МИНОБРНАУКИ РОССИИ

**Федеральное государственное бюджетное образовательное учреждение высшего образования «Волгоградский государственный технический университет»**

**Факультет электроники и вычислительной техники**

УТВЕРЖДЕНО
Факультет электроники и вычислительной техники
Декан Авдеюк О.А.
г.

# Компьютерная лингвистика

## рабочая программа дисциплины (модуля, практики)

| | |
|---|---|
| Закреплена за кафедрой | **Системы автоматизированного проектирования и поискового конструирования** |
| Учебный план | Направление 09.03.01 Информатика и вычислительная техника<br>Профиль "Вычислительные машины, комплексы, системы и сети" |
| Профиль | |
| Квалификация | **Бакалавр** |
| Срок обучения | **4 года** |

Форма обучения **очная** Общая трудоемкость **5 ЗЕТ**

Виды контроля в семестрах: экзамены 5

| Семестр(Курс.Номер семестра на курсе) | **5(3.1)** | | Итого | |
|---|---|---|---|---|
| | УП | ПП | УП | ПП |
| Лекции | 32 | 32 | 32 | 32 |
| Лабораторные | 32 | 32 | 32 | 32 |
| Итого ауд. | 64 | 64 | 64 | 64 |
| Контактная работа | 64.35 | 64.35 | 64.35 | 64.35 |
| Сам. работа | 80 | 95.65 | 80 | 95.65 |
| Часы на контроль | 35.65 | 20 | 35.65 | 20 |
| Практическая подготовка | 0 | 0 | 0 | 0 |
| Итого трудоемкость в часах | 180 | 180 | 0 | 0 |

ЛИСТ ОДОБРЕНИЯ, СОГЛАСОВАНИЯ И АКТУАЛИЗАЦИИ РАБОЧЕЙ ПРОГРАММЫ

Разработчик(и) программы:

Доцент Коробкин Д.М. к.т.н.

Рецензент(ы):
(при наличии)

Рабочая программа дисциплины (модуля, практики)

**Компьютерная лингвистика**

разработана в соответствии с ФГОС ВО:

Федеральный государственный образовательный стандарт высшего образования - бакалавриат по направлению подготовки 09.03.01 Информатика и вычислительная техника (приказ Минобрнауки России от 19.09.2017 г. № 929)

составлена на основании учебного плана:

Направление 09.03.01 Информатика и вычислительная техника
Профиль "Вычислительные машины, комплексы, системы и сети"

Профиль:

утвержденного учёным советом вуза от 05.06.2019 протокол № 12.

Рабочая программа одобрена на заседании кафедры

**Системы автоматизированного проектирования и поискового конструирования**

номер протокола 2019 г.
Зав. кафедрой Щербаков Максим Владимирович

СОГЛАСОВАНО:
Факультет электроники и вычислительной техники
Председатель НМС факультета: Авдеюк О.А.

Протокол заседания НМС от
г. №

Рабочая программа дисциплины (модуля, практики) актуализирована 31.08.2023

| 1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ). ВИД, ТИП ПРАКТИКИ, СПОСОБ И ФОРМА (ФОРМЫ) ЕЕ ПРОВЕДЕНИЯ. |
|---|
| Цель изучения дисциплины - Целью дисциплины является подготовка специалистов к проектированию и разработке трансляторов искусственных языков, изучение основных методов трансляции, основанных на теории формальных языков. |
| Задачи изучения дисциплины: |
| - знать теории формальных языков; |
| - знать синтаксически-ориентированные методы трансляции; |
| - знать методы организации информации при конструировании трансляторов. |
| |

| 2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ) В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ | |
|---|---|
| Цикл (раздел) ОП: | Б1.В |
| **2.1** | **Требования к предварительной подготовке обучающегося:** |
| 2.1.1 | Вычислительная математика |
| 2.1.2 | Системный анализ |
| 2.1.3 | Математическая логика и теория алгоритмов |
| **2.2** | **Дисциплины (модули) и практики, для которых освоение данной дисциплины (модуля) необходимо как предшествующее:** |
| 2.2.1 | Выполнение и защита выпускной квалификационной работы |
| 2.2.2 | Информационная культура студента |
| 2.2.3 | Производственная практика: Преддипломная практика |

