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

Информационные технологии в филологии

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:

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

ПК-2 Представление результатов научных исследований в сфере филологии профессиональному сообществу..

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

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-1.2 Реализует под руководством более квалифицированного работника план решения исследовательской задачи, используя необходимые информационные ресурсы, методы получения данных, при необходимости производя его корректировку в свете полученных промежуточных результатов.

ИПК-2.2 Использует современные технологии для представления результатов научных исследований в сфере филологии.

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 языке.

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

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

– практические задания;

– устный / письменный опрос;

– тесты;

– контрольная работа;

– написание кода;

– самостоятельная работа.

Практические задачи

(ИУК-4.3, ИПК-1.2, ИПК-2.2):

1. При помощи гугл коллаб загрузите текстовый документ и осуществите предобработку текста.

2. Обучите модель word2vec и fastText. Опишите их отличия, преимущества и недостатки.

3. Придумайте гипотезу по своим текстам и проверьте ее на основе арифметических операций над векторами.

4. Создайте модель классификации текстов (позитивный отзыв, негативный, нейтральный) BERT и оцените ее формальными метриками.

5. Обучите модель генерации GPT-3, сгенерируйте отзыв.

6. Создайте словарь в программе Mystem.

7. Проанализируйте звучащий файл на предмет присутствия интерференции.

8. Создайте формальные грамматики, извлекающие коллокаты в массиве текстов

9. Напишите код модели VOSK или Whisper, преобразующий звучащую речь в текст.

10. Организуйте систему хранения и вставки в текстовый редактор списка литературы при помощи библиографического менеджера.

Критерии оценивания практической работы:

Оценка	Критерии
	1. Понимание и логика алгоритма работы 2. Наличие или отсудив ошибок в коде 3. Полнота решения практических задач 4. Своевременность выполнения; 5. Умения связать практический материал с теоретическим; 6. Понимание базовых формул обработки естественного языка и программирования;
«зачтено»	Основные требования к решению практических задач выполнены. Продемонстрированы умение анализировать алгоритмы и находить оптимальное количество решений, умение работать с информацией, в том числе умение затребовать дополнительную информацию, необходимую для уточнения реализации алгоритма, навыки разработки программного кода; Основные требования к решению практических задач выполнены, но при этом допущены недочеты. В частности, недостаточно раскрыты навыки стиля, недостаточно комментариев Имеются существенные отступления от решения практических задач. В частности отсутствуют навык и умения моделировать решения в соответствии с заданием, представлять различные подходы к разработке алгоритмов, ориентированных на конечный результат
«не зачтено»	Задача не решена, обнаруживается существенное непонимание проблемы

Устный / письменный опрос (УК-4, ПК-1, ПК-2, ИУК-4.1, ИУК-4.2**)**

1. Какие языковые модели широко используются в обработке естественного языка?
2. Понятие искусственного интеллекта в задачах филологии
3. Какие задачи решаются при помощи определения тональности текста в NLP?
4. Какие методы используются для извлечения ключевых слов из текста в NLP?
5. Опишите роль и задачи контекстно-свободных формальных грамматик
6. Какие задачи решаются при помощи именованного сущностного распознавания ?

Критерии оценивания ответа на теоретический вопрос

Оценка	Критерии
	1. Понимание и логика высказывания изученного материала 2. Представление взаимосвязей процесса и взаимосвязи теоретических модулей изучаемого предмета 3. Полнота данных ответов; 4. Аргументированность данных ответов; 5. Правильность ответов на вопросы;

«зачтено»	Полно и аргументировано даны ответы по содержанию задания. Обнаружено понимание материала, может обосновать свои суждения, применить знания на практике, привести необходимые примеры не только по учебнику, но и самостоятельно составленные. Изложение материала последовательно и правильно Ответы обучающегося удовлетворяют тем же требованиям, что и для оценки «отлично», но допускается 1-2 ошибки, которые сам же исправляет. Обучающийся демонстрирует знание и понимание основных положений данного задания, но: 1) излагает материал неполно и допускает неточности в определении понятий или формулировке правил; 2) не умеет достаточно глубоко и доказательно обосновать свои суждения и привести свои примеры; 3) излагает материал непоследовательно и допускает ошибки.
«не зачтено»	Демонстрация незнания ответа на соответствующее задание, допускаются ошибки в формулировке определений и правил, искажающие их смысл, беспорядочно и неуверенно излагается материал; отмечаются такие недостатки в подготовке, которые являются серьезным препятствием к успешному овладению последующим материалом.

