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

Филологический факультет

УТВЕРЖДЕНО:
Декан
И. В.Тубалова

Рабочая программа дисциплины

Технологии обработки звучащей речи

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 дисциплины является изучение основных принципов и методов автоматической обработки текстов на естественном языке (ЕЯ)

Целью освоения дисциплины является формирование следующих компетенций:

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

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 7, экзамен.

5. Входные требования для освоения дисциплины

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

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: «Введение в языкознание», «Общая фонетика», «Общая морфология», «Общий синтаксис», «Общая семантика», «Информационные технологии и основы информационной культуры в лингвистике», «Информатика и основы программирования», «Квантитативные методы лингвистики», «Вероятностные модели», «Лингвистические базы данных».

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:
– лекции: 12 ч.;
– семинарские занятия: 0 ч.
– практические занятия: 22 ч.;
– лабораторные работы: 0 ч.
в том числе практическая подготовка: 22 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины (модуля), структурированное по темам

Тема 1. Автоматическая обработка устной речи

Цель и задачи автоматической обработки устной речи. Метод Байеса. Архитектура систем автоматической обработки текстов

Тема 2. Автоматы, формальные грамматики и языки

Лингвистический автомат. Уровневое построение систем АОТ и ЛА. Подблок опознания формата текста и его частей, а также определение их жанровой и тематической принадлежности

Тема 3. Морфологический анализ в системах автоматической обработки текста

Использование и запуск морфологического анализатора mystem в языке программирования R. Квантитативный анализ частей речи, их визуализация и анализ

Тема 4. Синтаксический анализ в системах автоматической обработки текста

Использование и запуск морфологического анализатора udpipe в языке программирования R. Квантитативный анализ частей речи, их визуализация и анализ

Тема 5. Семантический анализ в системах автоматического анализа текста

Формальные грамматики, извлечение сущностей из текста (нейронные сети и/или формальныеграмматики). Tomita parser, Spacy

Тема 6. Словарная поддержка. Типы словарей. Компьютерные (электронные) словари.

Создание тематических словарей для классификации текстов (sentiment analysis)

Тема 7. Синтез текстов на естественном языке

Понятие нейронной сети, принципы работы. Искусственные нейронные сети: keras, ruGPT-3, seq2seq

Тема 8. Морфологическая разметка корпусов текстов. Корпус русского языка

Принципы разметки, виды и типы морфологических теггеров.

Тема 9. Автоматическая обработка данных в корпусах русского языка и BNC

Поиск и сравнение лексем в корпусах, метрики сравнения: IPM, TF-IDF, LL-score, коэффициент Жуйана.

Тема 10. Итоговая презентация проекта

9. Текущий контроль по дисциплине

Текущий контроль образовательной программы (темы, раздела, модуля) требованиям образовательных стандартов по направлениям подготовки/специальностям. 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:

Примерные тестовые задания по 1 модулю

1. С работой каких органов связано деление согласных на глухие и звонкие?

А. Языка,

Б. Голосовых связок

В. твердого неба,

Г. альвеол.

2. Сила (интенсивность) звука зависит от:

А. Частоты колебаний,

Б. Протяженности его во времени,

В. амплитуды, или размаха, колебаний.

Г. громкости.

3.Определите гласный звук русского языка по следующим признакам: переднего ряда, среднего подъема, нелабиализованный.

А.Э,

Б. А,

В. О,
Г. Ы.
4. Укажите язык вокалический:
А. Русский,
Б. Немецкий,
В. французский,
Г. Английский.
5.Найдите полную регрессивную контактную ассимиляцию:
А. Ланно из ладно,
Б. В чечении из теченбии,
В. дважды из дватды,
Г. балалайка из балабайка.
6. Укажите регрессивную контактную диссимиляцию:
А. Февраль из феврарь,
Б. Мести из метти,
В.Пролубь из прорубь,
Г. ярмарка из ярманка.
7. Укажите гаплологию:
А. Табакур из табакокур,
Б. Наждак из наджак,
В. ндрав из нрав,
Г. вострый из острый.
8. Укажите эпентезу:
А. Знаменосец из знаменоносец,
Б. Ларивон из Ларион,
В. вена (укр.) из она (рус),
Г. вузол (укр.) из узел (рус).
9. С работой каких органов связано деление согласных на твердые и мягкие?
А. Голосовых связок,
Б. Альвеол,
В. Средней части спинки языка и твердого неба
Г. кончика языка и альвеол.
10. Частота колебаний за единицу времени определяет:
А.Силу звука,
Б. Высоту звука,
В.громкость звука,
Г. длительность звука.
11. Какие фонетические процессы наблюдаются в словах типа: сжатый, просьба?
А. Диссимиляция,
Б. Аккомодация,
В. ассимиляция,
Г. метатеза.

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 сформированность следующих компетенций: ПК-4, ИПК-4.1.

Примерный перечень практических вопросов

1. Дайте определение направлению “акустическая теория речеобразования”

2. Какие типы звуков (с точки зрения АТР) существуют? В чем их отличительная черта?

3. Дайте определение форманты. Какие типы звуков можно отнести к данному явлению?

