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

Филологический факультет

УТВЕРЖДЕНО:
Декан
И. В. Тубалова

Рабочая программа учебной практики

**Проектно-технологическая практика**

по направлению подготовки

**45.04.03 Фундаментальная и прикладная лингвистика**

Направленность (профиль) подготовки:
**«Компьютерная и когнитивная лингвистика»**

Форма обучения
**Очная**

Квалификация
**Магистр**

**Год приема**
**2024**

СОГЛАСОВАНО:
Руководитель ОП
З.И. Резанова

Председатель УМК
Ю.А. Тихомирова

Томск – 2025

**1. Цель практики**

Целями учебной практики по получению первичных умений и навыков являются: закрепление теоретических знаний и приобретение практических навыков, полученных при изучении базовых дисциплин; изучение организационной структуры лаборатории (далее предприятия) и действующих на нем систем управления; ознакомление с содержанием основных работ и исследований, выполняемых на предприятии или организации по месту прохождения практики.

Цели направлены на формирование следующих компетенций:

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

УК-3 Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели

УК-4 Способен применять современные коммуникативные технологии, в том числе на иностранном(ых) языке(ах), для академического и профессионального взаимодействия

УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки

ОПК-1 Способен решать профессиональные задачи, применяя основные понятия, категории и положения лингвистических теорий и актуальные концепции в области лингвистики

ОПК-2 Способен анализировать, сопоставлять и критически оценивать различные лингвистические направления, теории и гипотезы при решении задач профессиональной деятельности

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

ОПК-4 Способен расширять сферу научной деятельности, участвовать в междисциплинарных исследованиях на стыке наук

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

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-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.

**2. Задачи практики**

Задачами учебной практики по получению первичных знаний и умений являются – – Изучить организационную структуру базы практики как объекта информатизации, его особенности функционирования. (ИУК-4.2)

– Изучить особенности имеющихся на предприятии информационных систем, а также сбора, обработки, структуризации и передачи информации (ИУК-4.2, ИУК-4.1, УК-4).

– Собрать учебный материал для выполнения выпускных работ в процессе дальнейшего обучения в ВУЗе (ОПК-3, ИОПК-3.3, ПК-2, ИПК-2.2, ИПК-2.3, ПК-3, ИПК-3.1, ПК-4, ИПК-4.3). В частности:

-обработка текстов на естественном языке в научных целях (лингвистическая разметка, глоссирование, аннотирование, рубрикация, реферирование, редактирование);

-планирование, организация, проведение и обработка данных психолингвистических экспериментов;

-планирование организация сбора и обработки данных психолингвистических экспериментов;

-разработки, совершенствования и корректирования содержания электронных языковых ресурсов (текстовых, речевых и мультимодальных корпусов, словарей, тезаурусов, фонетических, лексических, грамматических и иных баз данных и баз знаний);

-сбор и структуризация таксового массива данных с целью формирования базы для дальнейшего исследования;

-обработка естественного языка современными методами компьютерной лингвистики.

**3. Место практики в структуре образовательной программы**

Практика относится к обязательной части образовательной программы.

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

Семестр 1, зачет.

Семестр 2, зачет с оценкой.

**5. Входные требования для освоения практики**

Для успешного освоения практики требуются результаты обучения по следующим дисциплинам: «Основные направления лингвистического обеспечения новых информационных технологий», «Профессиональный иностранный язык», «Статистические методы в гуманитарных исследованиях», «Экспериментальные методы лингвистического исследования», «Психолингвистика», «Язык программирования Python», «Язык программирования R».

**6. Способы и формы проведения практики**

Практика проводится на базе ТГУ – Лаборатория лингвистической антропологии, лаборатория «Когнитивные исследования языка». Способы проведения: стационарная.

Форма проведения: распределенная.

**7. Объем и продолжительность практики**

Объем практики составляет 6 зачётных единицы, 216 часов, из которых:

– лекции: 0 ч.;

– иная контактная работа: 6,8 ч.

