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

Филологический факультет

УТВЕРЖДАЮ:
Декан филологического факультета
_____ Тубалова И.В.
« 31 » 08 _____ 2023 г.

Рабочая программа дисциплины

Информационные технологии в филологии

по направлению подготовки

45.04.01 Филология

Направленность (профиль) подготовки:
Русский язык как иностранный

Форма обучения
Очная

Квалификация
Магистр

Год приема
2023

Код дисциплины в учебном плане: Б1.О.03

СОГЛАСОВАНО:
Руководитель ОПОП
_____М.М. Угрюмова

Председатель УМК
_____Ю.А. Тихомирова

Томск – 2023

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

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

– ПК-1 способен проводить самостоятельные исследования и получать новые научные результаты в области междисциплинарных лингвистических исследований

– 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:

ИУК-1.3 Предлагает и обосновывает стратегию действий с учетом ограничений, рисков и возможных последствий

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

ИУК-4.2 Применяет современные средства коммуникации для повышения эффективности академического и профессионального взаимодействия, в том числе на иностранном (ых) языке (ах)

ИУК-4.3 Оценивает эффективность применения современных коммуникативных технологий в академическом и профессиональном взаимодействиях

ИОПК-3.1 Демонстрирует знание существующих подходов и методов 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-6.3 Разрабатывает и отлаживает программный код, направленный на решение лингвистических и междисциплинарных задач с применением современных языков программирования

ИПК-1.1 Обнаруживает знания об актуальных направлениях междисциплинарных лингвистических исследований в избранной научной сфере.

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

2. Задачи освоения дисциплины

Целями освоения учебной дисциплины «Основные направления лингвистического обеспечения новых информационных технологий» – систематизация знаний по проблемам алгоритмизации и моделирования для решения лингвистических и исследовательских задач, с помощью современных информационных технологий, организации ведения результатов исследовательской деятельности с применением специализированных программ.

– Освоить инструменты ведения отчетности и написания статей.

– Научиться применять основные программные средства с целью увеличения эффективности обработки естественного языка.

– Научиться владеть формальными грамматиками, автоматизировать работу извлечения сущностей из текстового массива данных.

– Уметь интегрировать средства автоматизации обработки естественного языка в программный код.

– Знать и уметь пользоваться операционными системами (UNIX)

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к обязательной части образовательной программы.

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

Первый семестр, зачет

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования. Пререквизитами дисциплины являются следующие дисциплины: курс семантики и фонетики, основы программирования. Постреквизиты: Язык программирования P, учебная и производственная практики, Text mining с R

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 2 з.е., 72 часов, из которых:
-лекции: 6 ч.
-практические занятия: 20 ч.
в том числе практическая подготовка: 0 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Введение в информационные системы
1.1. Информационные технологии в современном мире
1.2. История информационных технологий
1.3. Архитектура и устройство персональных компьютеров
Тема 2. Типы операционных систем
2.1. Операционная система Windows
2.2. Операционная система Linux (Deb)
2.3. Работа в среде Linux
2.4. Создание и работа в системе контроля репозиторий (git)
2.5. Основные команды git
Тема 3. Морфологический анализ данных
3.1. Морфологические анализаторы
3.2. Работа в морфологическом анализаторе mystem, pymorphy
3.3. Использование сценариев и формальных грамматик. Работа ПО Tomita-parser
Тема 4. Организация научных исследований
4.1. Формирование структуры (отчета, диссертации, исследования) и содержания с помощью автоматизации в MS Office
4.2. Автоматическое формирование литературы (Mendeley, Zotero)
Тема 5. Организация научных исследований в Latex
5.1 Введение в Latex
5.2 Организация структуры документа
5.3 Работа с изображениями
5.4 Работа с формулами
5.5 Создание презентаций

Тема 6. Интеллектуальные алгоритмы и лингвистика

6.1. Тенденции развития современных технологий искусственного интеллекта

6.2. Использование интеллектуальных алгоритмов для решения лингвистических задач

6.3. Типы предварительной обработки текста для применения интеллектуальных алгоритмов

6.4. Проблема получения обучающих данных

Тема 7. Примеры интеллектуальных алгоритмов

7.1. Простейшая нейросеть на примере перцептрона

7.2. Обучение модели и оценка качества

7.3. Синонимическая близость и алгебраические операции над векторами

7.4. Рекуррентная модель и кодирование в вектор предложения

7.5. Модели глубокого обучения

Тема 8. Инструменты для разработки интеллектуальных алгоритмов

8.1. Язык Python и технология CUDA

8.2. TensorFlow и репозиторий моделей Google

8.3. Библиотека глубокого обучения PyTorch

8.4. HuggingFace и библиотека Transformers

8.5. Библиотеки NLTK и Pymorphy

Тема 9. Трансдукционные модели и трансформер

9.4. BERT как кодирующая половина трансформера

9.5. Идея Fine tuning

9.6. Генеративные модели и GPT как декодирующая половина трансформера

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, тестов по лекционному материалу, деловых игр по темам, выполнения домашних заданий, …, и фиксируется в форме контрольной точки не менее одного раза в семестр.

