Классификация текстов по заданным типам, извлечение именованных сущностей из текстов документов.

2. Раздел «Основы автоматической обработки текста»

2.1. Тема

Общие понятия и краткая история автоматической обработки текстов. Автоматическая обработка естественного языка (NLP). Краткая история развития автоматической обработки текстов. Машинный перевод как первое практическое приложение систем АОТ.

2.2. Тема

Уровни автоматического анализа языка. Фонетический уровень: распознавание и генерация звучащей речи. Лексический уровень: токенизация текста, токен; лемматизация текста, лемма. Синтаксический уровень: построение синтаксического дерева.

2.3.Тема

Автоматический морфологический анализ текста.

Теоретическая часть.

Принцип работы автоматической морфологической разметки текста. Морфологический тэг. Частеречная разметка текста. Обзор нескольких готовых решений по морфологическому разбору. Проблема омонимии и полисемии слов, возможности их решения.

Практическая часть.

Практика морфологической разметки текста и понимания морфологических тэгов на основе морфологического парсера pymorphy2.

2.4. Тема

Автоматический синтаксический анализ текста. Теоретическая часть.

Понятие синтаксического дерева разбора. Вершина дерева и зависимые. Глубина дерева.

Практическая часть.

Практика создания синтаксического дерева предложения. Сравнение результата со школьной системой синтаксического разбора.

2.5. Тема

Дополнительные возможности библиотек автоматической обработки текстов.

Теоретическая часть.

Предобработка текстов. Анализ биграмм. Поиск коллокаций. Поиск именованных сущностей в тексте.

Практическая часть.

Работа с библиотеками NLTK и Spacy.

3. Раздел «Практика автоматической классификации текстов»

3.1 Тема

Теоретическая часть. Автоматическая классификация текстов. Формулировка практической проблемы определения эмоциональной окраски (тональности) текста, описание задачи. Обзор возможных эталонных, обучающих коллекций для этой задачи.

3.2 Тема

Практическая часть. Практика классификации текстов по эмоциональной окраске. Проверка обучающей коллекции текстов, предобработка текстов. Выбор метода классификации текстов. Метрики качества полученной модели. Анализ результатов работы модели, пути оптимизации результата. Ограничения работы модели.

4. Раздел «Большие языковые модели: принципы работы, возможности и ограничения»

4.1. Тема

Теоретическая часть.

Принципы работы больших генеративных текстовых моделей. Понятие языковой модели. Краткая история развития языковых моделей. Обучающие данные для современных языковых моделей. Роль лингвистики в создании современных языковых моделей.

4.2. Тема

Практическая часть.

Применение и оценки качества результатов работы больших генеративных моделей. Обзор и характеристики основных доступных чатов на основе больших языковых моделей. Понятие промпта. Практика написания и

редактуры промптов для решения учебных задач. Эффект галлюцинации нейросети, практика проверки фактов, предлагаемых нейросетью.

5. Раздел «Корпус как особый вид языковых данных»

5.1. Тема

Теоретическая часть.

Введение в корпусную лингвистику. Виды больших текстовых данных: датасет, корпус, онтология, словарь. Краткая история появления электронных корпусов текстов. Типы корпусов, методы поиска и практические приложения. Национальный корпус русского языка.

5.2. Тема

Эффективный поиск в корпусе.

Теоретическая часть.

Виды поиска в корпусе. Лексико-грамматический поиск, поиск по семантическим категориям. Анализ конкордансов. Виды подкорпусов. Инструмент создания графиков частотности слова. Язык запросов CQL.

Практическая часть.

Практика составления запросов к корпусу и анализу результатов на материале НКРЯ

6. Раздел «Создание и анализ собственного корпуса»

6.1. Тема

Методы проектирования собственного корпуса текстов. Поиск подходящих данных. Сбор и хранение больших текстовых данных. Разметка текстовых данных: метатекстовая и текстовая разметка. Краудсорсинговая разметка данных.

6.2. Тема

Анализ собственной коллекции текстов с помощью корпусного менеджера.

Теоретическая часть.

Понятие корпусного менеджера. Обзор возможностей корпусного менеджера AntConc. Поиск в собственном корпусе. Особенности построения конкордансов и настройки выдачи результатов. Методика составления и интерпретации частотных списков слов. Ограничения работы программы.

Практическая часть.

Практика составления запросов и анализа результатов в корпусном менеджере AntConc.

7 Итоговое занятие

Практическая часть. Практическая самостоятельная работа под руководством преподавателя по созданию и анализу собственной коллекции текстов на тему, выбранную учащимся. Представление результатов самостоятельной практической работы.

4 ОРГАНИЗАЦИОННО-ПЕДАГОГИЧЕСКИЕ УСЛОВИЯ РЕАЛИЗАЦИИ ПРОГРАММЫ

Учебно-методические условия реализации программы

Реализация программы «**Цифровая лингвистика для школьников**» предполагает следующие формы организации образовательной деятельности: лекция, практическое занятие.

Программа реализуется с применением электронного обучения и дистанционных образовательных технологий с использованием систем дистанционного обучения.

При реализации программы используются следующие образовательные технологии: личностно-ориентированное обучение, технология сотрудничества, технологии развивающего обучения.

Также могут быть использованы дистанционные образовательные технологии.

При реализации программы используются следующие методы обучения: словесные, наглядные, практические.

Воспитывающий компонент программы

Воспитание является важным аспектом в образовании, который логично встраивается в содержание учебной программы.

На первом занятии учащиеся знакомятся с историей и традициями Института Пушкина.