Объем самостоятельной работы студента определен учебным планом.

Практика проводится в форме практической подготовки.

Продолжительность практики составляет 38 недель.

**8. Планируемые результаты практики**

Результатами прохождения практики являются следующие индикаторы достижения компетенций:

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

ИУК-3.1 Формирует стратегию командной работы на основе совместного обсуждения целей и направлений деятельности для их реализации;

ИУК-3.2 Организует работу команды с учетом объективных условий (технология, внешние факторы, ограничения) и индивидуальных возможностей членов команды;

ИУК-3.3 Обеспечивает выполнение поставленных задач на основе мониторинга командной работы и своевременного реагирования на существенные отклонения.

ИУК-4.1 Обосновывает выбор актуальных коммуникативных технологий (информационные технологии, модерирование, медиация и др.) для обеспечения академического и профессионального взаимодействия;

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;

ИУК-6.1 - Разрабатывает стратегию личностного и профессионального развития на основе соотнесения собственных целей и возможностей с развитием избранной сферы профессиональной деятельности;

ИУК-6.3 - Оценивает результаты реализации стратегии личностного и профессионального развития на основе анализа (рефлексии) своей деятельности и внешних суждений;

ИОПК-1.2 - Решает профессиональные задачи применяя основные понятия, категории и положения лингвистических теорий;

ИОПК-2.3 - Совершает выбор лингвистического направления, теории на основе их самостоятельного поиска и анализа, сопоставления, критической оценки при решении задач профессиональной деятельности;

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

ИПК-1.3 - Последовательно реализует исследовательскую программу, получает новые научные результаты;

ИПК-2.2 - Применяет имеющееся программное обеспечение и оборудование при проведении экспериментов в соответствии с технологическими возможностями для достижения цели исследования.

ИПК-2.3 Проводит экспериментальные исследования в соответствии с этическими нормами взаимодействия с респондентами, хранения и обработки данных.

ИПК-2.4 Применяет методы статистической обработки полученных экспериментальных данных и осуществляет их интерпретацию в соответствии с имеющимися теориями;

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-4.3 - 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.

**9. Содержание практики**

| **Этапы практики** | **Виды работ, связанные с будущей профессиональной деятельностью** | **Часы всего (в т.ч. контактные)** |
|---|---|---|
| 1. Организационный | 1. Проведение собрания по организации практики: | 4 (2) |

| | | |
|---|---|---|
| | – знакомство с целями, задачами, требованиями к практике и формами отчетности по практике (программой практики);<br>– знакомство с графиком проведения практики;<br>– подготовка дневников практиканта. | |
| 2. Ознакомительный | 1. Знакомство с правилами внутреннего распорядка и иными локальными нормативными актами ТГУ.<br>2. Инструктаж по технике безопасности и охране труда, соблюдению правил противопожарной безопасности, санитарно-эпидемиологических правил и гигиенических нормативов в ТГУ<br>3. Знакомство с проектной деятельностью лаборатории, ее целью и задачами, актуальными проектами. | 4 (2) |
| 3. Проектный | В зависимости от специализации студента, текущих целей лаборатории может применен дифференцированный подход формировании задач практики, учитывающий развитие компетенций, предусмотренной текущей программой практики<br>Производственный этап может включать следующую проектную деятельность:<br>I. Создание корпуса устной речи носителей русско-тюркского билингвизма (ИУК-2.3, ИПК-2.2, ИПК-2.3, ИПК-4.3, ИУК-4.1, ИУК-4.2)<br>Изучение литературы по корпусной лингвистике, знакомство с теоретическим обоснованием и особенностями создаваемой в проекте Лаборатории лингвистической антропологии.<br>1. Формирование реестра билингвов, носителей русско-тюркского билингвизма информантов;<br>2. Проведение анкетирования билингвов на основании социолингвистической и психолингвистической анкет.<br>3. Архивация данных анкет.<br>4. Записи устной речи носителей русско-тюркского билингвизма для формирования создаваемого в проекте Лаборатории лингвистической антропологии корпуса<br>5. Транскрибирование записей устной речи носителей русско-тюркского билингвизма.<br>6. Автоматическая разметка текстов записей устной речи носителей русско-тюркского билингвизма.<br>7. Ручная правка автоматической разметки текстов записей устной речи носителей русско-тюркского билингвизма.<br>8. Ручная разметка отклонений от речевого стандарта в текстах записей устной речи носителей русско-тюркского билингвизма. | 96 (4) |

