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

Text Mining с применением 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.1.1.7

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:

– ПК-3 способен 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)

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

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

– ПК-4 способен 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:

ИОПК-3.3 Способен решать конкретные научные и прикладные задачи в области лингвистики и информационных технологий на основе самостоятельного выбора оптимальных подходов и методов их решения

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

ИПК-3.2 Разрабатывает лингвистические компоненты электронных ресурсов (лингвистические корпуса, словари).

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-4.3 Способен участвовать в исследованиях и прикладных проектах в сфере междисциплинарного взаимодействия лингвистики и наук гуманитарного, математического и естественно-научного циклов.

2. Задачи освоения дисциплины

– освоить аппарат математических и лингвистических методов решения профессиональных задач с применением языков программирования;

– научится выбирать оптимальные подходы и методы решения конкретных научных и прикладных задач в области лингвистики и информационных технологи;

– овладеть навыками разработки программного кода, направленного на решение междисциплинарных прикладных задач; 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).

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

Третий семестр, экзамен

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: 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», «Языка программирования R», «Базы данных».

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 5 з.е., 180 часов, из которых:
-лекции: 10 ч.
-практические занятия: 44 ч.
в том числе практическая подготовка: 0 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Задачи и основные направления компьютерной лингвистики в изучении естественного языка.

Цели, задачи, проблемы обработки естественного языка. Интеграция методов лингвистики и машинного обучения

Тема 2. Первичная обработка текста. Основные операции с текстом. Сегментация текста

Библиотеки tm и quanteda, препроцессинг текстового массива данных.

Тема 3. Выделение ключевых слов и словосочетаний. WordEmbedding

Векторное представление слов: OneHotEncoding, BagOfWords, WordEmbeddings. Извлечение ключевых слов: создание словарей, триммирование матрицы, tf-idf

Тема 4. Частотный анализ лексических единиц с помощью математической статистики. Анализ N-грамм

Частотное построение n-грам в библиотеке quanteda. Анализ и визуализация данных

Тема 5. Автоматический морфологический анализ. Типы морфологических анализаторов. Интеграция морфологического анализатора Яндекса «Mystem» в язык программирования R

Тема 6. Машинное обучение: без учителя. Метрические алгоритмы классификации и кластеризации

Кластерный анализ, меры расстояний, методы кластеризации.

Тема 7. Методы сокращения выборки

Математические модели Lasso, PCA – метод главных компонент

Тема 8. Машинное обучение с учителем.

Виды и типы, наивный Байесовский классификатор.

Тема 9. Метод опорных векторов (SVM)

Цели, задачи и принципы работы SVM

Тема 10. Решающие списки и деревья

Принципы работы дерева решений: энтропия Шеннона, индекс Джинни, информационный прирост

Тема 11. Оценка результатов машинного обучения. и выбор моделей.

Формальные метрики оценки модели: accuracy, precision, recall, f1-score,

Тема 12. Нейронные сети. Типологизация НС. Генерация текстов с помощью НС.

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:

Изучите код, запускающий лемматизацию массивов текстов. Примените код на своих текстах, которые были скачены в прошлом семестре

```
data_path = "D:/stem"
list.files(path = "D:/stem", "txt")
# all subjects
#subject_id = c(c(1:3))
subject_id = list.files(path = "D:/stem", "txt")
# stem_id = c(c(1:3))
# write_id = c(c(1:3))
#filename = paste0(data_path, "/", subject_id[2])
for ( id in subject_id ) {
#filename = paste0(data_path, "/dlg_m_", id, ".txt")
filename = paste0(data_path, "/", id)
print(filename)
text <- read.table(filename, header = FALSE, sep = "\t", encoding = "russian",quote="",comment='')
ldf <- lapply(as.matrix(text$V1), function(x)x)
text2 <- as.character(ldf)
#text <- as.character(text$V1)
text3 <- enc2utf8(text2)
cmd <- "D:/stem/mystem.exe -cnild -e UTF-8"
#if (!is.null(fixlist) && file.exists(fixlist))
# cmd <- paste(filename, + id)
#system(cmd, intern = TRUE, input = read.table(filename, header = TRUE, sep = "\t", encoding = "UTF-8"))
# stem <- system(cmd, intern = TRUE, input = as.character(read.table(filename, header = TRUE, sep = "\t", encoding = "UTF-8"), output = write(stem,"0_")))
stem <- system(cmd, intern = TRUE, input = text3)
#for(id in subject_id){
myfile<-paste0("0_", id)
# write.table(stem,myfile,sep="\t",row.names=FALSE, col.names=F,)
```

```
#}
write.table(stem, myfile, row.names=F,col.names=F,sep="\t")
}
```

Пример задания темы 8

```
####Bayes classifier####
# Loading package
library(e1071)
library(caTools)
library(caret)
# Fitting Naive Bayes Model
# to training dataset
classifier_auth <- naiveBayes(Author ~ ., data = author.numtr)

classifier_auth

# Predicting on test data'
y_pred <- predict(classifier_auth, newdata = author.numte)

# Confusion Matrix
cm <- table(author.numte$Author, y_pred)
cm
```

Примените алгоритм классификации Байеса на своих данных (частотной матрицы).

