|  |  |
| --- | --- |
| Национальный  исследовательский университет  «Высшая школа экономики»  **Лицей** | **Приложение 153**  УТВЕРЖДЕНО  педагогическим советом  Лицея НИУ ВШЭ  протокол от 04.12.2017 № 1 |

Рабочая программа по учебному предмету (курсу)

«Лингвистика для математиков»

**10 класс**

**Авторы:**

Скоринкин Даниил Андреевич

Преподаватель школы лингвистики НИУ ВШЭ

**1. Содержание учебного предмета**

Курс знакомит учащегося с историей и современным состоянием компьютерной лингвистики как научной дисциплины и автоматической обработки языка как прикладной сферы деятельности и высокотехнологичной отрасли экономики. В рамках курса учащиеся получат представление об актуальных задачах автоматической обработки языка, таких как информационный поиск, машинный перевод, извлечение информации (знаний) из текста, анализ тональности, разрешение анафоры, автоматическое порождение языка, создание диалоговых агентов, а также о современных подходах к их решению. Часть тем предполагает практикумы, во время которых учащиеся получат опыт самостоятельного решения (под руководством преподавателя) таких задач.

Отдельный блок внутри курса посвящен лингвистическим корпусам и корпусным исследованиям. Учащиеся получат навык самостоятельно производить исследования в Национальном корпусе русского языка, освоят интерфейс поискового запроса НКРЯ, смогут осмысленно использовать грамматические, семантические и другие признаки, доступные в корпусе, получат представление о составе и подкорпусах НКРЯ. Кроме того, будет изучен потенциал использования Интернета и поисковых систем как корпуса для лингвистических и социолингвистических исследований.

Освоение программы курса предполагает и освоение некоторых элементов лингвистической теории, необходимых для понимания проблематики автоматической обработки языка. Учащиеся получат представление о предмете и проблематике таких разделов науки о языке, как морфология и синтаксис.

Особое внимание в курсе будет уделено практическому применению достижений компьютерной лингвистики. Учащиеся получат представление о том, как компьютерная лингвистика помогает исследователям из других областей и в каких массовых продуктах находят применение ее методы. Кроме того, в рамках курса предполагаются мастер-классы сотрудников коммерческих компаний, работающих в области автоматической обработки текстов.

**2. Личностные, метапредметные и предметные результаты освоения учебного предмета**

ФГОС выделяет три группы требований к результатам освоения среднего (полного) общего образования: личностные и метапредметные, единые для всего образовательного процесса, и предметные, специфические для отдельно взятого предмета.

**Личностные:**

1. Формирование ответственного отношения к учению, готовности и способности обучающихся к саморазвитию и самообразованию на основе мотивации к обучению и познанию, осознанному выбору и построению дальнейшей индивидуальной траектории образования на базе ориентировки в мире профессий и профессиональных предпочтений, с учетом устойчивых познавательных интересов, а также на основе формирования уважительного отношения к труду, развития опыта участия в социально значимом труде;
2. Формирование целостного мировоззрения, соответствующего современному уровню развития науки и общественной практики, учитывающего социальное, культурное, языковое, духовное многообразие современного мира;
3. Формирование коммуникативной компетентности в общении и сотрудничестве со сверстниками, детьми старшего и младшего возраста, взрослыми в процессе образовательной, общественно полезной, учебно-исследовательской, творческой и других видов деятельности;

**Метапредметные:**

1. Умение определять понятия, создавать обобщения, устанавливать аналогии, классифицировать, самостоятельно выбирать основания и критерии для классификации, устанавливать причинно-следственные связи, строить логическое рассуждение, умозаключение (индуктивное, дедуктивное и по аналогии) и делать выводы;
2. Умение создавать, применять и преобразовывать знаки и символы, модели и схемы для решения учебных и познавательных задач;
3. Формирование и развитие компетентности в области использования информационно-коммуникационных технологий (далее ИКТ – компетенции); развитие мотивации к овладению культурой активного пользования словарями и другими поисковыми системами;