| 3. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ) |
|---|
| **УК-1: Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач** |
| *УК-1.1: Знать: методики сбора и обработки информации; анализа и обобщения его результатов для решения поставленной задачи; информационные ресурсы для поиска информации в соответствии с поставленной задачей.* |
| Результаты обучения: знает подходы в рамках дисциплины "Компьютерная лингвистика" к методикам сбора и обработки информации; анализу и обобщению его результатов для решения поставленной задачи; информационным ресурсам для поиска информации в соответствии с поставленной задачей. |
| *УК-1.2: Уметь: применять методики поиска, сбора и обработки информации; осуществлять критический анализ и синтез информации, полученной из разных источников; использовать системный подход для решения поставленных задач.* |
| Результаты обучения: умеет принимать теоретические и практические подходы в рамках дисциплины "Компьютерная лингвистика" применять методики поиска, сбора и обработки информации; осуществлять критический анализ и синтез информации, полученной из разных источников; использовать системный подход для решения поставленных задач. |
| *УК-1.3: Владеть: методами поиска, сбора и обработки, критического анализа и синтеза информации; методикой системного подхода для решения поставленных задач; логичным и последовательным изложением выявленной информации со ссылками на информационные ресурсы.* |
| Результаты обучения: владеет практическими навыками применения методами поиска, сбора и обработки, критического анализа и синтеза информации; методикой системного подхода для решения поставленных задач; логичным и последовательным изложением выявленной информации со ссылками на информационные ресурсы. |
| **ПК-3: Способен проектировать программное обеспечение и базы данных, разрабатывать и отлаживать программный код, преобразовать программный код и структуры данных для повышения их эффективности, проводить рефакторинг программных приложений и баз данных** |
| *ПК-3.1: Знать: методы проектирования программного обеспечения и баз данных, разработки и настройки программного кода, преобразований программного кода и структуры данных для повышения их эффективности, проведения рефакторинга программных приложений и баз данных.* |
| Результаты обучения: знает подходы в рамках дисциплины "Компьютерная лингвистика" к методам проектирования программных обеспечений и базам данных, разработкам и настройкам программного кода, преобразованиям программного кода и структурам данных для повышения их эффективности, проведения рефакторинга программных приложений и баз данных. |
| *ПК-3.2: Уметь: проектировать программное обеспечение и базы данных, разрабатывать и отлаживать программный код, преобразовать программный код и структуры данных для повышения их эффективности, проводить рефакторинг программных приложений и баз данных.* |
| Результаты обучения: умеет принимать теоретические и практические подходы в рамках дисциплины "Компьютерная лингвистика" проектировать программное обеспечение и базы данных, разрабатывать и отлаживать программный код, преобразовать программный код и структуры данных для повышения их эффективности, проводить рефакторинг программных приложений и баз данных. |

*ПК-3.3: Владеть навыком: проектирования программного обеспечения и базы данных, разработок и настройки программного кода, преобразовывать программный код и структуры данных для повышения их эффективности, провождения рефакторинга программных приложений и баз данных.*

Результаты обучения: владеет практическими навыками применения проектирования программного обеспечения и базы данных, разработок и настройки программного кода, преобразовывать программный код и структуры данных для повышения их эффективности, провождения рефакторинга программных приложений и баз данных.

