|  |  |
| --- | --- |
| **Национальный** **исследовательский университет** **«Высшая школа экономики»****Лицей** | **Приложение 601**УТВЕРЖДЕНОпедагогическим советом Лицея НИУ ВШЭпротокол № 10 от 26.04.2023 |

Рабочая программа по учебному предмету (курсу)

«Компьютерная лингвистика»

10-11 класс

Автор:

Нестеренко Л.В.

1. **Планируемые результаты освоения учебного предмета (курса)**

К **личностным результатам**, на становление которых оказывает влияние изучение курса «Компьютерная лингвистика», можно отнести:

* понимание основ поведения в сфере компьютерной лингвистики для осуществления осознанного выбора будущей специализации;
* принятие и реализация ценностей здорового и безопасного образа жизни, бережное, ответственное и компетентное отношение к собственному физическому и психологическому здоровью;
* готовность обучающихся к конструктивному участию в принятии решений, затрагивающих их права и интересы, в том числе в различных формах общественной самоорганизации, самоуправления, общественно значимой деятельности;
* нравственное сознание и поведение на основе усвоения общечеловеческих ценностей, толерантного сознания и поведения в поликультурном мире, готовности и способности вести диалог с другими людьми, достигать в нем взаимопонимания, находить общие цели и сотрудничать для их достижения;
* развитие компетенций сотрудничества со сверстниками, детьми младшего возраста, взрослыми в образовательной, общественно полезной, учебно-исследовательской, проектной и других видах деятельности.
* мировоззрение, соответствующее современному уровню развития науки, значимости науки, готовность к научно-техническому творчеству, владение достоверной информацией о передовых достижениях и открытиях мировой и отечественной науки, заинтересованность в научных знаниях об устройстве мира и общества;
* готовность и способность к образованию, в том числе самообразованию, на протяжении всей жизни; сознательное отношение к непрерывному образованию как условию успешной профессиональной и общественной деятельности;
* уважение ко всем формам собственности, готовность к защите своей собственности;
* осознанный выбор будущей профессии как путь и способ реализации собственных жизненных планов;
* готовность обучающихся к трудовой профессиональной деятельности как к возможности участия в решении личных, общественных, государственных, общенациональных проблем.

**Метапредметные результаты** освоения образовательной программы по курсу «Компьютерная лингвистика» отражают овладение универсальными учебными действиями — познавательными, коммуникативными, регулятивными.

*Универсальные познавательные действия*

* искать и находить обобщенные способы решения задач, в том числе, осуществлять развернутый информационный поиск и ставить на его основе новые (учебные и познавательные) задачи;
* критически оценивать и интерпретировать информацию с разных позиций, распознавать и фиксировать противоречия в информационных источниках;
* использовать различные модельно-схематические средства для представления существенных связей и отношений, а также противоречий, выявленных в информационных источниках;
* находить и приводить критические аргументы в отношении действий и суждений другого; спокойно и разумно относиться к критическим замечаниям в отношении собственного суждения, рассматривать их как ресурс собственного развития;
* выходить за рамки учебного предмета и осуществлять целенаправленный поиск возможностей для широкого переноса средств и способов действия.

*Универсальные коммуникативные действия*

* осуществлять деловую коммуникацию как со сверстниками, так и со взрослыми (как внутри образовательной организации, так и за ее пределами), подбирать партнеров для деловой коммуникации исходя из соображений результативности взаимодействия, а не личных симпатий;
* координировать и выполнять работу в условиях реального, виртуального и комбинированного взаимодействия;
* развернуто, логично и точно излагать свою точку зрения с использованием адекватных (устных и письменных) языковых средств.

*Универсальные регулятивные действия*

* самостоятельно определять цели, задавать параметры и критерии, по которым можно определить, что цель достигнута;
* оценивать возможные последствия достижения поставленной цели в деятельности, собственной жизни и жизни окружающих людей, основываясь на соображениях этики и морали;
* ставить и формулировать собственные задачи в образовательной деятельности и жизненных ситуациях;
* оценивать ресурсы, в том числе время и другие нематериальные ресурсы, необходимые для достижения поставленной цели;
* выбирать путь достижения цели, планировать решение поставленных задач, оптимизируя материальные и нематериальные затраты;
* организовывать эффективный поиск ресурсов, необходимых для достижения поставленной цели;
* сопоставлять полученный результат деятельности с поставленной заранее целью.