| | | |
|---|---|---|
| | II. Создание психолингвистической базы данных (ИУК-2.3, ИОПК-2.3, ИОПК-3.3, ИОПК-4.3, ИПК-2.2, ИПК-2.3, ИПК-4.3, ИУК-4.1, ИУК-4.2)<br>1. Знакомство с принципами создания психолингвистических баз данных, с теоретическими основами психолингвистической базы данных, создаваемой в настоящее время в Лаборатории лингвистической антропологии.<br>2. Проведение анкетирования с целью выявления оценки отражения в семантике языковых пяти модальностей.<br>3. Архивация полученных данных<br>4. Разработка программного кода базы данных.<br>5. Создание интерфейса психолингвистической базы данных.<br>III. Проведение психолингвистических экспериментов, направленных на выявление разных аспектов взаимодействия языков в процессах порождения и восприятия речи. (ИУК-2.3, ИОПК-2.3, ИОПК-3.3, ИПК-2.2, ИПК-2.3, ИПК-4.3, ИУК-4.1, ИУК-4.2)<br>1. Знакомство с принципами проведения поведенческих психолингвистических экспериментов, изучение литературы.<br>2. Знакомство с принципами работы этического комитета Международного цента исследований развития человека.<br>3. Участие в организации и проведении психолингвистических экспериментов в качестве испытуемых<br>4. Организация и проведение психолингвистических экспериментов в разных группах испытуемых<br>5. Подготовка стимульного материала для проведения психолингвистических экспериментов разного типа.<br>6. Разработка скриптов для проведения поведенческих экспериментальных исследований языка при помощи специализированного программного обеспечения E-Prime, окулографического оборудования.<br>7. Обработка и интерпретация полученных данных.<br>IV. Парсинг и структуризация данных (ИУК-2.3, ИПК-3.1, ИПК-4.3, ИУК-4.1, ИУК-4.2)<br>1. Подготовительный этап.<br>2. Постановка задачи и выбор источников парсинга<br>3. Написание программного кода для выбранного сайта<br>4. Набор данных в соответствии с | |

| | | |
|---|---|---|
| | поставленными целями<br>5. Структуризация скаченных данных в виде таблиц с метаинформацией<br>6. Препроцессинг скаченных массивов<br>V. Извлечение фактов из текста (ИУК-2.3, ИПК-3.1, ИПК-4.3, ИУК-4.1, ИУК-4.2)<br>1. Подготовительный этап<br>2. Ознакомление с формальными грамматиками<br>3. Знакомство с современным ПО<br>4. Написание основных правил<br>5. Процесс извлечения фактов (написание правил, словарей)<br>VI. Анализ скаченных или предоставленных текстов (ИУК-2.3, ИПК-3.1, ИПК-4.3, ИУК-4.1, ИУК-4.2)<br>1. Подготовительный этап<br>2. Аналитика скаченных текстов в соответствии с текущими задачами лаборатории<br>VII. Искусственный интеллект в задачах обработки естественного языка<br>1. Структуризация данных<br>2. Предобработка текстовых данных<br>3. Генеративные модели<br>4. Модели суммаризации<br>5. ИИ в задачах классификации<br>6. Формальные метрики оценивания моделей<br>7. Обработка звучащей речи: модели vosk и whisper | |
| 5. Заключительный | 1. Подготовка отчета и подготовка материалов, необходимых для его защиты (презентация, методическая разработка и т.д.).<br>2. Защита отчета по итогам практики. | 4 (2) |