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»

Экзамен проводится в седьмом семестре в письменной форме по билетам. Экзаменационный билет состоит из двух частей. Продолжительность экзамена 1,5 часа.

Первая часть представляет собой тест из 2 вопросов, проверяющих ПК-4. Ответы на вопросы первой части даются путем выбора из списка предложенных.

Вторая часть содержит один практический вопрос, проверяющий ПК-4, ИПК-4.1. и предполагает решение задач и краткую развернутую интерпретацию полученных результатов.

Примерный перечень теоретических вопросов

1. Перечислить направления компьютерной лингвистики.

2. Сформулировать общие принципы построения автоматизированных систем обработки звучащей речи.

3. Разъяснить принципы работы системы TTS.

4. Перечислить принципы и методы работы STT.

Пример практического задания:

Напишите функцию сегментирования и декодирования звучащей речи:

```
from pydub import AudioSegment
import os

def mp3_to_wav(source, skip=0, excerpt=False):

    sound = AudioSegment.from_mp3(source) # load source
    sound = sound.set_channels(1) # mono
    sound = sound.set_frame_rate(16000) # 16000Hz

    if excerpt:
        excrept = sound[skip*1000:skip*1000+60000] # 30 seconds - Does
not work anymore when using skip
        output_path = os.path.splitext(source)[0]+"_excerpt.wav"
        excrept.export(output_path, format="wav")
    else:
        audio = sound[skip*1000:]
        output_path = os.path.splitext(source)[0]+".wav"
        audio.export(output_path, format="wav")
```

```
        return output_path
```

Критерии экзамена 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 «хорошо» выставляется за счет демонстрации полученных компетенций, владение и понимание кода, теоретических аспектов его применения в практике работы с текстовыми массивами данных допускаются недочеты в понятийном аппарате математики. Отметка «удовлетворительно» позволяет допустить ошибки в разработке кода, но учитывает последовательную логику изложения структуры кода, его интерпретацию, связь теоретических аспектов лингвистики и математики, демонстрация понимания хода обработки текста. Минимальный порог оценки «отлично» составляет 90-100 баллов, хорошо 75-89, удовлетворительно «55-74» ниже 55 – «неудовлетворительно»

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=12998

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

в) План семинарских / практических занятий по дисциплине.

Семинар №1

1. Объект, цели и методы исследования дисциплины.
2. Устройство произносительного аппарата человека.
3. Устройство слуховой системы человека.
4. Артикуляторные методы исследования

Семинар №2

1. Артикуляторный аспект фонетики.

Классификация гласных и согласных звуков.

2. Акустический аспект фонетики:

1) Частота колебаний и высота звуков

2) Сила и громкость звук а

3) Спектр звука и тембр

Общие параметры частоты основного тона

Частотный диапазон

Частотный интервал

Скорость изменения ЧОТ

Частные параметры ЧОТ

Семинар №3

1. Артикуляторный аппарат человека и его роль в образовании звуковой волны.
2. Способы получения спектров.
3. Общие сведения о спектральной структуре звуков речи:

Гласные звуки

Согласные звуки

4.Осциллографический анализ речи.

5. Звуки речи на осциллограмме.

Временные характеристики речи

Физиологические временные константы

Физический коррелят временной характеристики

–длительность

Относительные значения длительности
Феномен паузации
Перцептивные корреляты временной характеристики.
Дополнительные вопросы
Дайте определение форманты.
Какие согласные образуются с импульсным/турбулентным источником звука?
Перечислите основные методы акустического
анализа звуков речи.
Инструментальный анализ фраз при помощи компьютерной программы
анализа звучащей речи
VoiceScan
Семинар №4
1. Фонологический и фонетическ ий слог.
2. Фонетическая структура слога.
3. Теории слога:
1) Акустические теории слога
2) Артикуляторные теории слога
3) Три вида критериев при слогоделении:
a. Правила
b. Интуиция
c. Механизмы
4) Восприятие просодических характеристик речи
Дополнительные вопросы
Определение слога как произносительной единицы
Определение слога с точки зрения акустических теорий слога.
Сравните слогоделение в английском и русском языках.
Семинар №6
1. Слово как основная единица.
2. Фонемный состав слова
3. Акцентно-ритмическая структура слова
4. Инструментальный анализ фраз.
5. Сегментация интонограмм
Семинар №7
1. Синтагма как интонационная единица.
2. Основные фонетические средства интонации:
Мелодика
Темп речи
Интенсивность звука
3. Акустический аналог перцептивных характеристик:
Частота основного тона
Скорость речи
Сила звукового давления
4. Интонационные типы и их классификация
5. Способы связи синтагмы внутри одного высказывания.
6. Инструментальный анализ фраз.
Семинар №8
1. Произносительная норма английского языка.
2. Соотношение языковой нормы и системы.
3. Кодификация языковой нормы.
4. Инструментальный анализ обработки текста и звучащей речи.
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:

1. Запустите программу PRAAT, сделайте анализ звуков по их типу (голосовой, турбулентный, импульсный)

2. Сравните слово/предложение носителя языка со своим. Найдите интерференционные признаки в речи

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