## 4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ)

| Код занятия | Наименование разделов и тем /вид занятия/ | Семестр / Курс | Часов | Форма контроля |
|---|---|---|---|---|
| 1 | **Раздел 1. Компьютерная морфология и синтаксис** | | | |
| 1.1 | Компьютерная морфология /Тема/ | 5 | 0 | |
| 1.1.1 | Компьютерная морфология /Лек/ | 5 | 4 | Эк |
| 1.2 | Компьютерный синтаксис /Тема/ | 5 | 0 | |
| 1.2.1 | Компьютерная морфология /Лек/ | 5 | 4 | Эк |
| 1.2.2 | Контрольной работы /Ср/ | 5 | 20 | К |
| 1.2.3 | Морфологический и синтаксический анализ /Лаб/ | 5 | 8 | Эк |
| 2 | **Раздел 2. Машинное обучение в лингвистике** | | | |
| 2.1 | Машинное обучение в лингвистике /Тема/ | 5 | 0 | |
| 2.1.1 | Машинное обучение в лингвистике /Лек/ | 5 | 4 | Эк |
| 2.2 | Корпусная лингвистика /Тема/ | 5 | 0 | |
| 2.2.1 | Корпусная лингвистика /Лек/ | 5 | 4 | Эк |
| 2.2.2 | Контрольной работы /Ср/ | 5 | 20 | К |
| 2.2.3 | Корпусная лингвистика и машинное обучение /Лаб/ | 5 | 8 | Эк |
| 3 | **Раздел 3. Информационный поиск и извлечение информации** | | | |
| 3.1 | Информационный поиск /Тема/ | 5 | 0 | |
| 3.1.1 | Информационный поиск /Лек/ | 5 | 4 | Эк |
| 3.2 | Извлечение информации из текста /Тема/ | 5 | 0 | |
| 3.2.1 | Извлечение информации из текста /Лек/ | 5 | 4 | Эк |
| 3.2.2 | Контрольной работы /Ср/ | 5 | 20 | К |
| 3.2.3 | Статистический анализ текстовых данных /Лаб/ | 5 | 8 | Эк |
| 4 | **Раздел 4. Применение компьютерной лингвистики** | | | |
| 4.1 | Диалоги и чат-боты /Тема/ | 5 | 0 | |
| 4.1.1 | Диалоги и чат-боты /Лек/ | 5 | 4 | Эк |
| 4.2 | Анализ тональности текста /Тема/ | 5 | 0 | |
| 4.2.1 | Анализ тональности текста /Лек/ | 5 | 4 | Эк |
| 4.2.2 | Анализ тональности текста /Ср/ | 5 | 20 | К |
| 4.2.3 | Анализ тональности текста /Лаб/ | 5 | 8 | Эк |
| 4.2.4 | Опрос /КоРа/ | 5 | 0.35 | КоРа |
| 4.2.5 | Подготовка к экзамену /Экзамен/ | 5 | 20 | Эк |
| 4.2.6 | Контрольная работа /Контр.раб./ | 5 | 15.65 | К |

Примечание. Формы контроля: Эк – экзамен, К- контрольная работа, Ко- контрольный опрос, Сз- семестровое задание, З-зачет, ОП-отчет по практике.

## 5. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства планируемых результатов обучения представлены в виде фондов оценочных средств (ФОС), разработанных в соответствии с локальным нормативным актом университета. ФОС может быть представлен в Приложении к рабочей программе.

5.1 Контрольные вопросы и задания
Виды анализа слов
Анализ морфологии на основе правил
Статистические методы анализа слов
Морфологический анализатор Pymorphy2
Анализатор Mystem
Понятие парсинга
Грамматики зависимостей
Статистический парсинг
Лингвистический процессор ЭТАП

Парсер Stanford NLP
Парсер Link Grammar Parser
Парсер NLTK
Методы машинного обучения
Регрессионные методы
Искусственные нейронные сети
Кластеризация
Основные свойства корпуса
Разметка корпусов.
Поисковый индекс
Релевантность, полнота, точность
Фильтрация и ранжирование
Какую информацию извлекаем?
Распознавание сущностей
Разрешение анафоры и кореферентности
Установление референта. Автоматические подходы.
Извлечение отношений. Обучение моделей на размеченных текстах.
Извлечение событий
Особенности диалога на естественном языке
Архитектура диалоговых систем
Принципы работы чат-ботов
Обучение диалоговых систем на реальных диалогах.
Как измерить тональность текста
Оценка качества работы алгоритмов

5.2 Темы письменных работ (контрольная работа)

Контрольная работа по компьютерной лингвистике представляет собой законченную работу, включающую проектные решения по разработке компонента информационной системы. Выбор темы контрольной работы осуществляется в соответствии с перечнем заданий.