**Предметные** **результаты** освоения учебного предмета включают в себя:

* понимание смысла основных теоретических положений компьютерной лингвистики;
* знание основных понятий лингвистики, таких как: уровни языка, языковые единицы, языковые значения;
* понимание взаимосвязи языковых уровней с задачами и подзадачами компьютерной лингвистики;
* умение применять лингвистические знания для решения задач, связанных с автоматической обработкой текста;
* умение использовать различные ресурсы и инструменты для решения задач, связанных с автоматической обработкой текста;
* знание основных этапов в обработке естественного языка, понимание их назначения и осознание их вклада в процесс автоматической обрабтк текстов;
* знание особенностей различных задач компьютерной лингвистики таких как: морфологический анализ, синтаксический анализ, создание корпусов, извлечение фактов, сентимент анализ, распознавание речи, генерация текстов, распознавание анафорических связей;
* знание различных методов, используемых при решении задач компьютерной лингвистики
* умение тестировать технологии автоматической обработки текста

**2. Содержание учебного предмета**

**Тема 1. Введение в компьютерную лингвистику. Задачи в области обработки естественного языка**

Предмет компьютерной лингвистики. Текстовые данные как особый вид данных. Текстовый vs. Нетекстовый формат. Формальные характеристики текста, о чем они говорят и как их можно использовать.

Лингвистика. Уровни языка и разделы лингвистики. Задачи компьютерной лингвистики. Соотношение уровней языка и задач компьютерной лингвистики.

**Тема 2. Поиск в тексте. Использование регулярных выражений**

Работа с текстовыми файлами. Единицы в текстах, устроенные шаблонно.

Функция регулярных выражений. Возможности регулярных выражений. Символы, группы символов, диапазоны, модификаторы.

**Тема 3. Лингвистические ресурсы: корпуса**

Что такое корпус? Предназначение корпуса. Корпус как инструмент лингвиста. Корпус как материал для разработчика.

Особенности устройства корпуса. Морфология, синтаксис семантика. Виды разметки в корпусе. Условные обозначения в разметке. Проблема омонимии. Виды омонимии. Снятие омонимии.

Национальный корпус русского языка (далее НКРЯ). Состав НКРЯ. Возможности НКРЯ. Поиск в корпусе. Параметры поиска. Поиск точных форм. Лексико-грамматический поиск. Выдача в корпусе. Формат KWIC.

Корпус Leed's. Возможности корпуса. Поиск по регулярным выражениям. Поиск коллокаций.

Интерпретация результатов поиска в корпусе.

Использование размеченных корпусов для инженерных целей. Преимущества использования корпусов с разметкой, возможности, которые они дают.

**Тема 4. Ключевые слова. Частотный словарь. N-граммы**

Выделение ключевых слов в коллекции текстов. Построение частотного словаря. Выделение коллокаций. Работа с ПО Antconc. Подсчет tf.idf. N‑граммы: что это такое и где их используют.

**Тема 5.** **Морфология и автоматический морфологический анализ**

Введение в морфологию. Части речи. Грамматические значения. Морфологическая омонимия.

Автоматический морфологический анализ. POS-tagging, лемматизация. Дизамбигуация: снятие морфологической омонимии. Морфологические анализаторы (pymorphy, mystem, pos\_tag NLTK).

Тестирование морфологических анализаторов.

**Тема 6. Программирование для задач компьютерной лингвистики: Python**

Типы данных. Арифметические операции. Модуль math. Cписки, индексы и цикл for. Cтроки, операции со строками (поиск, замены, слайсы). Модуль регулярных выражений. Условия if, elif, else. Файловый ввод-вывод. Функции. Модуль NLTK. Pymorphy. WordNet и FrameNet. Модули numpy, scipy.

**Тема 7. Автоматическая обработка текстов (АОТ). Препроцессинг**

Препроцессинг как один из этапов построения систем. Необходимость процедуры предобработки сырых текстовых данных. Токенизация, сплиттинг.

**Тема 8. Генерация текстов. Чатботы. Введение. Подготовка системы. Планирование. Подготовка входных данных. Первичная реализация**

Генерация текстов и вопросно-ответные системы. Входные данные для систем генерации и вопросно-ответных систем. Архитектура систем генерации и вопросно-ответных систем. Знакомство с различными системами генерации и вопросно-ответными системами. Особенности построения чатботов. Параметры, которые следует учитывать при построении чатботов.

**Тема 9. Направления компьютерной лингвистики**

Морфологический анализ, синтаксический анализ, распознавание анафорических связей, создание корпусов, извлечение именованных сущностей и фактов, анализ тональности (sentiment analysis), распознавание речи, генерация текстов, классификация и кластеризация текстов.

