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

Язык программирования R

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.1.1.3

СОГЛАСОВАНО:
Руководитель ОПОП
_____З.И. Резанова

Председатель УМК
_____Ю.А. Тихомирова

Томск – 2022

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

– ОПК-6 способность осуществлять эффективное управление разработкой программных средств и информационных проектов в сфере своей профессиональной деятельности

– ОПК-3 способность выбирать оптимальные подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий

– ПК-2 способность самостоятельно планировать и проводить научные эксперименты (в том числе, при наличии подобного оборудования, с использованием высокоточных методов регистрации мозговой активности и движений глаз) ПК

– 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).

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

– ИОПК-3.2 Критически сопоставляет и оценивает существующие подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологий

– ИОПК-6.1 Аргументированно выбирает математические и лингвистические методы решения профессиональных задач с применением языков программирования

– ИОПК-6.2 Разрабатывает алгоритмы и программы для решения лингвистических и междисциплинарных задач в том числе с применением высокопроизводительных вычислительных технологий

– ИОПК-6.3 Разрабатывает и отлаживает программный код, направленный на решение лингвистических и междисциплинарных задач с применением современных языков программирования

– ИПК-2.1 Разрабатывает дизайн эксперимента, формирует стимульный материал в соответствии с целями исследования

– ИПК-3.1 Разрабатывает системы автоматической обработки звучащей речи и письменного текста на естественном языке.

2. Задачи освоения дисциплины

– Изучить основы программирования на языке R.

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

– Освоить основные задачи и методы сбора, структуризации текстового массива данных, их векторизацию пред и пост обработку.

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

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

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор. Дисциплина входит в модуль «Компьютерная лингвистика».

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

Второй 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 анализ естественного языка (NLP)», «Статистические методы в гуманитарных исследованиях», «Основные направления лингвистического обеспечения новых инф. технологий», «Лингвистика в контексте современного гуманитарного и естественнонаучного знания».

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 2 з.е., 72 часов, из которых:
-лекции: 6 ч.
-практические занятия: 24 ч.
в том числе практическая подготовка: 0 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Основы языка программирования R

Синтаксис, объекты, классы, переменные, структуры данных.

Тема 2. Сложные типы данных и работа с ними. Вектор, матрица, массив

Представление структур данных (переменных) в среде R: вектор, матрица, датафрейм, лист.

Тема 3. Управляющие структуры. Условные операторы, циклы, функции

Основные типы управляющих структур, их применение в разных структурах, синтаксис и алгоритм написания

Тема 4. Парсинг и структуризация текстовых данных с помощью языка программирования R

Парсинг web-страниц, основные положения, логика посика и структуризации информации. Библиотека rvest. Извлечение информации через API (Библиотека Rcurl)

Тема 5. Препроцессинг текстовых массивов: токенизация, лемматизация, единый регистр, удалении «шума»

Удаление стоп-слов, лемматизация текстов с помощью стеммеров (mystem)

Тема 6. Словарная поддержка. Типы словарей. Создание словарей

Создание тематических словарей для классификации текстов (sentiment analysis, topic modeling)

Тема 7. Визуализация текстовых данных в R.

Частотный анализ, построение гистограмм,

Тема 8. Автоматический анализ частей речи в библиотеки UdPipe

Принципы разметки, виды и типы морфологических теггеров.

Тема 9. Описательная статистика

Поиск и сравнение лексем в корпусах, метрики сравнения: IPM, TF-IDF, LL-score, коэффициент Жуйана. Базовая статистика, боксплот, тип распределения, корреляции

Тема 10. Разработка чат-бота

Принципы и методы создания час ботов. Разработка чат бота для ПО Telegram

Тема 11. 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 2 модулю:

#1. Дан вектор целых чисел. Исключить из него a) максимальный b) минимальный элемент.

vec <- c(2,60,4,10)

#2. Дан вектор целых чисел, в котором есть два нулевых элемента. Исключить нулевые элементы.

vec <- c(3,0,7,0,0,3)

#3. Дан вектор X целых чисел и целое число b. Исключить из вектора элементы, равные b.

#4. Дан вектор целых чисел и числа A1, A2 и A3. Включить эти числа в массив, расположив их после второго элемента.

