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

Филологический факультет

УТВЕРЖДЕНО:
Декан
И. В.Тубалова

Оценочные материалы по дисциплине

Большие данные в филологических исследованиях

по направлению подготовки

45.04.01 Филология

Направленность (профиль) подготовки:
Академическая филология: современные исследовательские технологии

Форма обучения
Очная

Квалификация
Магистр

Год приема
2025

СОГЛАСОВАНО:
Руководитель ОП
Н.А. Мишанкина

Председатель УМК
Ю.А. Тихомирова

Томск – 2025

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

ОПК-2 Способен использовать в профессиональной деятельности, в том числе педагогической, знания современной научной парадигмы в области филологии и динамики ее развития, системы методологических принципов и методических приемов филологического исследования.

ОПК-3 Способен владеть широким спектром методов и приемов филологической работы с различными типами текстов..

ПК-1 Выполнение отдельных заданий в рамках решения исследовательских задач в сфере филологии под руководством более квалифицированного работника.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИОПК-2.1 Демонстрирует знание современной научной парадигмы в области филологии и динамики ее развития, системы методологических принципов и методических приемов филологического исследования.

ИОПК-3.1 Демонстрирует углубленные знания в избранной конкретной области филологии и владение характерным для нее спектром методов и приемов филологической работы с различными типами текстов.

ИПК-1.1 Владеет методами и способами решения научных задач по тематике проводимого исследования, под руководством более квалифицированного работника намечает путь решения исследовательской задачи, методологию и методику исследования.

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

-- тесты;
-- практические кейс-задания;
-- аналитический отчет.

2.1. Тест (ИОПК-2.1, ИОПК-3.1)
Вариант 1

1. Какое из приведенных ниже определений наиболее точно характеризует термин «метаданные» в лингвистическом исследовании?
а) Полные тексты художественных произведений
б) Данные о данных: информация о тексте (автор, год издания, жанр, источник)
в) Статистические формулы для анализа частотности слов
г) Субъективные читательские рецензии

2. Какой тип визуализации наиболее уместен для представления динамики изменения частоты употребления слова по десятилетиям?
а) Круговая диаграмма
б) Линейный график
в) Диаграмма рассеяния (scatter plot)
г) Схема дерева

3. Что подразумевается под «очисткой данных» (data cleaning) при подготовке лингвистического датасета?
а) Удаление всех данных, кроме имен собственных
б) Перевод всех текстов на английский язык
в) Приведение данных к единообразному виду (регистр, устранение опечаток, формат дат)

г) Написание литературного комментария к текстам

Ключ к тесту: 1 – б, 2 – б, 3 – в.
Критерии оценивания: тест считается пройденным на «отлично» при 100% правильных ответов, «хорошо» – при 66% (2/3), «удовлетворительно» – при 33% (1/3). Менее 33% – «неудовлетворительно».

2.2. Практическое кейс-задание (ИОПК-3.1, ИПК-1.1)
Тема: «Создание и очистка датасета для анализа лингвистических метаданных»

Задание:
Вам предоставлен CSV-файл с метаданными 150 научных статей по лингвистике, извлеченными из базы данных OPENALEX. Файл содержит ошибки и несоответствия. Используя приложение для работы с таблицами, выполните следующее:
1. Приведите названия всех журналов к единому регистру (например, нижнему).
2. Устраните дубликаты записей.
3. Столбец «Год издания» содержит текстовые и числовые значения. Приведите все значения к числовому формату, ошибочные значения удалите.
4. В столбце «Ключевые слова» разделите слова, записанные через запятую с пробелом, на отдельные ячейки с помощью инструмента «Разделить текст на столбцы».

Критерии оценивания:
«Отлично»: Все задачи выполнены полностью и корректно. Использованы адекватные инструменты (формулы, инструменты обработки). Результат представлен в виде ссылки на корректно отформатированную таблицу.
«Хорошо»: Задачи выполнены с незначительными погрешностями (например, остались неучтенные дубликаты).
«Удовлетворительно»: Выполнена только часть заданий, процесс очистки не завершен.
«Неудовлетворительно»: Задание не выполнено или выполнено неверно.

2.3. Аналитический отчет (ИОПК-2.1, ИОПК-3.1)
Тема для отчета: «Сравнительный анализ метаданных двух лингвистических журналов по выбору студента за последние 5 лет».

Структура отчета:
1. Краткое описание объекта исследования и источника данных (например, «Вопросы языкознания» и «Научный диалог» на elibrary.ru).
2. Описание методов сбора и обработки метаданных (количество статей, отобранные параметры: год, авторство, количество страниц, ключевые слова).
3. Визуализация (2-3 графика, например, распределение статей по годам, облако ключевых слов).
4. Письменная интерпретация результатов: какие выводы о фокусе, динамике развития и научной политике журналов можно сделать на основе анализа метаданных?