Текущий контроль по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, разработки кода, тестов по лекционному материалу, деловых игр по темам, выполнения домашних заданий. Контроль успеваемости обучающихся направлен на определение соответствия результатов обучения после освоения элемента по дисциплине и фиксируется в форме контрольной точки не менее одного раза в семестр. Примерные задания текущего контроля:

Первоначальный смысл английского слова «компьютер»:

Выберите один ответ:

a. вид АЛУ

b. электронно-лучевая трубка

c. набор ламп, выполняющих различные функции

d. человек, производящий расчеты

e. электронный аппарат

Машины первого поколения были созданы на основе…

Выберите один ответ:

a. электронно-вакуумных ламп

b. транзисторов

c. интегральных микросхем

d. зубчатых колес

Основной элементной базой ЭВМ третьего поколения являются…

Выберите один ответ:

a. микропроцессор

b. интегральные микросхемы

c. электромеханические схемы
d. транзисторы
Основной элементной базой ЭВМ четвертого поколения являются…
Выберите один ответ:
a. электромеханические схемы
b. полупроводники
c. электровакуумные лампы
d. микропроцессор

1. Написать небольшой код (Hello, world!) на Python и запустить его в терминале (код и скриншоты прикрепить в мудл)

2. Скачать книги (wget). Предварительно нужно проверить ссылки на корректность (li.txt). Этапы команд - скриншоты.

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. Примерный перечень практических заданий:
1. Разметка текста
Дан текст:

Что такое Git и зачем он нужен?
Git - это консольная утилита, для отслеживания и ведения истории изменения файлов, в вашем проекте. Чаще всего его используют для кода, но можно и для других файлов. Например, для картинок - полезно для дизайнеров.

С помощью Git-а вы можете откатить свой проект до более старой версии, сравнивать, анализировать или сливать свои изменения в репозиторий.
Репозиторием называют хранилище вашего кода и историю его изменений. Git работает локально и все ваши репозитории хранятся в определенных папках на жестком диске (ХДД).
Так же ваши репозитории можно хранить и в интернете. Обычно для этого используют три сервиса:
GitHub
GitLab
Каждая точка сохранения вашего проекта носит название коммит (commit). У каждого commit-а есть hash (уникальный id) и комментарий. Из таких commit-ов собирается ветка. Ветка - это история изменений. У каждой ветки есть свое название. Репозиторий может содержать в себе несколько веток, которые создаются из других веток или вливаются в них.
Как работает
Если посмотреть на картинку, то становиться чуть проще с пониманием. Каждый кружок, это commit. Стрелочки показывают направление, из какого commit сделан следующий. Например C3 сделан из C2 и т. д. Все эти commit находятся в ветке под названием main. Это основная ветка, чаще всего ее называют master . Прямоугольник main* показывает в каком commit мы сейчас находимся, проще говоря указатель.
Задача:
а. Сохраните текст в формате .txt
б. Создайте словарь для двух любых неизвестных лексем
в. Выполните анализ текста со следующими опциями: грамматическая информация, снятие омонимии, фикслист, не печатать исходные словоформы. Сохраните аутпут-файл в формате .json
- Дан шаблон конференции по компьютерной лингвистике «Диалог»

```
\documentclass{dialogue}

\begin{document}

\begin{otherlanguage}{english}
\begin{center}
{\Large\bfseries{A paper on\\news headline generation}}

\medskip

Surname N. P. (\texttt{user@domain.tld}), Surname N. P.
(\texttt{user@domain.tld}),\\Surname N. P.
(\texttt{user@domain.tld})

\medskip

Affiliation, City, Country
\end{center}

Your abstract in English: describe your system briefly\medskip

\textbf{Key words:} text summarization, headline generation,
Russian language (write your own keywords!)
\end{otherlanguage}

\bigskip

\begin{otherlanguage}{russian}
\begin{center}
{\Large\bfseries{Статья о\\генерации заголовков}}

\medskip

Фамилия И. О. (\texttt{user@domain.tld}), Фамилия И. О.
(\texttt{user@domain.tld}),\\Фамилия И. О.
(\texttt{user@domain.tld})

\medskip

Организация, город, страна
\end{center}

Аннотация на русском: кратко опишите ваши методы и
модели.\medskip

\textbf{Ключевые слова:} автореферирование текстов  (перечислите
ключевые слова)
\end{otherlanguage}

\selectlanguage{english}

\section{Introduction}
Describe (briefly!):
```

```
\begin{enumerate}
    \item the task and the dataset
    \item your approach and cite some major works, it was based
on
    \item the structure of your paper
\end{enumerate}
```