#6. Вывести все элементы вектора, стоящие до максимального элемента

Дан вектор из 20-ти чисел и число A. Вычислить сумму тех отрицательных элементов вектора, значения которых больше, чем A. Подсчитать также количество таких элементов.

#7. Дан вектор из 10-ти чисел. Вычислить среднее арифметическое положительных элементов этого вектора и среднее арифметическое отрицательных элементов этого вектора

#8. Исключить из вектора элементы, расположенные между максимальным и минимальным.

#9. Дан вектор vec = c(2,3,8,7,4). Найдите пары чисел, сумма которых равна 5 (s=5). (Учитываются несколько алгоритмов решения задачи). Включая варианты комбинаторного сравнения элементов вектора, работы с памятью и циклами, бинарный подход.

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. Итоговый проект представляет собой парсинг, препроцессинг и первичный анализ массива текстов.

1. Составьте словарь слов для поиска в корпусе
2. Определите тип распределения
3. Постройте корреляционный анализ

Измените структуру кода для своих данных:

```
# install.packages("rvest") – установка библиотеки
library(rvest)
##парсинг новостных текстов
webpage <- read_html("https://news.vtomske.ru/c/tomsk?up=1635835980")
results <- webpage %>% html_nodes(".news-small") %>% html_attr("href")
page_start <- "https://news.vtomske.ru"
#функции парсинга элементов странцы
scarp_url <- function(url){
  url <- read_html("https://news.vtomske.ru/c/tomsk?up=1637118900")
  results <- url %>% html_nodes(".news-small") %>% html_attr("href")
```

```
  return(results)
}
scarp_text <- function(url){
  txt_link = read_html(url)
  text = txt_link %>% html_nodes(".full-text") %>% html_text()
  return(text)
}
scarp_head <- function(url){
  txt_link = read_html(url)
  text = txt_link %>% html_nodes(".material-title") %>% html_text()
  return(text)
}
scarp_date <- function(url){
  txt_link = read_html(url)
  text = txt_link %>% html_nodes(".info") %>% html_text()
  return(text)
}
#создание структуры DataFrame для хранения новостей и метаинформации
df <- data.frame(bodyNws = NA,
                 titleNws = NA,
                 dateNws=NA)
#цикл парсинга новости
i=1
for (i in 1:length(link)) {
  bodyNws = scarp_text(link[i])
  titleNws <- scarp_head(link[i])
  dateNws <- scarp_date(link[i])
  df <- rbind(df, cbind(titleNws,bodyNws,dateNws))
}
df[2,3]
#изменение даты публикации новости (дата системы)
grepl("Сегодня", df[4,3])
Sys.Date()
df$bodyNws[2]
df$bodyNws <- gsub("Дмитрий Кандинский / vtomske.ru",
    "", df$bodyNws)
i=1
for (i in 1:length(df[,3])) {
  grepl("Сегодня", df[i,3])
  if (grepl("Сегодня", df[i,3])){
    df[i,3] <- gsub("Сегодня",
        Sys.Date(), df[i,3])
  }else if(grepl("Вчера", df[i,3])){
    df[i,3] <- gsub("Вчера",
                    Sys.Date()-1, df[i,3])
  }
}

i=1
for (i in 1:length(df$dateNws)) {
  grepl("Сегодня", df$dateNws[i])
  if (grepl("Сегодня", df$dateNws[i])){
```

```
   df[i,3] <- gsub("Сегодня",
                   Sys.Date(), df$dateNws[i])
  }else if(grepl("Вчера", df$dateNws[i])){
   df[i,3] <- gsub("Вчера",
                   Sys.Date()-1, df$dateNws[i])
  }
}


i=1
for (i in 1:length(df$titleNws)) {
  if (grepl("Сегодня", df$dateNws[i])){
    df$dateNws[i] <- gsub("Сегодня", Sys.Date(), df$dateNws[i])

  }
}
i=1
for (i in 1:length(df$titleNws)) {
  if (grepl("Вчера", df$dateNws[i])){
    df$dateNws[i] <- gsub("Вчера", Sys.Date()-1, df$dateNws[i])

  }
}
grepl("Вчера", df$dateNws[5])
gsub("Вчера", Sys.Date()-1, df$dateNws[5])
if ()
write.csv(df, "vtomske2021.csv") #сохранение структурированных данных
```

- Напишите парсер для своего источника. При написании парсера следует обратить особое внимание на метаинформацию: дата создания, класс (оценка, количество просмотров, рубрика, автор, дата создания и пр.).
- Проведите лемматизацию скаченных текстов в структуре DataFrame при помощи программы Mystem

Выполнение этапов проекта демонстрирует овладение ИОПК-3.2, ИОПК-6.1, ИОПК-6.2, ИОПК-6.3, ИПК-2.1, ИПК-3.1 .

