Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Филологический факультет

УТВЕРЖДЕНО:
Декан
И. В.Тубалова

Рабочая программа дисциплины

Обработка естественного языка на Python

по направлению подготовки

45.03.03 Фундаментальная и прикладная лингвистика

Направленность (профиль) подготовки:
Фундаментальная и прикладная лингвистика

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
А.В. Васильева

Председатель УМК
Ю.А. Тихомирова

Томск – 2024

1. Цель и планируемые результаты освоения дисциплины (модуля)

Целью освоения дисциплины является формирование у студентов целостной системы теоретических знаний и практических навыков в области анализа текстов и текстовых массивов с применением математических методов и формирует следующие компетенции:

ПК-4. Способен разрабатывать программный код при решении задач автоматической обработки текстов

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.1. Применяет способы формализации и алгоритмизации поставленных задач в сфере автоматической обработки текстов

ИПК-4.2. Создает программный код с использованием языков программирования и манипулирования данными в сфере автоматической обработки текстов

2. Задачи освоения дисциплины

– Изучение методов обработки естественного языка, применение междисциплинарных методов в обработке исследовательских данных.

– Научиться применять понятийный математический аппарат в области лингвистики для решения практических задач профессиональной деятельности.

– Приобрести навыки хранения, структуризации, анализа и визуализации текстового массива данных

3. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Семестр 3, зачет.

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 2 з.е., 72 часа, из которых:
– лекции: 8 ч.;
– семинарские занятия: 0 ч.
– практические занятия: 24 ч.;
– лабораторные работы: 0 ч.
в том числе практическая подготовка: 24 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины (модуля), структурированное по темам

Тема 1. Квантитативная лингвистика, статистические методы анализа в лингвистических исследованиях. Закон Ципфа

Квантитативная лингвистика, статистические методы анализа в лингвистических исследованиях. Основные статистические категории: выборка и совокупности, типы переменных, их классификация применительно к соответствующему уровню лингвистического анализа. Частота как характеристика употребительности слова в тексте. Закон Ципфа. Уточнение закона Ципфа: закон Ципфа-Мандельброта. Закон Ципфа и структура реального текста. Статистические методы лингвистических исследований: корреляционный анализ; дисперсионный анализ (ANOVA); кластерный анализ; факторный анализ. Описательная статистика в лингвистических исследованиях.

Тема 2. Квантитативные методы в компаративистике.

Поиск и сравнение лексем в корпусах, метрики сравнения: IPM, TF-IDF, LL-score, коэффициент Жуйана. Квантитативные методы в компаративистике. Глоттохронология. Опыт применения лексикостатистического исчисления глубины дивергенции применительно к индоевропейским языкам. Исчисление относительной хронологии дивергенции группы языков (германской, романской) лингвостатистическим методом.

Тема 3. Квантитативная типология Дж. Гринберга

Применение статистических методов в основных разделах лингвистики. Фоностатистика. Статистико-комбинаторные, дистрибутивностатистические и дешифровочные методы в грамматике. Квантитативная типология Дж. Гринберга. Опыт квантитативного обоснования морфологических типов (корреляции между морфологическими признаками). Квантитативные характеристики морфологии изучаемых языков в области словоизменения и словообразования. Исчисление индекса Дж. Гринберга на материале текстов родного и изучаемого языков.

Тема 4. Квантитативные методы и корпусная лингвистика.

Квантитативные методы и корпусная лингвистика. Статистические методы выделения терминов, устойчивых словосочетаний, синонимических групп, семантических полей. Корпусы языков (изучаемых и родных), википедии (практическое изучение).

Тема 5. Предмет компьютерной лингвистики.

Автоматическое понимание смысла текста. Автоматический синтез речи. Декларативные средства. Процедурные средства.

Тема 6. Автоматизация составления и лингвистической обработки машинных словарей

Объём текстов для автоматической выборки. Особенности составления словарей словосочетаний. Установление парадигматических отношений между терминами.

Тема 7. Автоматизация процессов обнаружения и исправления ошибок при вводе текстов в ЭВМ.

Принципы распознавания орфографических ошибок. Принципы распознавания синтаксических ошибок.

Тема 8. Автоматическое индексирование документов и информационных запросов.

Присвоение классификационных индексов. Поисковые образы документов.

Тема 9. Аппаратное и программное обеспечение информационных технологий.

Составные части ЭВМ. Языки программирования.