10. Порядок проведения и критерии оценивания промежуточной аттестации

Экзамен в третьем семестре проводится в письменной форме по билетам. Экзаменационный билет состоит из трех частей.

Первая часть представляет собой тест из 2 вопросов, проверяющих ИПК-4.3, ИОПК-6.1. Ответы на вопросы первой части даются путем выбора из списка предложенных.

Вторая часть содержит один вопрос, проверяющий ИПК-4.1 и ОПК-3. Ответ на вопрос второй части дается в развернутой форме.

Третья часть содержит 1 вопрос, проверяющего следующие компетенции: ОПК-6, ИОПК-6.2, ИОПК-6.3, ИПК-3.1, ИПК-3.2 и оформленные в виде практических задач. Ответы на вопросы третьей части предполагают решение задач и краткую интерпретацию полученных результатов.

Примерный перечень теоретических вопросов

1. Опишите понятие энтропии Шеннона, дайте примеры использования. В каком методе машинного обучения используется энтропия?

2. Чем отличаются цели классификации и регрессии в машинном обучении?

3. Какие методы векторизации используются в репрезентации текстового массива данных? Опишите преимущества и недостатки методов

4. В каких случаях применяется индекс прироста информации? Опишите алгоритм его работы.

5. Опишите формальные метрики точности работы классификаторов. В чем преимущества и недостатки формальных метрик?

Примеры задач:

1. Задача 1. Построение линейных классификаторов

Дано: Матрица для обучения с признаками, исходный код двух классификаторов:

```
##LDA
author.lda<-lda(Author~.,data=author.numtr)
print(author.lda$scaling)
summary(author.lda)
```

```
author.ldapredicttrain<-predict(author.lda,author.numtr)
tldatrain<-table(author.numtr$Author,author.ldapredicttrain$class)
error(tldatrain)
f1(tldatrain)
author.ldapredicttest<-predict(author.lda,author.numte)
summary(author.ldapredicttest)
tldatest<-table(author.numte$Author,author.ldapredicttest$class)
error(tldatest)
print(paste0(c("text: ", 55)))
x1 <- "string1"
x2 <- "sring2"
paste0(x1, x2)
f1(tldatest)
plot(author.lda,col=author.train$colour)
plot(author.lda,dimen=1,type="both")
plot(author.lda,dimen=1,type="density")

##Logistic Regression
lr<-multinom(Author~.,data=author.numtr)
help(multinom)
lr.train<-predict(lr,author.numtr,type = "class")
error(table(author.numtr$Author,lr.train))

lr.test<-predict(lr,author.numte,type = "class")
error(table(author.test$Author,lr.test))
##SVM
# author.numtr$colour = NA
# author.numtr$colour[author.numtr$Author == "Austen"] = 0
# author.numtr$colour[author.numtr$Author == "London"] = 1
# author.numtr$colour[author.numtr$Author == "Milton"] = 2
# author.numtr$colour[author.numtr$Author == "Shakespeare"] = 3
# author.numtr <- author.numtr[,-15]
author.numtr$Author <- as.factor(author.numtr$Author)
author.svm<-svm(Author~.,data=author.numtr,
kernel="sigmoid")
help("svm")
summary(author.svm)

author.pred<-predict(author.svm,
author.numtr,decision.values=T)
ttrain<-table(author.numtr$Author,author.pred)
error(ttrain)

author.predtestsvm<-predict(author.svm,author.numte,decision.values=T)
ttest<-table(author.numte$Author,author.predtestsvm)
error(ttest)
```

Требуется: подобрать оптимальные гиперпараметры модели, влияющих на работу классификацараю