Критерии оценивания:
«Отлично»: Отчет имеет четкую структуру. Визуализации релевантны и технически правильно выполнены. Интерпретация глубокая, демонстрирует умение переводить данные в научные выводы.
«Хорошо»: Структура соблюдена, но интерпретация поверхностна или есть недочеты в визуализациях.
«Удовлетворительно»: Отчет формальный, интерпретация отсутствует или сводится к описанию графиков.

«Неудовлетворительно»: Отчет не представлен или не соответствует заданию.

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Промежуточная аттестация по дисциплине проводится в форме зачета с оценкой и представляет собой защиту итогового кейс-проекта.

Структура итогового контроля:
Зачет проводится в устной форме. Студент представляет презентацию по результатам выполнения итогового кейс-проекта. Продолжительность презентации – 7-10 минут, продолжительность ответа на вопросы – 5-7 минут.

Тематика итоговых кейс-проектов (на выбор студента):
1. Анализ дискурса в социальных сетях: Сбор метаданных (дата, время, лайки, репосты) и текстов постов из тематического сообщества (паблика) о лингвистике/русском языке. Анализ активности аудитории и визуализация сетевых взаимодействий.
2. Эволюция терминологического поля: Анализ метаданных и ключевых слов статей из журналов по лингвистике за 10-летний период для выявления тенденций в тематике научных исследований.
3. Корпусный анализ на метауровне: Сравнение метаданных двух корпусов текстов (например, Национального корпуса русского языка и корпуса современного медиадискурса) по параметрам: источник текстов, жанровое разнообразие, временной охват.

Критерии оценивания итогового проекта:
«Отлично» (5): Исследовательский вопрос сформулирован четко и соответствует возможностям анализа метаданных. Выбор методов и инструментов сбора/анализа адекватен. Визуализации профессиональны и наглядны. Интерпретация результатов глубокая, продемонстрировано владение методологией филологического исследования.
«Хорошо» (4): Проект выполнен полностью, но интерпретация результатов недостаточно глубокая или имеются незначительные ошибки в методологии.
«Удовлетворительно» (3): Проект выполнен формально, исследовательский вопрос слишком общий, интерпретация результатов отсутствует или сводится к описанию проделанной работы.
«Неудовлетворительно» (2): Проект не представлен или не соответствует заявленной теме, методы анализа применены неверно.

Влияние текущего контроля на итоговую оценку:
Итоговая оценка выставляется на основе оценки за итоговый кейс-проект. Результаты текущего контроля (тесты, практические задания) являются допуском к зачету. Студент, не выполнивший более 50% заданий текущего контроля, к сдаче зачета не допускается.

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

4.1. Тест (ИОПК-2.1, ИОПК-3.1)
1. Distant reading – это методологический подход, который:
а) Предполагает углубленное медленное чтение одного текста
б) Основывается на количественном анализе больших массивов текстовых данных для выявления масштабных patterns
в) Используется исключительно для анализа поэтических текстов
г) Заменяет собой все традиционные филологические методы

2. Для визуализации взаимосвязей между соавторами научных статей наиболее подходит:
а) Линейный график
б) Столбчатая диаграмма (гистограмма)
в) Сетевая диаграмма (graph)
г) Круговая диаграмма

3. При работе с метаданными этическим принципом является:
а) Публикация всех персональных данных авторов без ограничений
б) Использование данных без указания источника
в) Обеспечение конфиденциальности и анонимности там, где это необходимо
г) Отсутствие необходимости в этических нормах при работе с данными

Ключ: 1 – б, 2 – в, 3 – в.

4.2. Практическая задача (ИПК-1.1)
Задача:
Вам поручено подготовить датасет для анализа частоты публикаций по корпусной лингвистике. В raw-файле содержится список статей с полями: `Название; Автор; Год; Журнал`.
Обнаружены ошибки: в поле «Год» для некоторых статей проставлено значение «н.д.», в поле «Журнал» один и тот же журнал назван «ВЯ», «Вопр. язык.» и «Вопросы языкознания».
Опишите последовательность действий для очистки данных и приведения их к единообразному виду.
Правильный ответ должен включать упоминание функций `НАЙТИ И ЗАМЕНИТЬ` для унификации названий, сортировки и фильтрации для поиска аномалий, условного форматирования или формул для поиска и удаления/замены значений «н.д.».

Информация о разработчиках

Кашпур Валерия Викторовна, канд. филол. наук, доцент кафедры романо-германской и классической филологии НИ ТГУ