|  |  |  |  |
| --- | --- | --- | --- |
| **Национальный**  **исследовательский университет**  **«Высшая школа экономики»**  **Лицей** | | | **Приложение 89**  УТВЕРЖДЕНО  педагогическим советом  Лицея НИУ ВШЭ  протокол от 04.12.2017 |
|  |  | | |

**Рабочая программа учебного предмета (курса)**

**«Компьютерная лингвистика»**

**10-11 класс**

**Автор:**

Нестеренко Л.В.

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

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

В соответствии с Федеральным государственным образовательным стандартом среднего общего образования (10-11 кл.) освоение учебного предмета «Компьютерная лингвистика» предполагает достижение личностных, метапредметных и предметных результатов.

Личностные результаты освоения учебного предмета включают в себя:

* понимание основ поведения в сфере компьютерной лингвистике для осуществления осознанного выбора будущей специализации;
* использование приобретенных знаний и умений в практиче­ской деятельности и повседневной жизни для решения практических задач, связанных с жизненными ситуа­циями; совершенствования собственной познавательной деятельно­сти;
* владение навыками познавательной рефлексии как осознания совершаемых действий и мыслительных процессов, их результатов и оснований, границ своего знания и незнания, новых познавательных задач и средств их достижения.
* готовность и способность к самостоятельной информационно-познавательной деятельности, включая умение ориентироваться в различных источниках информации, критически оценивать и интерпретировать информацию, получаемую из различных источников;
* владение навыками познавательной, учебно-исследовательской и проектной деятельности, навыками разрешения проблем; способность и готовность к самостоятельному поиску методов решения практических задач, применению различных методов познания;
* умение самостоятельно определять цели деятельности и составлять планы деятельности; самостоятельно осуществлять, контролировать и корректировать деятельность; использовать все возможные ресурсы для достижения поставленных целей и реализации планов деятельности; выбирать успешные стратегии в различных ситуациях;

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 компьютереной 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

Метапредметные результаты освоения учебного предмета включают в себя:

* владение навыками познавательной, учебно-исследовательской и проектной деятельности, навыками разрешения проблем; способность и готовность к самостоятельному поиску методов решения практических и аналитических задач;
* готовность и способность к самостоятельной информационно-познавательной деятельности, включая умение ориентироваться в различных источниках информации, критически оценивать и интерпретировать информацию, получаемую из различных источников;
* умение использовать средства информационных и коммуникационных технологий (далее – ИКТ) в решении когнитивных, коммуникативных и организационных задач с соблюдением требований эргономики, техники безопасности, гигиены, ресурсосбережения, правовых и этических норм, норм информационной безопасности;

**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 выраженый. 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**

|  |  |  |  |
| --- | --- | --- | --- |
| **№ темы** | **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 обозначенной преподавателем проблематике. |
| 2. | Поиск в тексте. Использование регулярных выражений. | 4 | Лекция и практические занятия. |
| 3. | Корпусная лингвистика. Национальный корпус русского языка. | 5 | Лекция и практические занятия. |
| 4. | Программирование для задач компьютерной лингвистики: Python. Введение. Типы данных. Арифметические операции. | 2 | Практическое занятие. |
| 5. | Программирование для задач компьютерной лингвистики: Python. Введение. Файловый ввод-вывод. Cтроки. | 2 | Практическое занятие. |
| 6. | Программирование для задач компьютерной лингвистики: Python. Введение. Cтроки. Условия if, elif, else | 4 | Практическое занятие. |
| 7. | Ключевые слова. Частотный словарь. N-граммы. Подсчет tf.idf | 4 | Лекция и практические занятия. |
| 8. | Программирование для задач компьютерной лингвистики: Python. Введение. Cписки и цикл for | 4 | Практические занятия. |
| 9. | Препроцессинг. Нормализация. Обработка сырых текстовых данных в Python. | 2 | Лекция и практические занятия. |
| 10. | Морфология и автоматический морфологический анализ. Часть 1. Pymorphy2, pymystem3 | 4 | Лекция и практическое занятие. |
| 11. | Морфология и автоматический морфологический анализ. Часть 2. | 3 | Практическое занятие. |
| 12. | Программирование для задач компьютерной лингвистики: Python. Введение. Повторение, подведение итогов. | 3 | Практическое занятие. |
| 13. | Автоматическая обработка языка. Препроцессинг (токенизация, стемминг). NLTK | 4 | Практическое занятие. |
| 14. | Морфология и автоматический морфологический анализ. Часть 3. pos\_tag в NLTK | 4 | Практическое занятие. |
| 15. | Оценивание систем АОТ, Precision, Recall, F-measure | 2 | Практическое занятие. |
| 16. | Введение в синтаксис. Структуры зависимостей. Структуры составляющих. | 2 | Лекция |
| 17. | Синтаксис в компьютерной лингвистике | 2 | Практическое занятие. |
| 18. | Этапы построения систем АОТ. | 2 | Лекция и практические занятия. |
| 19. | Генерация текстов. Чатботы. Введение. Подготовка системы. Планирование. Подготовка входных данных. Первичная реализация. | 8 | Лекция и практические занятия. |
| 20. | Повторение. Лингвистические ресурсы. | 2 | Лекция и практические занятия. |
| 21. | Представление проектов | 2 | Презентации учеников |
|  | **Итого в 10 классе** | **68** |  |
| **11 класс** | | | |
| 22. | Направления компьютерной лингвистики: повторение и новые знакомства. | 2 | Лекция и дисскуссия по обозначенной тематике. |
| 23. | Лингвистические ресурсы: WordNet и FrameNet. Работа с ними online и в  NLTK  Повторение: регулярные выражения в Python | 4 | Лекция и практические занятия. |
| 24. | Повторение: Python ключевые навыки | 4 | Лекция и практические занятия. |
| 25. | Морфологический анализ. Повторение. Оценка анализаторов (Precision, Recall, F-measure). | 4 | Лекция и практические занятия. |
| 26. | Программирование для задач компьютерной лингвистики: Python. Numpy. | 2 | Практическое занятие. |
| 27. | Машинное обучение: основы. | 4 | Лекция и практические занятия. |
| 28. | Векторное представление документов, TF-IDF | 4 | Лекция и практические занятия. |
| 39. | Case study: классификация и кластеризация текстов | 2 | Практическое занятие. |
| 30. | Лингвистические ресурсы: WordNet и FrameNet. Работа с ними online и в NLTK | 4 | Лекция и практические занятия. |
| 31. | Дистрибутивная семантика, word2vec | 2 | Лекция. |
| 32. | Извлечение именованных сущностей. | 4 | Практическое занятие. |
| 33. | Синтаксис (повторение). Структуры зависимостей. Структуры составляющих. | 4 | Лекция и практические занятия. |
| 34. | Синтаксическая разметка. Синтаксические анализаторы. Оценка синтаксических анализаторов. | 4 | Лекция и практические занятия. |
| 35. | Технологии shallow parsing. Технологии chunking. | 4 | Лекция и практические занятия. |
| 36. | Digital Humanties | 4 | Лекция и практические занятия. |
| 37. | Повторение | 6 | Лекция и практические занятия. |
| 38. | Подготовка проектов. Обсуждение стратегий реализации. | 6 | Лекция и практические занятия. |
| 39. | Представление проектов. | 4 | Презентации учеников |
|  | **Итого в 11 классе** | **68** |  |
|  | **Итого всего** | **136** |  |

**Дополнительные материалы**

**Учебно-методическое обеспечение образовательной деятельности**

Базовыми учебниками учебного курса являются:

1. 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.
2. 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.