**Предметные:**

***(Филология)***

*Русский язык*

1. расширение и систематизацию научных знаний о языке; осознание взаимосвязи его уровней и единиц; освоение базовых понятий лингвистики, основных единиц и грамматических категорий языка;
2. формирование навыков проведения различных видов анализа слова (фонетического, морфемного, словообразовательного, лексического, морфологического), синтаксического анализа словосочетания и предложения, а также многоаспектного анализа текста;
3. обогащение активного и потенциального словарного запаса, расширение объема используемых в речи грамматических средств для свободного выражения мыслей и чувств адекватно ситуации и стилю общения;

**3. Тематическое планирование с указанием количества часов,**

**отводимых на освоение каждой темы**

|  |  |  |  |
| --- | --- | --- | --- |
| № | Тема урока | Количество часов | Контрольные виды деятельности констатирующего типа |
| 1 | Компьютерная лингвистика как часть науки об искусственном интеллекте. Философские основания. 1. Способность к общению как (не)достаточный признак интеллекта. Тест Тьюринга. Эксперимент Джона Серля "китайская комната" как контраргумент для теста Тьюринга. 2. Роботы-собеседники из научной фантастики. Что нужно «знать» роботу, чтобы понимать человека и взаимодействовать с человеком. | 1 |  |
| 2 | История компьютерной лингвистики и автоматической обработки языка 1. 40-е – 50-е; послевоенное развитие computer science и теории алгоритмов; формальные лингвистические модели. Тьюринг, Шеннон, Хомский. Начало машинного перевода, Джорджтаунский эксперимент IBM.  2. Оптимизм 60-х. Диалоговый агент ELIZA. Программа SHRDLU. Brown Corpus.  3. AI winter 70-х и возрождение интереса в (конце) 80-х.MUC. Появление русскоязычных корпусов. Уппсальский корпус. Машинный фонд русского языка.  4. 90-е – веб, поисковики, большие корпуса. Начинают работать статистические подходы на машинном обучении. Бум в информационном поиске. 5. 2000-е – триумф машинного обучения. Параллельные корпуса и статистический MT.  6. Наши дни. Deep learning и Нейросети. | 1 | Задание на работу с кодировками |
| 3 | Современное состояние, последние разработки (state of the art) в сфере автоматической обработки языка. Перспективы.  1. Практическая сторона – поиск, переводчики, распознавание текста, извлечение знаний, анализ тональности в соцсетях. Демонстрация. Краткий рассказ про распознавание речи, информационный поиск, извлечение знаний. Слово о ближайшем будущем.  2. Научная сторона. Исследование текстов с разными целями (политология, социология, литературоведение, криминалистика), исследование самого языка. Демонстрация.  3. Практика: используем Ngram viewer. | 2 |  |
| 4 | Введение в статистические подходы к анализу естественного языка.  1. Поверхностное объяснение сути машинного обучения («обучение на примерах»)  2. Демонстрация применимости машинного обучения к задачам комп.лингвистики  3. Совместная работа по отбору релевантных признаков для задачи классификации текстов. | 2 | Сбор датасета для тренировки классификатора |
| 5 | Виды классификаторов. Деревья решений, Наивный байесовский классификатор, метод k-Nearest Neighbours | 1 | Самостоятельное построение (абстрактного) дерева решений. |
| 6 | Нейросети и глубинное обучение в задачах автоматической обработки языка | 1 |  |
| 7 | Лингвистические данные. Корпуса 1. Типы лингвистических данных. i) Полевые данные ii) Словари  iii) Грамматики iv) Корпуса (текстовые или мультимедийные) 2. Корпуса.  i) Чем корпус отличается от набора текстов. ii) Типы корпусов. iii) Известные корпуса. НКРЯ. | 1 |  |
| 8 | Национальный корпус русского языка. 1. Как устроен поиск в НКРЯ.  2. Обзор подкорпусов НКРЯ 3. Какого типа исследования можно проводить на базе НКРЯ (практическое занятие). | 1 |  |
| 9 | Национальный корпус русского язык (продолжение). 1. Проводим исследование изменения значений слов на базе НКРЯ.  2. Анализ проблем (морфологическая и лексическая омонимия, дисбалансы в корпусе) | 1 |  |
| 10 | Частотность. Частотности единиц текста и их важность для лингвистических исследований. 1. Что может сказать о тексте количество употреблений определенного слова. 2. Пробуем считать Tf/idf. Извлекаем ключевые слова из текста. | 2 | В начале урока - небольшая контрольная по теме "лингвистические корпуса" (предыд.тема) |
| 11 | Интернет как лингвистический корпус для исследования. 1. Что можно исследовать прямо в поисковике. Синтаксис и возможности расширенного поискового запроса.  2. Веб как спеллчекер.  3. Говорят ли так? Веб как помощник переводчика (помимо Google Translate) 4. Исследуем региональные вариации языка с помощью расширенного запроса. (в Украине vs на Украине; поребрик vs бордюр) 5. Веб как источник знаний о мире (методы майнинга структурированной информации) 6. ГИКРЯ и его возможности. | 2 |  |