**10. Формы отчетности по практике**

По итогам прохождения практики обучающиеся в срок до завершения периода практики по календарному графику предоставляют руководителю практики от ТГУ:

– заполненный дневник практики;
– отчет о прохождении практики с приложениями, в которых включены этапы выполнения практических заданий (программный код, работа с эксериментами и пр.);
– программный код в соответствии с поставленными задачами;
– дизайны экспериментов;
– результаты обработки текстовых массивов данных;
– результаты проведенных экспериментов;
– иные виды работ, предусмотренные в рамках практики в лаборатории.

**11. Организация промежуточной аттестации обучающихся**

11.1 Порядок и форма проведения промежуточной аттестации

Промежуточная аттестация проводится в форме зачета с оценкой путем публичной защиты обучающимися индивидуальных отчетов о прохождении практики на итоговом

учебном занятии перед комиссией из не менее трех научно-педагогических работников, включая руководителя практики от ТГУ.

11.2 Процедура оценивания результатов обучения

Оценка сформированности результатов обучения осуществляется руководителем практики (комиссией) на основе анализа предоставленных отчетных документов, выступления обучающегося и его ответов на вопросы. Подведение итогов по этапам проведения (разделам) практики осуществляется по следующим критериям: а) практическая подготовленность обучающегося к решению конкретных профессиональных задач (соответствующих формируемым компетенциям); б) рефлексивность обучающегося: способность критически оценивать свою работу входе практики (в том числе – с точки зрения этических норм, в аспекте собственного личностного роста, с точки зрения возможности применения полученного опыта в предстоящей профессиональной деятельности и т.д.); в) сформированность универсальных / общекультурных компетенций, проявляющихся в том числе в своевременности, аккуратности и полноте выполнения всех видов работ на протяжении всех этапов практики, предусмотренных настоящей программой. Результаты текущего контроля каждого обучающегося отражаются в его дневнике практики, заверяются подписью руководителя практики.

Обучающиеся, не выполнившие программу практики без уважительной причины или получившие отрицательную оценку, считаются имеющими академическую задолженность и обязаны ликвидировать академическую задолженность в порядке, установленном в локальных документах Университета.

11.3 Критерии оценивания результатов обучения

Результаты прохождения практики определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно» / «зачтено», «незачтено». Результат прохождения практики может оценен на «отлично» (зачтено), если программа практики выполнена в полном объеме; практическая деятельность проведена на высоком научном и организационно-методическом уровне, формулировались и эффективно решались практические задачи, рационально применялись разнообразные методы и приемы практической деятельности; студент проявил глубокое знание теоретического материала и творческую самостоятельность в подборе материала при построении, проведении и анализе отчетной документации; студент показал в полной мере личностные качества специалиста в области компьютерной и когнитивной лингвистики; своевременно предоставил качественно оформленную отчетную документацию по практике, в которой предоставлен глубокий анализ результатов практики. Компетенции, закреплённые за практикой, сформированы на уровне – высокий

Оценка «Хорошо» (зачтено). Программа практики выполнена в полном объеме; практическая деятельность проведена на высоком научном и организационно-методическом уровне; однако недостаточно эффектно формулировались и решались практические задачи, применялись разнообразные методы и приемы практической деятельности; студент показал достаточные знания теоретического материала, самостоятельность в подборе материала при построении, проведении и анализе отчетной документации; достаточно успешно справляется с выполнением исследовательских процедур и на теоретическом, и на эмпирическом уровне (осознанно и грамотно); своевременно предоставил качественно оформленную отчетную документацию по практике. К недостаткам можно отнести: содержание предоставленной отчетной документации характеризуется недостаточно глубоким самоанализом деятельности. Компетенции, закреплённые за практикой, сформированы на уровне – хороший(средний)