1. Задача 2. Энтропия Шеннона

Дано: Исходные классы в векторе:

```
train <- c(1:20)
x <- train
```

```
target = c(
0, 1, 1, 1,
1, 0, 0, 0,
0, 1, 1, 1,
1, 0, 0, 0,
0, 0, 0, 1
)
ones <- c()
i=1
j =1
for (i in 1:length(target)){
if (target[i]==1){
ones[j] = target[i]
i=i+1
j=j+1
}else{
i=i+1
}
}
zeros <- c()
i=1
j =1
for (i in 1:length(target)){
if (target[i]==0){
zeros[j] = target[i]
i=i+1
j=j+1

}else{
i=i+1
}
}
p_1 <- length(zeros)/length(target)
p_0 <- length(ones)/length(target)
p <- c(p_0, p_1)
S0 = sum((- (p * log2(p))))

target2 <- c(target[1:13])
ones <- c()
i=1
j =1
for (i in 1:length(target2)){
if (target2[i]==1){
ones[j] = target2[i]
i=i+1
j=j+1
}else{
i=i+1
}
}
zeros <- c()
i=1
```

```
j =1
for (i in 1:length(target2)){
if (target2[i]==0){
zeros[j] = target2[i]
i=i+1
j=j+1

}else{
i=i+1
}
}
p_1 <- length(zeros)/length(target2)
p_0 <- length(ones)/length(target2)
p <- c(p_0, p_1)
S1 = sum((- (p * log2(p))))

# Прирост информации
IG = S0 - sum(target)/length(x) * S1 - sum(target2)/length(x)
```

Задача: подобрать оптимальную энтропию Шеннона и прирост информации

Результаты экзамена определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

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 «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=14707

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

в) План семинарских / практических занятий по дисциплине.

Тема 1. Задачи и основные направления компьютерной лингвистики в изучении естественного языка.

Найдите источник текстов в интернете. Напишите парсер и скачайте структурированный массив текстов.

Тема 2. Первичная обработка текста. Основные операции с текстом. Сегментация текста

Изучение работы библиотеки qanteda, токенизация, сегментация, первичный анализ текстов (kwiq - формат).

Тема 3. Выделение ключевых слов и словосочетаний. WordEmbedding

Векторное представление слов: OneHotEncoding, BagOfWords, WordEmbeddings. Извлечение ключевых слов: создание словарей, триммирование матрицы, tf-idf при помощи библиотеки quanteda

Тема 4. Частотный анализ лексических единиц с помощью математической статистики. Анализ N-грамм

Частотное построение n-грам в библиотеке quanteda. Анализ и визуализация данных

Тема 5. Автоматический морфологический анализ. Типы морфологических анализаторов. Интеграция морфологического анализатора Яндекса «Mystem» в язык программирования R. Создание колонки лемматизированных текстов для метода BagOfWords

Тема 6. Машинное обучение: без учителя. Метрические алгоритмы классификации и кластеризации

Создание кластерного анализа текстов разными методами кластеризации и метриками расстояний: Евклидово расстояние, ближайшие и дальние соседи, расстояние Манхеттоона.

Тема 7. Методы сокращения выборки

Разработка Математические модели Lasso, PCA – метод главных компонент

Тема 8. Машинное обучение с учителем.

Виды и типы машинного обучения с учителем, наивный Байесовский классификатор.

Тема 9. Метод опорных векторов (SVM)

Разработка классификатора SVM

Тема 10. Решающие списки и деревья

Создание дерева решений. Интерпретация энтропии Шеннона, индекса Джинни, информационного прироста

Тема 11. Оценка результатов машинного обучения. и выбор моделей.

Разработка функции формальных метрик оценки модели: accuracy, precision, recall, f1-score. Библиотека caret.

Тема 12. Нейронные сети. Типологизация НС. Генерация текстов с помощью НС.

Классификация текстов при помощи нейронной сети transformer LSTM в библиотеке karet. Глубокое обучение, настройка и обучение модели на базе cpu и gpu.

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:

На основе своего корпуса создайте:

- Частотную матрицу Bag of Words (ключевые слова/части речи/словарь). Веса слов должны быть выражены относительными величинами

- Создайте кластерный анализ по своим данным, используя различные метрики расстояний и методов построения кластеров. Опишите полученный результат, в соответствии со своей гипотезой

- Визуализируйте полученный результат в виде дендрограммы.