| 12 | Морфологический анализ.  1. Морфология как раздел науки о языке. Предмет морфологии.  2. Грамматическое vs лексическое значение. Словоизменительная парадигма. Словарь Зализняка. 3. Стеммеры. Стеммер Портера. 4. Морфологические анализаторы. Mystem. 5. Морфологическая омонимия 6. Морфологическая разметка в НКРЯ и других корпусах. Практикуем поиск в НКРЯ с опорой на морфологию. | 2 |  |
| 13 | Конечные автоматы и регулярные выражения. | 2 |  |
| 14 | Синтаксический анализ.  1. Синтаксис как раздел науки о языке. Предмет синтаксиса.  2. Варианты иерархической связи между единицами речи. Линейный порядок. Древесная структура. 3. Синтаксические формализмы. Структура зависимостей. Структура составляющих.  4. Пробуем строить синтаксические деревья. Синтаксическая разметка в НКРЯ и других корпусах. Практикуем поиск в НКРЯ с использованием синтаксической информации. | 2 | Задание по построению синтаксических деревьев |
| 15 | Контекстно-свободные грамматики. | 2 |  |
| 16 | Блок "актуальные задачи компьютерной лингвистики". Задача извлечения информации из текста. История. DARPA и MUC. Простые инструменты извлечения информации с опорой на морфологию. Tomita парсер. GATE. Создаем собственную систему извлечения информации. | 4 | В начале первого урока - небольшая контрольная-опрос по базовым понятиям морфологии/синтаксиса (предыд.темы) |
| 17 | Мастер-класс сотрудника компании ABBYY о работе компьютерного лингвиста. Рассказ о технологиях извлечения информации, применяемых в индустрии. | 1 |  |
| 18 | Блок "актуальные задачи компьютерной лингвистики". Машинный перевод. История развития машинного перевода. Холодная война, Джорджтаунский эксперимент, ALPAC. Правиловый машинный перевод. Статистический машинный перевод, параллельные корпуса, выравнивание. Тестирование доступных систем MП и анализ ошибок. Создание примитивной системы машинного перевода. | 2 |  |
| 19 | Блок "актуальные задачи компьютерной лингвистики". Анализ тональности текста.  1. Как мы понимаем, что в тексте хвалят или ругают что-то или кого-то. Маркеры тональности. Инверторы и интенсификаторы. 2. Как это может понять компьютер.  3. Тестирование систем, анализ ошибок, делаем выводы о подводных камнях. | 2 |  |
| 20 | Блок "актуальные задачи компьютерной лингвистики". Автоматическое порождение языка (Natural language generation). Диалоговые агенты. Экспертные (вопросно-ответные) системы. Пробуем создать чат-бота и поговорить с ним. | 3 |  |
| 21 | Блок "актуальные задачи компьютерной лингвистики". Разрешение анафоры. Понятие референции, анафорическая связь, антецедент и анафор. Катафора. | 1 |  |
| 22 | Введение в статистические подходы в анализе естественного языка.  1. Поверхностное объяснение сути машинного обучения («обучение на примерах») 2. Демонстрация применимости машинного обучения к задачам комп.лингвистики 3. Совместная работа по отбору релевантных признаков для задачи классификации текстов. | 2 |  |
| 23 | Блок "актуальные задачи компьютерной лингвистики". Распознавание речи. Физические свойства звукового сигнала и их математическая формализация. Лингвистика в распознавании речи. | 1 |  |
| 24 | Лексикографические ресурсы и их применение в компьютерной лингвистике/автоматическом анализе текста.  1. Традиционные лексикографические ресурсы (словари). 2. Тезаурусы и лексические онтологии. Рутез, ворднет.  3. Меры семантической близости слов.  4. Считаем семантическую близость по ворднету. | 1 |  |
| 25 | Неоднозначность в языке. Разные уровни неоднозначности. Морфологическая, лексическая, синтаксическая. Способы моделирования значений и снятия неоднозначности. | 1 | Письменное задание в классе на приведение примеров неоднозначности разных типов. |
| 26 | Связь компьютерной лингвистики и языкового разнообразия: влияние особенностей языка (языковой семьи) на успешность решения определенных лингвистических задач. | 1 |  |
| 27 | Компьютерная лингвистика и педагогика. Применение методов автоматической обработки языка в образовании. | 1 |  |
| 28 | Компьютерная лингвистика и СМИ. Мониторинг медиа, измерение репутации, автоматический рерайт. | 1 |  |
| 29 | Forensic компьютерная лингвистика. Определение авторства текста. | 1 |  |
| 30 | Компьютерная лингвистика и литература. Автопоэты и автописатели. Цифровое литературоведение и "дальнее чтение". Компьютерная стилистика. Исследование: сравниваем язык Пушкина с языком Гоголя. | 2 |  |
|  |  | Итого часов: 44 |  |