**Тема 10. Лингвистические ресурсы: WordNet и FrameNet**

Семантика. Отношения синонимии, меронимии. Понятие семантического поля. Синсеты и фреймы как единицы в Wordnet и Framenet. Смысловая неоднозначность как одна из проблем в области АОТ, например, в машинном переводе. Использование Wordnet для снятия смысловой неоднозначности.

**Тема 11. Препроцессинг: продолжение**

Нормализация. Стоп-слова. Зачем необходимо избавляться от стоп-слов. Разное наполнение списка стоп-слов в зависимости от поставленной задачи. Стемминг. Задачи, которые требуют стемминга на этапе предобработки.

**Тема 12. Конечные автоматы. Использование их в морфологическом анализе**

Входной алфавит. Множество внутренних состояний. Начальное состояние. Множество заключительных состояний. Функция переходов. Инструмент FOMA для построения конечных автоматов. Конечные автоматы для морфологического анализа. Решение лингвистических задач.

**Тема 13. Программирование для задач компьютерной лингвистики: Python. Инструмент NLTK**

Модуль NLTK — Natural Language Toolkit инструмент лингвиста‑разработчика. Возможности NLTK. Модули для токенизации и сплиттинга. Модули для проведения морфологического анализа: pos\_tag, лемматизация. Корпуса в NLTK. Ресурсы WordNet и FrameNet.

**Тема 14. Case study: извлечение именованных сущностей**

Именованные сущности как особые объекты в тексте. Для чего необходимо извлечение именованных сущностей и чем могут быть полезны полученные результаты. Правила для извлечения именованных сущностей. Языковая информация, привлекаемая для извлечения именованных сущностей. Системы извлечения именованных сущностей. Тестирование системы извлечения именованных сущностей.

**Тема 15. Digital Humanities**

Компьютерные технологии в гуманитарной сфере. Методы сетевого анализа, компьютерный анализ художественного текста. Знакомство с различными электронными ресурсами: жанровые корпуса и электронные инструменты для гуманитарных исследований.

**Тема 16. Синтаксис в области АОТ**

Введение в синтаксис. Структуры зависимостей. Структуры составляющих. Преимущества одного и другого подхода к синтаксическому анализу. Технологии shallow parsing. Технологии chunking. Синтаксическая разметка. Синтаксические анализаторы. Оценка качества работы синтаксических анализаторов.

**Тема 17. Машинное обучение (основы)**

Обучение с учителем и без учителя. Области применения машинного обучения. Подхотовка данных для проведения машинного обучения. Классификаторы, используемые в машинном обучении. Модуль scipy в Python. Инструмент Weka. Оценка результатов машинного обучения.

**Тема 18. Case study: классификация и кластеризация текстов**

Зачем и когда необходима классификация текстов. Отличие классификации от кластеризации. Примеры применения классификации и кластеризации текстов (в области информационной безопасности, в системах аналитики, в системах мониторинга медиа ресурсов).