Тема 10. Машинный перевод текстов

Машинный перевод текстов с одних естественных языков на другие. Семантический анализ текста на ИЯ и пословный перевод. Фразовый перевод и «переводческая память»

9. Текущий контроль по дисциплине

Текущий контроль успеваемости обучающихся направлен на определение соответствия результатов обучения после освоения элемента по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, тестов по лекционному

материалу, разработки кода, выполнения домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр. Примерные задания текущего контроля:

– С помощью программных средств введите речевой сигнал в компьютер и отцифруйте его. При формировании звукового файла обратите внимание на частоту дискретизации (Гц), длительность звучания записываемого сигнала (с/мин), формат хранения (WAV, AU, AIFF, RAW и др.), количество записываемых каналов (моно/стерео), разрешение сигнала (16, 24, 32 бит).

– При выполнении задания необходимо варьировать показатели записи и выявлять различия.

– После оцифровки сигнала приступить к первичному редактированию в программе GoldWave (например, прослушать записанный сигнал, вырезать или вставить какие-либо отрезки звучания, соединить разные файлы и др.), подготовить файл для работы с другими программами.

– Сегментируйте речевой сигнал на фразы, слова, словосочетания, слоги, отдельные звуки.

– С помощью опционалов разных программ установите метки сегментации. 6 Отредактируйте сигнал, используя возможности разных программ (например, измените громкость сигнала, частоту дискретизации, поработайте с эффектами, модифицируйте частоту основного тона и др.). В качестве материала для работы используйте данные звуковых хрестоматий.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет проводится в письменной и устной форме по выбранному проекту. Проект предполагает логическое изложение теоретического блока с привязкой к практической деятельности и проверяет уровень сформированности следующих компетенций: ПК-4, ИПК-4.1, ИПК-2.

Примерный перечень теоретических вопросов

Тема 1. Квантитативная лингвистика, статистические методы анализа в лингвистических исследованиях. Закон Ципфа. 1. Основные понятия квантитативной лингвистики. 2. Основы статистических методов анализа в лингвистических исследованиях. 3. Основные статистические категории. 4. Частота употребительности слова в тексте. 5. Закон Ципфа и закон Ципфа-Мандельброта. 6. Статистические методы лингвистических исследований. 7. Описательная статистика в лингвистических исследованиях.

Тема 2. Квантитативные методы в компаративистике. Глоттохронология. 8. Основные характеристики квантитативных методов. 9. Особенности глоттохронологии. 10. Лексикостатистическое исчисление глубины дивергенции. 11. Примеры использования лингвостатистического метода.

Тема 3. Квантитативная типология Дж. Гринберга. 12. Применения статистических методов в лингвистике. 13. Основы фоностатистики. 14. Квантитативные методы в грамматике. 15. Основы типологии Дж. Гринберга. 16. Основы квантитативного обоснования морфологических типов. 17. Квантитативные характеристики морфологии в области словоизменения и словообразования. 18. Пример исчисления индекса Дж. Гринберга.

Тема 4. Квантитативные методы и корпусная лингвистика. 19. Характеристика и использование квантитативных методов. 20. Основные характеристики корпусной лингвистики. 21. Практическое применение статистических методов. 22. Использование корпусов языков.

Тема 5. Предмет компьютерной лингвистики. 23. Особенности автоматического понимания смысла текста. 24. Особенности автоматического синтеза речи. 25. Особенности декларативных средств. 26. Особенности процедурных средств.

Тема 6. Автоматизация составления и лингвистической обработки машинных словарей. 27. Требования к объёмам текстов для автоматической выборки. 28. Основные особенности и правила составления словарей словосочетаний. 29. Примеры состановления парадигматических отношений между терминами.

Тема 7. Автоматизация процессов обнаружения и исправления ошибок при вводе текстов в ЭВМ. 30. Основные принципы и особенности распознавания орфографических ошибок. 31. Основные принципы и особенности распознавания синтаксических ошибок.

Тема 8. Автоматическое индексирование документов и информационных запросов. 32. Виды классификационных индексов. 33. Правила присвоения классификационных индексов. 34. Особенности создания поисковых образов документов.

Тема 9. Аппаратное и программное обеспечение информационных технологий. 35. Аппаратное обеспечение ЭВМ. 36. Программное обеспечение ЭВМ. 37. Основные языки программирования.