**Учебно-методическое обеспечение образовательной деятельности:**

:

* Задание по работе с кодировками
* Сбор датасета для эксперимента по машинному обучению
* Создание дерева решений (абстрактного) по таблице с данными
* Контрольная по лингвистическим корпусам
* Применение учащимся одного из доступных открытых морфоанализаторов к собственному тексту

**3**. **Учебно-методическое обеспечение образовательной деятельности:**

*В. И. Беликов, Л. П. Крысин.* Социолингвистика. М.: Издательство РГГУ, 2001.

*С. А. Бурлак.* Происхождение языка: Факты, исследования, гипотезы. М.: Астрель, 2011.

*Н. Б. Вахтин, Е. В. Головко*. Социолингвистика и социология языка: Учебное пособие. СПб.: Издательство Европейского университета в Санкт-Петербурге, 2004.

*В. А. Плунгян*. Почему языки такие разные? М.: АСТ-ПРЕСС КНИГА, 2012.

*Г. С. Старостин, А. В. Дыбо, А. Ю. Милитарёв, И. И. Пейрос, Е. Я. Сатановский*. К истокам языкового разнообразия: Десять бесед о сравнительно-историческом языкознании. М.: Издательский дом «Дело» РАНХиГС, 2015.

*И. Фридрих*. История письма. М.: Наука, 1979.

2013.

*David Crystal*. English as a global language. 2nd edn. Cambridge: CUP, 2003.

*David Crystal*. Language and the Internet. Cambridge: CUP, 2001

*David Crystal*. Language death. Cambridge: CUP, 2000.