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

САЕ «Институт человек цифровой эпохи».

УТВЕРЖДАЮ:
Руководитель ОПОП:
_____ З.И. Резанова
« 31 » августа 2022 г.

Рабочая программа дисциплины

Базы данных

по направлению подготовки

45.04.03 Фундаментальная и прикладная лингвистика

Направленность (профиль) подготовки :
Компьютерная и когнитивная лингвистика

Форма обучения
Очная

Квалификация
Магистр

Год приема
2022

Код дисциплины в учебном плане: Б1.В.ДВ.1.1.5

СОГЛАСОВАНО:
Руководитель ОПОП
_____З.И. Резанова

Председатель УМК
_____Ю.А. Тихомирова

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины

В результате курса формируются следующие компетенции:

УК-2 Способен управлять проектом на всех этапах его жизненного цикла

ПК-3 Способность разрабатывать системы автоматической обработки звучащей речи и письменного текста на естественном языке, лингвистические компоненты электронных ресурсов и интеллектуальных электронных систем (лингвистические корпуса, словари, онтологии, базы данных)

ПК-4 Способность разрабатывать проекты прикладной направленности в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-4.3 Обеспечивает выполнение проекта в области когнитивной и компьютерной лингвистики с применением современных технических средств и информационных технологий, в том числе в области искусственного интеллекта, в соответствии с установленными целями, сроками и затратами..

ИПК-4.2 Разрабатывает программу действий по решению задач проекта в области когнитивной и компьютерной лингвистики с учетом имеющихся технических средств и информационных технологий, в том числе в области искусственного интеллекта..

ИПК-4.1 Формулирует цель проекта прикладной направленности в области когнитивной и компьютерной лингвистики, обосновывает необходимость применения современных технических средств и информационных технологий, в том числе в области искусственного интеллекта.

ИПК-3.3 Разрабатывает лингвистические компоненты интеллектуальных информационных систем (онтологии, базы данных).

ИУК-2.3 Обеспечивает выполнение проекта в соответствии с установленными целями, сроками и затратами.

2. Задачи освоения дисциплины

Задача дисциплины направлена на изучение проектирования и использования баз данных, формирование навыков проектирования и разработки прикладных лингвистических проектов с использованием современных СУБД, систематизирует знания о способах анализа, верификации и оценки полноты информации в ходе профессиональной деятельности

– Проектирование и разработка реляционных баз данных
– Создание и анализ er-диаграмм
– Создание SQL-запросов: создание, изменение данных, манипуляция данными
– Разработка лингвистических приложений, реализуемых в СУБД PostgreSQL
– Бекенд: создание запросов, вывод данных через html, javascript, flask

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к обязательной части образовательной программы.

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, является обязательной для изучения. Дисциплина входит в модуль Компьютерная лингвистика.

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

Второй семестр, зачет
Третий семестр, зачет

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».

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 5 з.е., 180 часов, из которых:
-лекции: 12 ч.
-практические занятия: 62 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Базы данных и их классификация

Тема 2. Основные понятия, связанные с лингвистическими информационными ресурсами. Классификация СБД и ИС.

Тема 3. Общие сведения о реляционной модели данных (РМД)

Тема 4. Структурная и целостная части РМД

Тема 5. Манипуляционная часть РМД

Тема 6. Язык структурированных запросов (SQL). DDL, DML

Тема 7. Оптимизация плана выполнения запросов. Индексирование данных

Тема 8. Проектирование БД. Обзор нотаций описания БД. CASE системы

Тема 9. Разработка хранимых функций, процедур, триггеров

Тема 10. Сравнение технологий доступа к данным.

Тема 11. Технологии клиент-сервер. Понятия тонкого и толстого клиентов.

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:

– Продумайте концепцию вашей базы данных, это может быть словарь, корпус, база респондентов и т.п.

– Создайте ER-диаграмму вашей базы

– При помощи СУБД создайте базу данных:

```
CREATE DATABASE [Cook_demo]
GO

ALTER DATABASE [Cook_demo]
  SET
    ANSI_NULL_DEFAULT OFF,
    ANSI_NULLS OFF,
    ANSI_PADDING OFF,
    ANSI_WARNINGS OFF,
```

```
    ARITHABORT OFF,
    AUTO_CLOSE OFF,
    AUTO_CREATE_STATISTICS ON,
    AUTO_SHRINK OFF,
    AUTO_UPDATE_STATISTICS ON,
    AUTO_UPDATE_STATISTICS_ASYNC OFF,
    COMPATIBILITY_LEVEL = 130,
    CONCAT_NULL_YIELDS_NULL OFF,
    CURSOR_CLOSE_ON_COMMIT OFF,
    CURSOR_DEFAULT GLOBAL,
    DATE_CORRELATION_OPTIMIZATION OFF,
    DB_CHAINING OFF,
    HONOR_BROKER_PRIORITY OFF,
    MULTI_USER,
    NESTED_TRIGGERS = ON,
    NUMERIC_ROUNDABORT OFF,
    PAGE_VERIFY CHECKSUM,
    PARAMETERIZATION SIMPLE,
    QUOTED_IDENTIFIER OFF,
    READ_COMMITTED_SNAPSHOT OFF,
    RECOVERY SIMPLE,
    RECURSIVE_TRIGGERS OFF,
    TRANSFORM_NOISE_WORDS = OFF,
    TRUSTWORTHY OFF
    WITH ROLLBACK IMMEDIATE
GO

ALTER DATABASE [Cook_demo]
  COLLATE Cyrillic_General_CI_AS
GO

ALTER DATABASE [Cook_demo]
  SET DISABLE_BROKER
GO
```