Контрольная работа состоит из 3 разделов.
1. парсинг текстовых данных
2. извлечение информации из текста
3. анализ тональности текста

В результате формируется пояснительная записка, содержащая результаты выполнения по разделам и содержащая описание алгоритмов, методов и стороннего использованного ПО.
Контрольная работа оформляется в соответствии с шаблоном и оценивается на защите. Каждый раздел оценивается максимум на 20 баллов.
Темы (системы для проектирования)
Вариант 1. https://v1.ru/
Вариант 2. https://v102.ru/
Вариант 3. https://bloknot-volgograd.ru/
Вариант 4. http://vpravda.ru/
Вариант 5. https://volgograd.kp.ru/
Вариант 6. https://gorvesti.ru/
Вариант 7. https://www.novostivolgograda.ru/
Вариант 8. http://www.volgograd.ru/
Вариант 9. https://www.volzskij.wx1.ru/
Вариант 10. https://riac34.ru/

- Раздел «Парсинг текстовых данных» предполагает создание программного краулер, который должен считывать новостную ленту с первой страницы сайта. Периодичность повторения устанавливается пользователем. Данные заполняются в БД MongoDB. Обязательные поля для текста новости: - Название новости, - Дата новости, - Ссылка на новость, - Текст новости, - Ссылка на видео (если есть), - Количество просмотров новости (если есть), - Количество комментариев новости (если есть).

- Раздел «Извлечение информации из текста» предполагает создание программного модуля для анализа новостей из БД. Необходимо выделить с помощью ПО «Томита-парсер» упоминание в тексте значимых персон Волгоградской области и достопримечательностей. Зафиксировать в БД предложения с их упоминанием для дальнейшего анализа тональности. С помощью ПО «Spark MlLib» осуществить анализ модели «word2vec» на всем объеме новостных статей из БД. Для персон Волгоградской области и достопримечательностей определить контекстные синонимы и слова, с которыми они упоминались в тексте.
Персоны – ресурс https://global-volgograd.ru/person
Достопримечательности- ресурс https://avolgograd.com/sights?obl=vgg

- Раздел «Анализ тональности текста» предполагает создание программного модуля для выявления тональности высказываний по отношению к персонам Волгоградской области и достопримечательностям. Необходимо использовать либо подход на основе правил и словарей, либо методы машинного обучения.

5.3 Фонды оценочных средств

Оценочные средства планируемых результатов обучения представлены в виде фондов оценочных средств (ФОС), разработанных в соответствии с локальным нормативным актом университета. ФОС представлен в Приложении к рабочей программе.

5.4 Перечень видов оценочных средств

Наименование оценочного средства 1: Контрольная работа. Контрольная работа по компьютерной лингвистике представляет собой законченную работу, включающую проектные решения по разработке компонента информационной системы. Позволяет оценить умения обучающихся самостоятельно конструировать свои знания в процессе решения практических задач и проблем, ориентироваться в информационном пространстве и уровень сформированности аналитических, исследовательских навыков, навыков практического и творческого мышления.

Наименование оценочного средства 2: Собеседование. Средство контроля, организованное как специальная беседа преподавателя с обучающимся на темы, связанные с изучаемой дисциплиной, и рассчитанное на выяснение объема знаний обучающегося по определенному разделу, теме, проблеме и т.п. Собеседование применяется на экзамене.

## 6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ)

### 6.1. Рекомендуемая литература

| | Авторы, составители | Заглавие | Издательство, год. | Электронный адрес |
|---|---|---|---|---|
| Л.1 | Городецкий Б. Ю. | Компьютерная лингвистика | М.: Прогресс, 1989 | |
| Л.2 | Гаврилова Т. А., Кудрявцев Д. В., Муромцев Д. И. | Инженерия знаний. Модели и методы: учебник | Санкт-Петербург: Лань, 2020 | https://e.lanbook.com/reader/book/147337/#3 |
| Л.3 | Григорьев Ю. Д., Мартыненко Г. Я. | Последовательности типа Фибоначчи. Теория и прикладные аспекты: учебное пособие | Санкт-Петербург: Лань, 2021 | https://e.lanbook.com/book/167431 |
| Л.4 | Исакова А. И. | Основы информационных технологий: учебное пособие | Москва: ТУСУ□, 2016 | https://e.lanbook.com/book/110256 |
| Л.5 | Переходько И. В. | Компьютерные технологии в переводе: учебное пособие | Оренбург: ОГУ, 2018 | https://e.lanbook.com/book/159856 |

### 6.2. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"

| | |
|---|---|
| Э1 | Онлайн-курс Компьютерная лингвистика в СДО «Moodle» |

### 6.3 Перечень программного обеспечения

| | |
|---|---|
| 6.3.1.1 | Adobe Acrobat Reader DC — бесплатное решение для просмотра файлов PDF |
| 6.3.1.2 | LibreOffice — офисный пакет |