Оценка «Удовлетворительно» (зачтено). Недостаточно эффективно применял теоретические, методологические и технологические методы и приемы, слабо активизировал познавательную деятельность, не всегда мог выполнить поставленные практические задачи (ошибки в коде, неточности и ошибки в проектировании и проведении экспериментов), при анализе собственной практической деятельности не видел своих

ошибок и недостатков; допущены серьезные ошибки при заполнении отчетной документации; нерационально организовывал свою практическую деятельность на рабочем месте в учреждении-базе практики; выявлена неорганизованность и недостаточная ответственность в практической деятельности; студент пропустил запланированные виды работ, без уважительной причины; может ориентироваться в основных характеристиках исследования, допуская при этом ошибки в трактовках и формулировании конкретных положений. Может действовать только по образцу; несвоевременно представил отчетную документацию, которая характеризуется неглубоким анализом, поверхностностью и тезисностью изложения итогов прохождения практики. Компетенции, закреплённые за практикой, сформированы на уровне – достаточный.

Оценка «неудовлетворительно» (не зачтено). Не может самостоятельно выполнять задания; студент не явился на занятие без уважительной причины и без предупреждения; студент проявил безответственность, недисциплинированность, халатность в ходе практики; не предоставил отчетную документацию. Компетенции, закреплённые за практикой, сформированы на недостаточном уровне или не сформированы.

**12. Учебно-методическое обеспечение**

а) Электронный учебный курс по практике в электронном университете «Moodle» - https://moodle.tsu.ru/course/view.php?id=14659

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

в) Методические указания по подготовке отчета по практике.

Практика является составной частью основной профессиональной образовательной программы подготовки обучающегося и обеспечивает профессионально-практическую подготовку студентов на базах практики: организациях и структурных подразделениях ТГУ. Обучающиеся в период прохождения практики: выполняют индивидуальные задания, предусмотренные программами практики; соблюдают правила внутреннего трудового распорядка; соблюдают требования охраны труда и пожарной безопасности. Обучающийся, не вышедший на базу практики и не выполнивший программу практики по уважительной причине. При подведении итогов (при промежуточной аттестации по практике) такому обучающемуся оценка автоматически снижается на один балл (по 5-балльной шкале). Обучающийся, приступивший к практике, однако систематически нарушающий учебную дисциплину (срыв запланированных диагностических, коррекционных и иных мероприятий, выход на базу без необходимой подготовки к выполнению практических заданий и т.д.), не соблюдающий внутренний распорядок базы и этические нормы профессиональной деятельности, снимается с базы практики, не аттестуется по практике и представляется к отчислению как не выполнил программу практики без уважительных причин

Обучающийся не выполнил программу практики без уважительных причин, или не представил отчёт о практике в установленный приказом срок, или при защите отчёта по практике на Комиссии получил неудовлетворительную оценку, подлежит отчислению из университета за невыполнение обязанностей по добросовестному освоению основной профессиональной образовательной программы и выполнению учебного плана.

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:

Создание словаря и его частотного распределения в текстах (пример):

```
library(readxl)
#library(quanteda.sentiment)
library(quanteda)
# install.packages("remotes")
# remotes::install_github("quanteda/quanteda.sentiment")
tmp <- read_excel("full_word_rating_after_coding.xlsx", col_names = TRUE)

df #head body stem class
mycorp <- corpus(df, text_field = "stem", )

dict2 <- dictionary(list(neg = c(tmp$word[tmp$value < 0]),
                         neut = c(tmp$word[tmp$value==0]),
                         pos = c(tmp$word[tmp$value>0])))
sent_pres <- mycorp_vk %>%
  corpus_subset(gnd == "f")
sent_pres2 <- mycorp_vk %>%
  corpus_subset(gnd == "m")
summary(mycorp_vk)
x_m <- tokens_lookup(tokens(sent_pres), dictionary = dict2) %>%
  dfm()
x_f <- tokens_lookup(tokens(sent_pres2), dictionary = dict2) %>%
  dfm()
x_m <- dfm_weight(x_m, scheme = "prop")
x_f <- dfm_weight(x_f, scheme = "prop")
x_full <- tokens_lookup(tokens(mycorp_vk), dictionary = dict2) %>%
  dfm()
x_f <-convert(x_f, to = "data.frame")
x_m <-convert(x_m, to = "data.frame")
x_m$gnd <- "f"
x_f$gnd <- "m"
x_full_abs_vkwall <- rbind(x_f, x_m)
x_full_abs_vkwall$ComType <- "vkw"
write.csv(x_full_abs_vkwall, "sentiment_vkwall_gnd.csv")
ggplot(x_full_abs_vkwall, aes(doc_id, neut, fill = gnd, group = gnd)) +
        geom_bar(stat='identity', position = position_dodge(), size = 1) +
scale_fill_brewer(palette = "Set1") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  ggtitle("Sentiment scores in twelve Sherlock Holmes novels") + xlab("")
x_m2 <-  convert(x_m, to = "data.frame")
x_m2 <- as.data.frame(x_m)
x_f2 <-  convert(x_f, to = "data.frame")
x_f2 <- as.data.frame(x_f)
```

Корреляционный анализ:

```
library(outliers)
grubbs.test(emot_int$neg, type = 10)

grubbs.test(emot_int$pos, type = 10)
```

```
grubbs.test(emot_int$neut, type = 10)
library(ggplot2)
p = ggplot(emot_int[,-1], aes(x=self))
(p <- p+geom_density(aes(fill=gnd), alpha=1/2))

# Sample data
data <- emot_int[, 2:4] # Numerical variables
groups <- as.factor(emot_int[, 5]) # Factor variable (groups)
# Plot correlation matrix
pairs(data)

# Equivalent with a formula
pairs(~ neg+pos+neut, data = emot_int)


pairs(data,                    # Data frame of variables
      labels = colnames(data),  # Variable names
      pch = 1,                 # Pch symbol
      bg = rainbow(2)[groups],  # Background color of the symbol (pch 21 to 25)
      col = rainbow(2)[groups], # Border color of the symbol
      main = "",    # Title of the plot
      row1attop = TRUE,         # If FALSE, changes the direction of the diagonal
      gap = 1,                  # Distance between subplots
      cex.labels = NULL,        # Size of the diagonal text
      font.labels = 1)          # Font style of the diagonal text

panel.hist <- function(x, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5))
  his <- hist(x, plot = FALSE)
  breaks <- his$breaks
  nB <- length(breaks)
  y <- his$counts
  y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = rgb(0, 1, 1, alpha = 0.5), ...)
  # lines(density(x), col = 2, lwd = 2) # Uncomment to add density lines
}

# Creating the scatter plot matrix
pairs(data,
      upper.panel = NULL,         # Disabling the upper panel
      diag.panel = panel.hist)    # Adding the histograms
# Function to add correlation coefficients
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  Cor <- abs(cor(x, y)) # Remove abs function if desired
  txt <- paste0(prefix, format(c(Cor, 0.123456789), digits = digits)[1])
  if(missing(cex.cor)) {
    cex.cor <- 0.4 / strwidth(txt)
```

```
  }
  text(0.5, 0.5, txt,
       cex = 1 + cex.cor * Cor) # Resize the text by level of correlation
}

# Plotting the correlation matrix
pairs(data,
      upper.panel = panel.cor,    # Correlation panel
      lower.panel = panel.smooth) # Smoothed regression lines

# install.packages("gclus")
library(gclus)

# Correlation in absolute terms
corr <- abs(cor(data))

colors <- dmat.color(corr)
order <- order.single(corr)

cpairs(data,                 # Data frame of variables
       order,                # Order of the variables
       panel.colors = colors,   # Matrix of panel colors
       border.color = "grey70", # Borders color
       gap = 0.45,              # Distance between subplots
       main = "Ordered variables colored by correlation", # Main title
       show.points = TRUE,      # If FALSE, removes all the points
       pch = 21,                # pch symbol
       bg = rainbow(2)[groups]) # Colors by group
```