Создайте свою базу данных

– Заполните существующую базу и добавьте свои данные:

```
CREATE TABLE [dbo].[Автор рецепта] (
  [ID_Автора] [int] IDENTITY,
  [Фамилия] [varchar](50) NOT NULL,
  [Имя] [varchar](50) NOT NULL,
  [Отчество] [varchar](50) NULL,
  [Пол] [char](1) NOT NULL,
  [Дата рождения] [date] NULL,
  [ФИО] AS (
    CONCAT(
      [Фамилия], ' ',
      LEFT([Имя],1), '.',
      ' '+LEFT([Отчество],1)+'.'
    )
  ),
  CONSTRAINT [PK_Автор рецепта] PRIMARY KEY CLUSTERED ([ID_Автора]) WITH (FILLFACTOR = 100),
```

```
  CONSTRAINT [CKC_Автор рецепта - Дата рождения] CHECK(
    [Дата рождения]>='1700.01.01' AND [Дата рождения]<=GetDate()
  ),
  CONSTRAINT [CKC_Автор рецепта - Пол] CHECK (
    [Пол] IN ('Ж', 'М')
  )
)
ON [PRIMARY]
```

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 овладения компетенциями ИПК-4.3, ИПК-4.2, ИПК-4.1, ИПК-3.3, ИУК-2.3

Зачет по дисциплине принимается на основе достижения рубежных показателей в рейтинге (не ниже 55 баллов), при выполнении практических заданий, тестов, посещения занятий.

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»

Рейтинг, баллы

1 – присутствие на лекции

1 – присутствие на занятии

1-3 – работа на занятии

1-36 – подготовка к занятию и работа на практическом занятии (в т.ч. д/з)

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

Примерный перечень теоретических вопросов

1. Понятие информационной системы, БД и их классификация.
2. Определение системы баз данных (СБД) и её назначение.
3. Основные этапы проектирования БД.
4. Трехуровневая архитектура БД.
5. Доступ к данным в трехуровневой архитектуре.
6. Моделирование предметной области. Модель сущность-связь: основные понятия и методы. Этапы моделирования Назначение модели. Свойства связей.
7. Графические нотации представления ER модели данных.
8. Понятие РМД. Основные концепции и термины. Фундаментальные свойства отношений. Понятие потенциального, первичного и альтернативного ключей.
9. Структурная часть реляционной модели данных (РМД).
10. Целостностная часть РМД. Виды ограничений целостности. Возможный и первичный ключи отношений, внешние ключи.
11. Манипуляционная часть РМД. Эквивалентность абстрактных реляционных языков.
12. Реляционная алгебра. Операции объединения, пересечения, разности, произведения, присвоения.
13. Реляционная алгебра. Операции выборки, создания проекций, деления.

14.Реляционная алгебра. Операция соединения (естественное соединение, тета-соединение, внешнее соединение).

15.Язык SQL. Структура запроса на выборку. Команды SELECT, FROM, WHERE. Использование операторов сравнения, логических операторов, операторов IN, BETWEEN, LIKE в команде WHERE.

16.Язык SQL. Структура запроса на выборку. Команда SELECT. Исключение избыточных данных в результирующих отношениях.

17.Язык SQL. Структура запроса на выборку. Упорядочивание выходных результатов.

18.Язык SQL. Структура запроса на выборку. Группировка данных: предложения GROUP BY и HAVING.

19.Язык SQL. Организация многотабличных запросов: естественное соединение, тета-соединение, внешнее соединение, соединение таблицы с самой собой.

20.Язык SQL. Структура запросов с подзапросами. Некоррелированные подзапросы. Использование DISTINCT, IN и агрегатных функций в подзапросах.

21.Структура запросов с подзапросами. Коррелированные подзапросы. Сравнение коррелированных подзапросов и запросов на соединение.

22.Язык SQL. Комбинирование результирующих таблиц. Создание запросов на объединение, пересечение и разность.

23.Язык SQL. Операторы языка манипулирования данными: DELETE, UPDATE, INSERT.

24.Язык SQL. Средства определения схемы базы данных. Общая структура, этапы определения таблицы, определение столбцов.

25.Язык SQL. Средства определения схемы базы данных. Общая структура, этапы определения таблицы, ограничительные условия на таблицу.

26.Операция соединения отношений. Примеры с использованием реляционной алгебры и решения с использованием средств языка SQL.