Исходный код:

```
library(readtext)
library(quanteda)
DATA_DIR_s = "D:/tmp/"
texts_s <- readtext(paste0(DATA_DIR_s, "*.txt"), docvarsfrom = "filenames",
docvarnames = "document", encoding = "UTF-8")

library(cluster)

author<-read.table("author.txt",sep=",",header = T)
write.csv(author, "author.csv")
#correlation
boxplot(author$all ~ author$Author)
cor(author[,1:10], method = "pearson")
cor.test(author$a,author$all, method = "pearson")
help(cor.test)
dim(author)

library(corrplot)
cor_matrix <- cor(author[,1:10],
use = 'complete.obs',
method = "pearson")
corrplot.mixed(cor_matrix, lower = "circle",
upper = "number", tl.pos = "lt",
diag = "u")
#Pearson’s Correlation
cor(author[,1:5], author[,1:5], method = 'spearman') # method depends on distribution
cor(author[,1:5], author[,1:5], method = 'pearson')
cor(author[,1:5], author[,1:5], method = 'kendall')
library(GGally)
ggcorr(author[,1:5])
ggcorr(author[,1:5],
nbreaks = 6,
low = "steelblue",
mid = "white",
high = "darkred",
geom = "circle")
```

```
ggcorr(author[,1:10],
nbreaks = 6,
label = TRUE,
label_size = 3,
color = "grey50")
library(ggcorrplot)
p.mat <- cor_pmat(author[,1:30] , method = "pearson")
help(cor_pmat)

ggcorrplot(p.mat)
ggcorrplot(p.mat, hc.order = TRUE, type = "lower",
lab = TRUE)
ggcorrplot(p.mat, hc.order = TRUE,
type = "lower", p.mat = p.mat, lab = TRUE) #insig = "blank"

t.test(author$a ~ author$Author)#parametric for 2 classes

kruskal.test(author$also ~ author$Author) #nonparametric for multiple classes

names(author)
dim(author)
table(author$Author)
error<-function(x){((sum(x)-sum(diag(x)))/sum(x))*100}

##CLUSTER

#K-means
fit <- kmeans(author.train2, 3)
# get cluster means
aggregate(author.train2,by=list(fit$cluster),FUN=mean)
# append cluster assignment
mydata <- data.frame(author.train2, fit$cluster)
# Ward Hierarchical Clustering
d <- dist(mydata, method = "euclidean") # distance matrix
fit <- hclust(d, method="ward.D")
help(hclust)

plot(fit) # display dendogram
groups <- cutree(fit, k=2) # cut tree into 5 clusters
# draw dendogram with red borders around the 5 clusters
rect.hclust(fit, k=3, border="red")

# K-Means Clustering with 5 clusters
fit <- kmeans(mydata, 2)
```

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Джеймс Г. Введение в статистическое обучение (с примерами на языке R) / Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. – М.: ДМК, 2016. – 449 с.

– Боярский К. К. Введение в компьютерную лингвистику / Боярский К. К., Спб: ИТМО, 2013. – 73 с.

– Dalgaard P. Introductory Statistics with R New York, NY : : Springer-Verlag New York, , 2008. [Электронный ресурс: http://dx.doi.org/10.1007/978-0-387-79054-1]

б) дополнительная литература:

– Thomas Rahlf. Data Visualisation with R. Springer International Publishing, New York, 2017.

– Lawrence Leemis. Learning Base R. Lightning Source, 2016.

– Matthias Kohl. Introduction to statistical data analysis with R. bookboon.com, London, 2015.

– Torsten Hothorn and Brian S. Everitt. A Handbook of Statistical Analyses Using R. Chapman & Hall/CRC Press, Boca Raton, Florida, USA, 3rd edition, 2014.

– An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics), Corr. 7th printing / G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2017,

– Буховец А. Г. Статистический анализ данных в системе R. Учебное пособие / А.Г. Буховец, П.В. Москалев, В.П. Богатова, Т.Я. Бирючинская; Под ред. проф. Буховца А.Г. - Воронеж: ВГАУ, 2010. - 124 с.

– Кабаков Р. R в действии. Анализ и визуализация данных на языке R / Роберт И. Кабаков, – М.: ДМК, 2016. – 587 с.

– 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.

– Spector P. Data Manipulation with R / P. Spector New York, NY : Springer Science+Business Media, LLC, 2008

– Ergül Ö. Guide to Programming and Algorithms Using R electronic resource / Ö.Ergül - London : Springer London : Imprint: Springer, 2013. - 182 p.

– Pathak A. Beginning Data Science with R electronic resource / A. Pathak, A.Manas - Springer International Publishing : Imprint: Springer, 2014. - 157 p.

– Larry A. R Recipes electronic resource : A Problem-Solution Approach / Larry A. Pace - Berkeley, CA : Apress : Imprint: Apress, 2014. - 264 p.

– Joshua F. Beginning R electronic resource : An Introduction to Statistical Programming / F. Joshua Wiley, Wiley, Joshua F., Larry A. Pace. - Berkeley, CA : : Apress : Imprint: Apress,, 2015. - 327 p.

– Буховец А. Алгоритмы вычислительной статистики в системе R / А. Буховец - Санкт-Петербург: Лань , 2015. - 147 с.

в) ресурсы сети Интернет:

– открытые онлайн-курсы: https://online.stanford.edu/lagunita-learning-platform

– 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 смешенном формате («Актру»).

15. Информация о разработчиках

Степаненко Андрей Александрович, ассистент кафедры общей, компьютерной и когнитивной лингвистики филологического факультета ТГУ.