Пример работы по генерации 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
```

Задача: сгенерировать заголовок новости, дообучив модель ruGPT-3, на основе датасета «riatomsk.csv»
Пример кода для генеративных моделей языка:

```
!pip3 install urllib3==1.26.4
!pip3 install transformers==2.8.0
!pip3 install wget

import wget

wget.download('https://raw.githubusercontent.com/sberbank-ai/ru-gpts/
master/generate_transformers.py', './')
wget.download('https://raw.githubusercontent.com/sberbank-ai/ru-gpts/
master/pretrain_transformers.py', './')

%%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')

l = pd.Series(data4['Text']).str.replace(r'^.*?\.', '', regex=True)
m = data4['Head']
data4 = pd.concat([m, l], axis=1)

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
```

Пример самостоятельной работы по преобразованию звучащей речи:

```
!git clone https://git.ffmpeg.org/ffmpeg.git ffmpeg
!pip install vosk
!pip install wget pydub wave tqdm
!apt-get ffmpeg

# Download Vosk model
!mkdir models
!wget -P models/ https://alphacephei.com/vosk/models/vosk-model-small-
ru-0.22.zip
!unzip models/vosk-model-small-ru-0.22.zip -d models/ && rm models/vosk-
model-small-ru-0.22.zip
!pip install pydub
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
from google.colab import drive
drive.mount('/content/gdrive')
mp3_to_wav('/content/gdrive/MyDrive/example.mp3', 0, True)
from vosk import Model, KaldiRecognizer, SetLogLevel
from tqdm.notebook import tqdm
import wave
import os
import json

def transcript_file(input_file, model_path):

    # Check if file exists
    if not os.path.isfile(input_file):
```

```
        raise FileNotFoundError(os.path.basename(input_file) + " not
found")

    # Check if model path exists
    if not os.path.exists(model_path):
        raise FileNotFoundError(os.path.basename(model_path) + " not
found")

    # open audio file
    wf = wave.open(input_file, "rb")

    # check if wave file has the right properties
    if wf.getnchannels() != 1 or wf.getsampwidth() != 2 or
wf.getcomptype() != "NONE":
        raise TypeError("Audio file must be WAV format mono PCM.")

    # Initialize model
    model = Model(model_path)
    rec = KaldiRecognizer(model, wf.getframerate())

    # Get file size (to calculate progress bar)
    file_size = os.path.getsize(input_file)

    # Run transcription
    pbar = tqdm(total=file_size)

    # To store our results
    transcription = []

    while True:
        data = wf.readframes(4000) # use buffer of 4000
        pbar.update(len(data))
        if len(data) == 0:
            pbar.set_description("Transcription finished")
            break
        if rec.AcceptWaveform(data):
            # Convert json output to dict
            result_dict = json.loads(rec.Result())
            # Extract text values and append them to transcription list
            transcription.append(result_dict.get("text", ""))

    # Get final bits of audio and flush the pipeline
    final_result = json.loads(rec.FinalResult())
    transcription.append(final_result.get("text", ""))

    transcription_text = ' '.join(transcription)

    return transcription_text
transcription =
transcript_file('/content/gdrive/MyDrive/example_excerpt.wav',
```