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

Критерии оценки выполнения заданий и парт на «зачтено»:

1. Программный код имеет структуру, грамотный синтаксис, комментарии.
2. Осуществляется дебаггинг и рефакторринг кода, позволяющий выявить уровень понимания решения задачи.
3. Все блоки выполнены в едином стиле, имеют систематизацию в виде итогового проекта (портфолио).
4. Осуществляется интеграция математических и лингвистических методов обработки естественного языка.

Условия получения зачета по курсу:

1. Посещение не менее 50% аудиторных занятий.
2. Выполнение не менее 50% заданий, предусмотренных по курсу.
3. Получение за презентацию итогового задания (портфолио).

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=14669

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

в) План семинарских / практических занятий по дисциплине.

Семинар №1

1. Основы языка программирования R
2. Переменные, структуры данных
3. Циклы, проверка условий

Семинар №2

1. Сбор и структуризация текстовых данных
2. Работа с файлами
3. Лемматизация текстов при помощи mystem

Семинар №3

1. Библиотека quanteda для векторизации и анализа текстовых массивов данных
2. Векторизация текстов. Принципы, методы
3. Составление словарей, n-грамы

Семинар №4

1. Статистический анализ матрицы слов текста
2. Описательная статистика
3. Корреляционный анализ
4. Проверка статистических гипотез
5. Кластерный анализ

Семинар №6

1. Сокращение пространства признаков
2. Визуализация и анализ текстовых данных

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.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Кабаков Р. R в действии. Анализ и визуализация данных на языке R / Роберт И. Кабаков Р., – М.: ДМК, 2016. – 587 с. [Электронный ресурс]: https://ez.lib.tsu.ru/login?url=https://e.lanbook.com/book/58703

– Шипунов А. Б. Наглядная статистика. Используем R. / А. Б. Шипунов, Е. М. Балдин, П. А. Волкова, А .И. Коробейников, С. А. Назарова, С. В. Петров, В. Г. Суфиянов, – М.: ДМК, 2017. – 296 с. [Электронный ресурс]: https://ez.lib.tsu.ru/login?url=https://e.lanbook.com/book/50572

– Мастицкий С.Э. Статистический анализ и визуализация данных с помощью R / Мастицкий С.Э., Шитиков В.К., - М.: ДМК, 2015. – 495с. [Электронный ресурс]: https://ez.lib.tsu.ru/login?url=https://e.lanbook.com/book/73072

б) дополнительная литература:

– Thomas Rahlf. Data Visualisation with R. Springer International Publishing, New York, 2017. ISBN 978-3-319-49750-1. [Электронный ресурс]: https://link.springer.com/content/pdf/bfm%3A978-3-319-49751-8%2F1.pdf?error=cookies_not_supported&code=3fd6d6f9-b6fd-478f-87b6-f2ecbb060f8c

– Matthias Kohl. Introduction to statistical data analysis with R. bookboon.com, London, 2015. ISBN 978-87-403-1123-5. [Электронный ресурс]: https://www.arma.org.au/wp-content/uploads/2017/03/introduction-to-statistical-data-analysis-with-r.pdf

– Torsten Hothorn and Brian S. Everitt. A Handbook of Statistical Analyses Using R. Chapman & Hall/CRC Press, Boca Raton, Florida, USA, 3rd edition, 2014. ISBN 978-1-4822-0458-2. [Электронный ресурс]: http://www.ecostat.unical.it/tarsitano/Didattica/LabStat2/Everitt.pdf

–JurafskyD., Martin J. Speech and Language Processing.An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.Prentice Hall, 2000 [Электронный ресурс]: https://home.cs.colorado.edu/~martin/slp.html

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, 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) – https://www.fedstat.ru/

– Справка ПО и библиотек R-CRAN https://cran.r-project.org/

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 смешанном формате («Актру»).

15. Информация о разработчиках

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