### 6.4 Перечень информационных справочных систем и электронных библиотечных систем (ЭБС)

| | |
|---|---|
| 6.3.2.1 | Библиотека (НТБ), http://library.vstu.ru/sci-nci |
| 6.3.2.2 | Электронная информационно-образовательная среда университета,http://eos.vstu.ru |
| 6.3.2.3 | ЭБС "Лань", https://e.lanbook.com/ |
| 6.3.2.4 | ЭБС "Book.ru", https://www.book |
| 6.3.2.5 | Электронная библиотека "Grebennikon", https://grebennikon.ru/ |

## 7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ) /ОБОРУДОВАНИЕ

| | |
|---|---|
| 7.1 | Мультимедийная учебная аудитория для проведения занятий лекционного и семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации. /Учебная доска, учебная мебель,интерактивная трибуна, видеопроектор. |
| 7.2 | Лаборатория информационных технологий. /Учебная мебель, компьютерная техника, оснащенная программным обеспечением, доступом в Интернет и в электронную информационно-образовательную среду университета. |
| 7.3 | Аудитория для самостоятельной работы обучающихся./Учебная мебель, компьютерная техника с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду университета (читальный зал информационно-библиотечного центра. |

## 8. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ, ПРАКТИКИ)

Организация образовательного процесса по данной дисциплине регламентируется учебным планом и расписанием учебных занятий. При формировании своей индивидуальной образовательной траектории обучающийся имеет право на перезачет дисциплины (переаттестации ее части), если она была освоена в процессе предшествующего обучения. Перезачёт (переаттестации ее части)освобождает обучающегося от необходимости повторного освоения дисциплины

(полностью или частично).

Учебный процесс при преподавании курса основывается на использовании традиционных, инновационных и информационных образовательных технологий. Традиционные образовательные технологии представлены лекциями и практическими занятиями. Инновационные образовательные технологии используются в виде широкого применения активных и интерактивных форм проведения занятий. Информационные образовательные технологии реализуются путем активизации самостоятельной работы студентов в электронной информационной образовательной среде.

Лекционный курс предполагает систематизированное изложение основных вопросов учебного плана. На первой лекции лектор информирует студентов о рекомендуемой литературе и электронных источниках информации по дисциплине, с указанием, какой учебник (учебное пособие) является базовым.

Лабораторные работы предполагают выполнение и отчет заданий по темам, рассмотренным на лекционных и закрепленных на практических занятиях. Каждому лабораторному занятию предшествует самостоятельная подготовка студента, включающая: ознакомление с содержанием лабораторной работы по методическим указаниям; проработку теоретической части по лекционному материалу и учебникам, рекомендованным в методических указаниях;

Самостоятельная работа студентов включает изучение законспектированного на лекционных занятиях материала, дополнение его с учетом рекомендованной по данной теме литературы, самостоятельную подготовку к лабораторным работам, самостоятельное выполнение и оформление заданий контрольной работы, аналогичных выполненным на занятиях.
Перечень методических указаний для освоения дисциплины представлен в таблице 6.1.3
В течении семестра для студентов проводятся групповые текущие консультации по учебной дисциплине, а также консультация перед экзаменом.

Методические рекомендации по обучению лиц с ограниченными возможностями здоровья и инвалидов
Профессорско-педагогический состав знакомится с психолого-физиологическими особенностями обучающихся инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ), индивидуальными программами реабилитации инвалидов (при наличии). При необходимости осуществляется дополнительная поддержка преподавания тьюторами, психологами, социальными работниками, прошедшими подготовку ассистентами.

В соответствии с методическими рекомендациями Минобрнауки РФ (утв. 8 апреля 2014 г. N АК-44/05вн), в курсе предполагается использовать социально-активные и рефлексивные методы обучения, технологии социокультурной реабилитации с целью оказания помощи в установлении полноценных межличностных отношений с другими студентами, создании комфортного психологического климата в студенческой группе. Подбор и разработка учебных материалов производятся с учетом предоставления материала в различных формах: аудиальной, визуальной, с использованием специальных технических средств и информационных систем.
Освоение дисциплины лицами с ОВЗ осуществляется с использованием средств обучения общего и специального назначения (персонального и коллективного использования). Материально-техническое обеспечение предусматривает приспособление аудиторий к нуждам лиц с ОВЗ (при необходимости).