Тема 10. Машинный перевод текстов. 38. Особенности и правила машинного перевода текстов. 39. Правила семантического анализа текста на ИЯ. 40. Плюсы и минусы пословного перевода. 41. Особенности фразового перевода. 42. Характеристики «переводческой памяти».

Примеры практических задач:

«Исследование ПО, использующего методы квантитативной лингвистики»

Цель работы: Закрепление теоретических знаний в вопросах использования существующего ПО для

решения различных задач квантитативной лингвистики.

Задачи работы: Ознакомиться с существующим ПО, разработанным для решения прикладных задач

квантитативной лингвистики. Выбрать один из существующих программных продуктов, изучить его

функционал. Выполнить учебное задание по анализу лингвистической информации с помощью

данного программного продукта.

Задание для работы

1. На основе выбранной задачи изучить технологии ее решения, подготовить данные для обработки.

2. Написать программу, реализующую решение выбранной задачи в автоматическом режиме.

«Анализ лингвистических явлений с точки зрения возможности их математического моделирования».

Цель работы: Проанализировать какое-либо лингвистическое явление и указать сложности, которые

могут возникнуть при его математическом моделировании.

Задачи работы:

1. Ознакомиться с выбранным лингвистическим явлением, решив одну или несколько

лингвистических задач;

2. Разработать методический материал, поясняющий суть явления, на основе решенных задач;

3. Добавить примеры данного явления из русского, английского, других языков;

4. Описать сложности, которые могут возникнуть при математическом моделировании данного

явления;

5. Выступить с докладом на семинаре; руководить работой группы по анализу данного явления и

сложностей его математического моделирования.
«Исследование ПО, использующего методы квантитативной лингвистики»

Цель работы: Закрепление теоретических знаний в вопросах использования существующего ПО для

решения различных задач квантитативной лингвистики.

Задачи работы: Ознакомиться с существующим ПО, разработанным для решения прикладных задач

квантитативной лингвистики. Выбрать один из существующих программных продуктов, изучить его

функционал. Выполнить учебное задание по анализу лингвистической информации с помощью

данного программного продукта.

Задание для работы

1. На основе выбранной задачи изучить технологии ее решения, подготовить данные для обработки.

2. Написать программу, реализующую решение выбранной задачи в автоматическом режиме.

Результаты зачета определяются оценками «зачтено», «не зачтено».

Критерии зачета обусловлены логической демонстрацией приобретенных компетенций в соответствии с текущей программой. Демонстрация предусматривает уверенное использование терминологии, понимание и корректное использование математического аппарата, предусматривает корректность написания кода, его понимание и корректное использование в нем математических методов. Отметка «зачтено» выставляется за счет демонстрации полученных компетенций в практиках, домашних работах и итоговом задании: уверенное владение и понимание работы кода, знание и демонстрация в практике теоретических основ баз данных. Минимальный порог зачета составляет 55 баллов, ниже 55 – «не зачтено»

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=31488

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

в) План семинарских / практических занятий по дисциплине.

1. Поиск в корпусе. Статистические гипотезы
2. Создание корпуса
3. Метрики LL score, коэффициент Жуйана, IPM
4. Работа со звуковым корпусом
5. Обработка звуков
7. Визуализация и анализ данных.

г) Методические указания по проведению лабораторных работ.

Подготовка к проведению лабораторных работ начинается в начале теоретического изложения изучаемой темы и продолжается по ходу её изучения при освоении материала на занятиях в рамках практических заданий и работе над ним в ходе самостоятельной подготовки дома и в библиотеках. Для качественного выполнения лабораторных работ студентам необходимо:

1) повторить теоретический материал по конспекту и учебникам;

2) ознакомиться с описанием лабораторной работы;

3) в специальной тетради для лабораторных работ записать название и номер работы, перечень необходимого программного обеспечения, подготовить алгоритм или код;

4) выяснить цель работы, четко представить себе поставленную задачу и способы её достижения, продумать ожидаемые результаты опытов;

5) ответить устно или письменно на контрольные вопросы по изучаемой теме или решить ряд задач;

6) изучить порядок выполнения лабораторной работы. Подготовить среду выполнения кода к работе. После проверки правильности алгоритма работы программы преподавателем можно начинать выполнение лабораторной работы.

д) Методические указания по организации самостоятельной работы студентов.