Задача:

Генерация заголовков новостей. Скачайте датасет новостей riatomsk.csv. В колонках есть следующие атрибуты: «lead» - лид новости, «title» -заголовок новости, «body» - тело новости. Обучите модель RuGPT-3 с целью генерации заголовка новости. Загрузите полученный код в мудл, прикрепите сгенерированные примеры заголовков в формате .txt.

Пример кода:

```
%%writefile setup.sh
git clone https://github.com/NVIDIA/apex
cd apex
pip install -v --disable-pip-version-check --no-cache-dir ./
!sh setup.sh
import re
import pandas as pd
from sklearn.utils import shuffle

data = pd.read_csv("/content/drive/MyDrive/news.csv",encoding='utf8',index
_col=0)

titles1 = data['Head']
print (titles1)

titles = titles1.dropna()
titles.convert_dtypes(convert_string=True)

texts1 = data['Text']
print (texts1)


texts = texts1.dropna()
texts.convert_dtypes(convert_string=True)

# создаем новый датафрейм
data2 = data[["Head", "Text"]]
# удаляем пропуски
data3 = data2.dropna(axis = 0, how = "any")
data4 = data3.astype('string')

titles1 = data['Head']
print (titles1)

titles = titles1.dropna()
titles.convert_dtypes(convert_string=True)

texts1 = data['Text']
print (texts1)
```

```
texts = texts1.dropna()
texts.convert_dtypes(convert_string=True)

# создаем новый датафрейм
data2 = data[["Head", "Text"]]
# удаляем пропуски
data3 = data2.dropna(axis = 0, how = "any")
data4 = data3.astype('string')

headlines = data4["Head"]
bodies = data4["Text"]
data5 = pd.concat([headlines, bodies])
data5 = shuffle(data5)

train = data5
valid = data5[:1000]
valid = shuffle(valid)

import numpy as np
import random
random.seed(1234)
np.random.seed(1234)

val_ind = random.sample(range(data5.shape[0]), 500)

with open("train.txt", "w") as file:
    file.write("\n".join(train))
with open("valid.txt", "w") as file:
    file.write("\n".join(valid))
!python3 pretrain_transformers.py \
    --output_dir=my_model \
    --model_type=gpt2 \
    --model_name_or_path=sberbank-ai/rugpt3small_based_on_gpt2 \
    --do_train \
    --train_data_file=train.txt \
    --do_eval \
    --fp16 \
    --eval_data_file=valid.txt \
    --per_gpu_train_batch_size 1 \
    --gradient_accumulation_steps 1 \
    --num_train_epochs 1 \
    --block_size 1024 \
    --overwrite_output_dir
```

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=14697

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

в) План семинарских / практических занятий по дисциплине.

Тема 1. Введение в информационные системы

1.1. Дать оценку текущему состоянию в области NLP

1.2. Подготовить презентацию своего исследования

Тема 2. Типы операционных систем

2.1. Установить Ubuntu на свой компьютер или виртуальную машину

2.2. Установка приложений

2.3. Работа с файловой системой и архивами

2.4. Создание и работа в системе контроля репозиторий (git)

2.5. Миграция проекта в репозеторий git

Тема 3. Морфологический анализ данных

3.1. Создание лемматезированного массива текстов

3.2. Создание фикслиста

3.3. Извлечение фактов в ПО Tomita-parser

Тема 4. Организация научных исследований

4.1. Формирование структуры (отчета, диссертации, исследования) и содержания с помощью автоматизации в MS Office

4.2. Автоматическое формирование литературы (Mendeley, Zotero) и интеграция в MS Office

Тема 5. Организация научных исследований в Latex

5.1 Введение в Latex, установка регистрация в сервисах латеха

5.2 Организация структуры документа, изучение тегов

5.3 Работа с изображениями: подготовка рисунка, вставка

5.4 Работа с формулами. Изучение тегов.

5.5 Создание презентаций. Работа по шаблонам

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.

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, 2019. 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:

– 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

– Система контроля версий ПО https://github.com/

– Создание документов Latex https://www.overleaf.com/

– Морфологический анализатор Mystem https://yandex.ru/dev/mystem/

– Извлечение фактов (формальные грамматики) Tomita-parser 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 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.).

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 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 общего славяно-русского языкознания и классической филологии