Примеры практических задач:

2. С помощью языка SQL разработать запрос согласно ER модели с использованием ограничения.

3. С помощью языка SQL разработать запрос согласно ER модели с использованием группировки.

4. С помощью языка SQL разработать запрос согласно ER модели с использованием соединения.

5. С помощью языка SQL разработать запрос на модификацию таблицы

6. С помощью языка SQL разработать запрос на удаление данных из таблицы.

7. С помощью средств реляционной алгебры составить запрос согласно ER модели.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=14690

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

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

в) План семинарских / практических занятий по дисциплине.

Тема 1. Введение в информатику. Понятие и представление информации. Операционные системы. Windows, Linux. Прикладные разделы компьютерной лингвистики.

Тема 2. Фонетический уровень языка. Основы акустической теории речеобразования. Типологизация звуков АТР. Фонетическая разметка в программе PRAAT. Графическое представление звуков: спектрограмма, осциллограмма. Частота общего фона, форманты. Анализ и синтез звучащей речи

Тема 3. Компьютерная морфология. Анализ морфологии на основе правил. Морфологический анализатор Pymorphy2, mystem, AOT, проект mystem+. Статистические методы анализа слов

Тема 4. Извлечение информации из неструктурированного текстового массива данных. Распознавание сущностей: Natasha, Tomita-parser. Формальные грамматики. Распознавание отношений

Тема 5. Компьютерный синтаксис. Современные подходы к анализу синтаксических структур. Современные синтаксические анализаторы. Лингвистический процессор ЭТАП. Stanford NLP, RASP, OpenNLP, NLTK

г) 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:

Скачайте текст, содержащий в себе неологизмы:

- проведите морфологический анализ лексем при помощи программы mystem
- создайте словарь неопределенных неологизмов в программе mystem

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Степанов А.Н. Информатика: учебник для вузов / А.Н. Степанов. – СПб.: Питер, 2015 – 720 с.

– Jurafsky Daniel, James H. Martin. Speech and Language Processing. / An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. Upper Saddle River, NJ, 2009. https://www.cs.colorado.edu/~martin/slp2.html

– Николаев И.С. / Прикладная и компьютерная лингвистика. Изд. 2 URSS. 2017. 320 с. ISBN 978-5-9710-4633-2

б) дополнительная литература:

– Щипицина Л. Информационные технологии в лингвистике: учеб. пособие / Л. Щипицина. – М.: Флинта, 2015. – 128 с.

– Кодзасов С.В. Алгоритмы преобразования русских орфографических текстов в фонетическую запись / С.В. Кодзасов М.: МГУ, 1970. 130 с/

– Коваль С. А. Лингвистические проблемы компьютерной морфологии. СПб., 2005. Леонтьева Н. Н. Автоматическое понимание текстов. Системы, модели, ресурсы. М., 2006. Ляшевская О. Н. и др. Оценка методов автоматического анализа текста: морфологические парсеры русского языка. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог–2010». Вып. 9(16). М., 2010.

– Шаров С. А., Беликов В. И., Копылов Н. Ю., Сорокин А. А., Шаврина Т. О. Корпус с автоматически снятой морфологической неоднозначностью: К методике лингвистических

исследований. Компьютерная лингвистика и интеллектуальные технологии. // Диалог. М., 2015. http://www.dialog-21.ru/digests/dialog2015/materials/pdf/SharoffSAetal.pdf

в) ресурсы сети Интернет:

– Система ЭТАП-3: http://proling.iitp.ru/ru/etap3
– Синтаксический анализатор АОТ: http://aot.ru/demo/synt.html
– FrameNet https://framenet.icsi.berkeley.edu
– Stanford NLP http://nlp.stanford.edu:8080/corenlp/process
– Парсер RASP в составе системы Gate: последняя версия доступна по ссылке http://ilexir.co.uk/applications/rasp/download/
– OpenNLP: https://opennlp.apache.org
– Link Grammar Parser: http://slashzone.ru/parser/parse.pl
– Лингвистический пакет NLTK: http://www.nltk.org/install.html
– AIIRE http://aiire.org, http://svn.aiire.org/repos/t
– Томита-парсер: https://yandex.ru/dev/tomita

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, Windows 7-10;
- публично доступные облачные технологии (Google Docs, Яндекс диск и т.п.).
- язык программирования R (RStudio) и Python;
- Программа Mystem.

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
- в) профессиональные базы данных:
- Университетская информационная система РОССИЯ – https://uisrussia.msu.ru/

– Единая межведомственная информационно-статистическая система (ЕМИСС) –

– Национальный корпус русского языка [Электронный ресурс]. URL: https://ruscorpora.ru/

Institute of Formal and Applied Linguistics [Электронный ресурс]. URL: http://ufal.mff.cuni.cz/udpipe

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

Дацюк Валерий Валентинович, НИ Томский государственный университет, старший преподаватель кафедры общей, компьютерной и когнитивной лингвистики