Формы самостоятельной работы студентов разнообразны. Они включают в себя:

– изучение и систематизацию практических и теоретических примеров в рамках выполнения текущих заданий по предмету;

– изучение учебной, научной и методической литературы, материалов периодических изданий с привлечением электронных средств официальной, статистической, периодической и научной информации;

– подготовку докладов и презентаций, написание программного кода и его отладка;

– участие в работе студенческих конференций, комплексных научных исследованиях.

Самостоятельная работа приобщает студентов к научному творчеству, поиску и решению актуальных современных проблем.

Примеры самостоятельной работы студентов:

– С помощью программных средств введите речевой сигнал в компьютер и отцифруйте его. При формировании звукового файла обратите внимание на частоту дискретизации (Гц), длительность звучания записываемого сигнала (с/мин), формат хранения (WAV, AU, AIFF, RAW и др.), количество записываемых каналов (моно/стерео), разрешение сигнала (16, 24, 32 бит).

– При выполнении задания необходимо варьировать показатели записи и выявлять различия.

– После оцифровки сигнала приступить к первичному редактированию в программе GoldWave (например, прослушать записанный сигнал, вырезать или вставить какие-либо отрезки звучания, соединить разные файлы и др.), подготовить файл для работы с другими программами.

– Сегментируйте речевой сигнал на фразы, слова, словосочетания, слоги, отдельные звуки.

– С помощью опционалов разных программ установите метки сегментации.

– Отредактируйте сигнал, используя возможности разных программ (например, измените громкость сигнала, частоту дискретизации, поработайте с эффектами, модифицируйте частоту основного тона и др.). В качестве материала для работы используйте данные звуковых хрестоматий.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Грудева, Е. В. Корпусная лингвистика [Электронный ресурс] : учеб. пособие / Е. В. Грудева. -2-е изд., стер. - М. : ФЛИНТА, 2012. - 165 с. http://znanium.com/catalog.php?bookinfo=455049

– Потапова Р.К. Новые информационные технологии и лингвистика : учебное пособие для студентов вузов, обучающихся по специальности 021800 'Теоретическая и прикладная лингвистика' направления 620200 'Лингвистика и новые информационные технологии'

– Баранов А. Н. Введение в прикладную лингвистику. М., 2001. URL: https://dislyget.ru/index.php?r=item/view&id=21065

б) дополнительная литература:

– Потапова ; Моск. гос. лингвист. ун-т .? Изд. 5-е .? Москва : URSS : [ЛИБРОКОМ, 2012] .? 364 с.

–Федотова Е.Л., Федотов А.А. Информационные технологии в науке и образовании : учебное пособие / Е. Л. Федотова, А. А. Федотов .? Москва : ФОРУМ : ИНФРА-М, 2011. 334 с.

–Щипицина, Л. Ю. Информационные технологии в лингвистике [Электронный ресурс] : учеб.пособие / Л. Ю. Щипицина. М. : ФЛИНТА, 2013. 128 с. http://znanium.com/catalog.php?bookinfo=462989

в) ресурсы сети Интернет:

– открытые онлайн-курсы

– Журнал «Эксперт» - http://www.expert.ru

– Официальный сайт Федеральной службы государственной статистики РФ - www.gsk.ru

– Официальный сайт Всемирного банка - www.worldbank.org

– Общероссийская Сеть КонсультантПлюс Справочная правовая система. http://www.consultant.ru

– Официальный сайт языка программирования R - www.r-cran.com

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).

– язык программирования R (RStudio) и Python;

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index

– ЭБС Лань – http://e.lanbook.com/

– ЭБС Консультант студента – http://www.studentlibrary.ru/

– Образовательная платформа Юрайт – https://urait.ru/

– ЭБС ZNANIUM.com – https://znanium.com/

– ЭБС IPRbooks – http://www.iprbookshop.ru/

в) профессиональные базы данных *(при наличии)*:

– Университетская информационная система РОССИЯ – https://uisrussia.msu.ru/

– Единая межведомственная информационно-статистическая система (ЕМИСС) – https://www.fedstat.ru/

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

Лаборатории, оборудованные компьютерами (не ниже i3, RAM 8Gb), проектором

Аудитории для проведения занятий лекционного и семинарского типа индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации в смешенном формате («Актру»).

15. Информация о разработчиках

Степаненко Андрей Александрович, НИ Томский государственный университет, ассистент кафедры общей, компьютерной и когнитивной лингвистики ТГУ