```
'/content/models/vosk-model-small-ru-0.22')
#/content/gdrive/MyDrive/example_excerpt.wav
transcription
```

‘только двое какого возраста сын да не тринадцать лет и дочери яна семь лет хорошо где выросли вы сказали что достаточно переехали в этих котик васкес с чем связано с у меня просто мама закончила институт родила меня на этом курсовых закончил институт они по их поехали времени работа тире просто направлению раньше же выдавали жилье заканчивали она учитель языка и литературы закончила и выделенный жилье всего родители сейчас подражать господа а какое воздействие что такое деревне свою счастливое детство тем занимались дети нас мы сначала жили на одной улице’

## 13. Перечень рекомендованной литературы и ресурсов сети Интернет

а) основная литература:


– Janyan, I. Vankov, O. Tsaregorodtseva, A. Miklashevsky (2015) Remember down, look down, read up: Does a word modulate eye trajectory away from remembered location? Cognitive Processing, 16 (Suppl. 1), 259-263. DOI: 10.1007/s10339-015-0718-5

– Гришина Е. А., Савчук С. О. Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 129—149.

– Миклашевский А.А. Проект психолингвистической базы данных: взаимоотношение между модальностью и другими характеристиками русских существительных //Когнитивные исследования языка. М., Тамбов, СПб., 2015. Вып. XXII: Язык и сознание в междисциплинарной парадигме исследований: материалы Международного конгресса по когнитивной лингвистике. 30 сентября - 2 октября 2015 г. С. 550-551

– Бровко С.Л. PR-мероприятия: методика подготовки и проведения эффективных событий: методология PR-мероприятий: концепция петербургской школы PR. – СПб., 2012. – 187 с.

– Катлип С., Центер А., Брум Г. Паблик рилейшнз: Теория и практика. – М.: Вильямс, 2003. – 614 с.

– Кузнецов В.Ф. Связи с общественностью: Теория и технологии: Учебник для вузов. М.: Аспект Пресс, 2009. – 391 с.

– Резанова З.И., Некрасова Е.Д., Миклашевский А.А. Исследование психолингвистических и когнитивных аспектов языкового контактирования в проекте "языковое и этнокультурное разнообразие Южной Сибири в синхронии и диахронии: взаимодействие языков и культур//Русин. 2018. № 2 (52). С. 107-117.

–. Сичинава Д. В. Обработка текстов с грамматической разметкой: инструкция разметчика // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 136—154.

– Сичинава Д. В. Обработка текстов с грамматической разметкой: инструкция разметчика // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 136—154.


б) дополнительная литература:


– Зубов А.В. Информационные технологии в лингвистике учебное пособие для студентов вузов, обучающихся по специальности 021800 – Теоретическая и прикладная лингвистика. - М.: Академия, 2004. – 205 с.

– Мишанкина Н.А. Базы данных в лингвистических исследованиях. Вопросы лексикографии, Выпуск № 1 (3), 2013 – С.25-33.

– Поляков А. Е. Технология подготовки информации в Национальном корпусе русского языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 175—192.

– Потапова Р.К. Новые информационные технологии и лингвистика. – М.: Ленанд, 2016. – 364 с.

– Хроленко А.Т., Денисов А.В. Современные информационные технологии для гуманитария: практическое руководство. – 3-е изд. – М.: Флинта: Наука, 2010. - 127, с.

– Информатика для гуманитариев: Учебник и практикум / Кедрова Г.Е. - Отв. ред. М.: Юрайт, 2016. 439 с.

в) ресурсы сети Интернет:

– открытые онлайн-курсы

– репозиторий github.com

**14. 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);

– E-prime v2.0 или v3.0

– 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)*:

– База данных RuWordPerception (ТГУ) – http://clingv.ru:3839/

**15. Материально-техническая база проведения практики**

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения экспериментальной части практики.

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.

Материально-техническая база профильной организации, включая перечень помещений, предоставленных профильной организацией в соответствии с приложением 2 к договору о практической подготовке обучающихся.

**16. Информация о разработчиках**

Степаненко Андрей Александрович, Томский государственный университет, старший преподаватель кафедры общей, компьютерной и когнитивной лингвистики.