**3. Тематическое планирование**

|  |  |  |  |
| --- | --- | --- | --- |
| **№****п/п** | **Тематическое содержание курса** | **Количество аудиторных часов** | **Основные виды деятельности** |
| **10 класс** |
| 1 | Введение в компьютерную лингвистику. Задачи в области обработки естественного языка. | 3 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 2 | Поиск в тексте. Использование регулярных выражений. | 4 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 3 | Корпусная лингвистика. Национальный корпус русского языка. | 5 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 4 | Программирование для задач компьютерной лингвистики: Python. Введение. Типы данных. Арифметические операции. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 5 | Программирование для задач компьютерной лингвистики: Python. Введение. Файловый ввод-вывод. Cтроки. | 2 | Выполнение практических заданий на компьютере. |
| 6 | Программирование для задач компьютерной лингвистики: Python. Введение. Cтроки. Условия if, elif, else | 4 | Выполнение практических заданий на компьютере. |
| 7 | Ключевые слова. Частотный словарь. N-граммы. Подсчет tf.idf | 4 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 8 | Программирование для задач компьютерной лингвистики: Python. Введение. Cписки и цикл for | 4 | Выполнение практических заданий на компьютере. |
| 9 | Препроцессинг. Нормализация. Обработка сырых текстовых данных в Python. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 10 | Морфология и автоматический морфологический анализ. Часть 1. Pymorphy2, pymystem3 | 4 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 11 | Морфология и автоматический морфологический анализ. Часть 2. | 3 | Выполнение практических заданий на компьютере. |
| 12 | Программирование для задач компьютерной лингвистики: Python. Введение. Повторение, подведение итогов. | 3 | Выполнение практических заданий на компьютере. |
| 13 | Автоматическая обработка языка. Препроцессинг (токенизация, стемминг). NLTK | 4 | Выполнение практических заданий на компьютере. |
| 14 | Морфология и автоматический морфологический анализ. Часть 3. pos\_tag в NLTK | 4 | Выполнение практических заданий на компьютере. |
| 15 | Оценивание систем АОТ, Precision, Recall, F-measure | 2 | Выполнение практических заданий на компьютере. |
| 16 | Введение в синтаксис. Структуры зависимостей. Структуры составляющих. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 17 | Синтаксис в компьютерной лингвистике | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 18 | Этапы построения систем АОТ. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 19 | Генерация текстов. Чатботы. Введение. Подготовка системы. Планирование. Подготовка входных данных. Первичная реализация. | 8 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 20 | Повторение. Лингвистические ресурсы. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 21 | Представление проектов | 2 | Презентации учеников |
|  | **Итого в 10 классе** | **68** |  |
| **11 класс** |
| 22 | Направления компьютерной лингвистики: повторение и новые знакомства. | 1 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 23 | Лингвистические ресурсы: WordNet и FrameNet. Работа с ними online и вNLTKПовторение: регулярные выражения в Python | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 24 | Повторение: Python ключевые навыки | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 25 | Морфологический анализ. Повторение. Оценка анализаторов (Precision, Recall, F-measure). | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 26 | Программирование для задач компьютерной лингвистики: Python. Numpy. | 1 | Выполнение практических заданий на компьютере. |
| 27 | Машинное обучение: основы. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 28 | Векторное представление документов, TF-IDF | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 39 | Case study: классификация и кластеризация текстов | 1 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 30 | Лингвистические ресурсы: WordNet и FrameNet. Работа с ними online и в NLTK | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 31 | Дистрибутивная семантика, word2vec | 1 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 32 | Извлечение именованных сущностей. | 2 | Выполнение практических заданий на компьютере. |
| 33 | Синтаксис (повторение). Структуры зависимостей. Структуры составляющих. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 34 | Синтаксическая разметка. Синтаксические анализаторы. Оценка синтаксических анализаторов. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 35 | Технологии shallow parsing. Технологии chunking. | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 36 | Digital Humanties | 2 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 37 | Повторение | 3 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 38 | Подготовка проектов. Обсуждение стратегий реализации. | 3 | Лекция и дискуссия по обозначаемой преподавателем проблематике. Решение задач. Выполнение практических заданий на компьютере. |
| 39 | Представление проектов. | 2 | Презентации учеников |
|  | **Итого в 11 классе** | **34** |  |
|  | **Итого всего** | **102** |  |

В воспитании обучающихся приоритетом является создание благоприятных условий для приобретения обучающимися опыта осуществления социально значимых дел:

— опыт дел, направленных на заботу о своей семье, родных и близких;

— трудовой опыт, опыт участия в производственной практике;

— опыт дел, направленных на пользу своему родному городу или селу, стране в целом, опыт деятельного выражения собственной гражданской позиции;

— опыт природоохранных дел;

— опыт разрешения возникающих конфликтных ситуаций в школе, дома или на улице;

— опыт самостоятельного приобретения новых знаний, проведения научных исследований, опыт проектной деятельности;

— опыт изучения, защиты и восстановления культурного наследия человечества, опыт создания собственных произведений культуры, опыт творческого самовыражения;

— опыт ведения здорового образа жизни и заботы о здоровье других людей;

— опыт оказания помощи окружающим, заботы о малышах или пожилых людях, волонтерский опыт;

— опыт самопознания и самоанализа, опыт социально приемлемого самовыражения и самореализации.

Выделение данного приоритета связано с особенностями обучающихся юношеского возраста: с их потребностью в жизненном самоопределении, в выборе дальнейшего жизненного пути, который открывается перед ними на пороге самостоятельной взрослой жизни.

**Дополнительные материалы**

**Учебно-методическое обеспечение образовательной деятельности**

1. Поляков К.Ю., Еремин, Е.А. Информатика. Углубленный уровень. Учебник для 11 класса. В 2-х частях. Часть 1. М.: БИНОМ Лаборатория знаний.
2. Поляков К.Ю., Еремин, Е.А. Информатика. Углубленный уровень. Учебник для 11 класса. В 2-х частях. Часть 2. М.: БИНОМ Лаборатория знаний.
3. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб.пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
4. Волкова И. А. Введение в компьютерную лингвистику. Практические аспекты создания лингвистических процессоров. Учебное пособие для студентов факультета ВМК МГУ. — Издательский отдел факультета ВМК МГУ им. М.В.Ломоносова Москва, 2006. — С. 44.