В процессе обучения приоритет отдается воспитанию бережного отношения к материалам и оборудованию, которые используются на занятиях. Педагоги уделяют особое внимание воспитанию культуры общения в детско-взрослом коллективе.

Оценка результатов воспитательной работы происходит в процессе педагогического наблюдения на протяжении всего периода обучения.

Содержание воспитания:

- традиции и уникальность Института Пушкина, коллектива;
- адекватность восприятия профессиональной оценки.

Перечень методического обеспечения к программе

№ п/п	Название раздела (темы) учебно-тематического плана	Название и форма методического материала
1.	**Введение**	слайды презентации, список доп. литературы
2.	**Основы автоматической обработки текста**	слайды презентации, исполняемый файл в программным кодом формата ipynb, список доп. литературы
3.	**Практика автоматической классификации текстов.**	слайды презентации, исполняемый файл в программным кодом формата ipynb, список доп. литературы
4.	**Большие языковые модели: принципы работы, возможности и ограничения.**	слайды презентации, список доп. литературы
5.	**Корпус как особый вид языковых данных.**	слайды презентации, сайт НКРЯ (электронный ресурс), список доп. литературы
6.	**Создание и анализ собственного корпуса.**	слайды презентации, сайт laurenceanthony.net (электронный ресурс), список доп. литературы

Для проведения занятий с применением электронного обучения и дистанционных образовательных технологий с использованием систем дистанционного обучения по каждой учебной теме разработаны информационные материалы и технологические карты (инструкции, памятки) по выполнению обучающимися практических заданий.

Материально-технические условия реализации программы

Требования к помещению для занятий:
Компьютерный класс с выходом в сеть Интернет.
Требования к мебели:
Удобные парты, стулья, экран проектора, школьная доска.
Оборудование:
1. компьютер
2. мультимедийный проектор
3. веб-камера,
4. микрофон,
5. специализированное ПО для видео связи
6. Доступ к сервису Google Colab
7. Предустановленная программа AntConc (бесплатная программа)

Учебно-информационное обеспечение программы

Нормативно-правовые акты и документы

1. Федеральный закон от 29 декабря 2012 г. № 273-ФЗ "Об образовании в Российской Федерации" (с изм. на 24 июня 2023 года).

2. Концепция развития дополнительного образования детей до 2030 года (с изм. на 15.05.2023 г.) (утверждена распоряжением Правительства Российской Федерации от 31 марта 2022 г. № 678-р).

3. Порядок организации и осуществления образовательной деятельности по дополнительным общеобразовательным программам (утвержден приказом Министерства просвещения Российской Федерации от 27 июля 2022 г. № 629).

4. Целевая модель развития региональных систем дополнительного образования детей (утверждена приказом Министерства просвещения Российской Федерации от 3 сентября 2019 г. № 467) (с изм. на 21.04.2023).

5. Методические рекомендации по проектированию дополнительных общеразвивающих программ (включая разноуровневые программы): приложение к письму Министерства образования и науки Российской Федерации от 18 ноября 2015 г. № 09-3242.

6. Методические рекомендации по реализации дополнительных общеобразовательных программ с применением электронного обучения и дистанционных образовательных технологий: приложение к письму Министерства просвещения Российской Федерации от 31 января 2022 г. № ДГ-245/06.

7. СП 2.4.3648-20 «Санитарно-эпидемиологические требования к организации воспитания и обучения, отдыха и оздоровления детей и молодежи» (утверждены постановлением Главного государственного санитарного врача Российской Федерации от 28 сентября 2020 г. № 28).

8. СанПиН 1.2.3685-21 «Санитарные нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» (утверждены постановлением Главного государственного санитарного врача Российской Федерации от 28 января 2021 г. № 2).

9. Приказ Департамента образования города Москвы от 17.12.2014 г. № 922 «О мерах по развитию дополнительного образования детей» (с изм. на 24.10.2022).

10. Приказ Департамента образования и науки города Москвы от 3.04.2023 г. № 271 «О внесении изменений в приказ Департамента образования и науки города Москвы от 17 декабря 2014 года № 922». - *для программ вводного уровня*

Литература:

1. Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова

Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. — М.: Изд-во НИУ ВШЭ, 2017. — 269 с.
2. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. — М.: ЛЕНАНД, 2016
3. Зализняк А.А. Грамматический словарь русского языка. — М., Русский язык, 1980.
4. Международные стандарты в области корпусной лингвистики. // Структурная и прикладная лингвистика. Выпуск 9. СПб., 2012 С. 201-221.
5. Ляшевская О.Н. и др. 2010. Оценка методов автоматического анализа текста: морфологические парсеры русского языка // Компьютерная лингвистика и интеллект. технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 9 (16) — М.: Изд-во РГГУ.
6. Прикладная и компьютерная лингвистика / Под. ред. Николаева И.С., Митрениной О.В., Ландо Т.М. — М.: ЛЕНАНД, 2016. — 320 с

Интернет-ресурсы:

1. http://www.pushkin.institute // Сайт Гос ИРЯ им. А.С. Пушкина
2. http://pushkininstitute.ru // Образовательный портал «Образование на русском).

Кадровое обеспечение программы

Программа «Цифровая лингвистика для школьников» реализуется квалифицированными научно-педагогическими кадрами системы высшего профессионального образования, имеющим профессиональное образование в области, соответствующей профилю программы, и постоянно повышающим уровень профессионального мастерства. Для обеспечения образовательного процесса необходимо привлечение следующих специалистов: преподаватели, имеющие специальность компьютерного лингвиста.