Примерный тест (ИУК-4.1, ИУК-4.2**)**

1. Что такое обработка естественного языка?
   - a) Изучение естественных языков
   +b) Анализ и обработка текстов на естественных языках
   - c) Создание искусственных языков
   - d) Программирование на языках программирования

2. Какой из следующих методов не используется в обработке естественного языка?
   - a) Машинное обучение
   - b) Статистический анализ
   - c) Синтаксический анализ
   + d) Численные методы

3. Как называется задача определения частей речи в предложении?
   - a) Лемматизация
   - b) Синтаксический анализ
   + c) Морфологический анализ

4. Что такое лемматизация?
   + a) Приведение слова к его базовой форме
   - b) Анализ синтаксической структуры предложения
   - c) Определение частей речи слова
   - d) Построение семантической модели текста

5. Как называется задача определения смысла слова или выражения в контексте?
   - a) Синтаксический анализ
   - b) Морфологический анализ
   + c) Семантический анализ

6. Какой из следующих методов используется для извлечения информации из текста?
- a) Морфологический анализ
+ b) Именованные сущности
- c) Синтаксический анализ

7. Как называется задача определения именованных сущностей в тексте?
+ a) Распознавание именованных сущностей
- b) Морфологический анализ
- c) Синтаксический анализ

8. Что такое машинный перевод?
+ a) Автоматический перевод текста с одного языка на другой
- b) Анализ синтаксической структуры предложения
- c) Определение частей речи слова
- d) Построение семантической модели текста

9. Какой из следующих методов используется в машинном переводе?
+ a) Статистический анализ
- b) Морфологический анализ
- c) Синтаксический анализ

10. Какой метод обработки естественного языка используется для определения тональности текста?
- a) Морфологический анализ
+ b) Анализ эмоциональной окраски
- c) Синтаксический анализ

11. Что такое корпус в обработке естественного языка?
- a) Набор правил для анализа текста
- b) Структура данных для хранения текстовых данных
+ c) Большой набор текстов для обучения и тестирования моделей

12. Какие методы могут использоваться для классификации текстов?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Машинное обучение

13. Что такое стемминг?
+ a) Процесс нахождения основы слова
- b) Анализ синтаксической структуры предложения
- c) Определение частей речи слова
- d) Построение семантической модели текста

14. Какой метод используется для автоматического генерирования текста?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Генеративные модели

15. Как называется задача определения связей между словами в предложении?
- a) Морфологический анализ
+ b) Синтаксический анализ

- c) Семантический анализ

16. Какой метод используется для определения схожести текстов?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Сравнение векторных представлений

17. Что такое векторное представление слова?
+ a) Числовое представление слова в виде вектора
- b) Анализ синтаксической структуры предложения
- c) Определение частей речи слова
- d) Построение семантической модели текста

18. Какой метод используется для генерации резюме на основе текста?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Автоматическое реферирование

19. Что такое автоматическое реферирование?
+ a) Генерация краткого содержания текста
- b) Анализ синтаксической структуры предложения
- c) Определение частей речи слова
- d) Построение семантической модели текста

Критерии оценивания тестирования:

Оценка	Критерии
	1. Полнота выполнения тестовых заданий; 2. Своевременность выполнения; 3. Правильность ответов на вопросы; 4. Самостоятельность тестирования
«зачтено»	Выполнено более 85 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос Выполнено более 70 % заданий предложенного теста, в заданиях открытого типа дан полный, развернутый ответ на поставленный вопрос; однако были допущены неточности в определении понятий, терминов и др. Выполнено более 54 % заданий предложенного теста, в заданиях открытого типа дан неполный ответ на поставленный вопрос, в ответе не присутствуют доказательные примеры, текст со стилистическими и орфографическими ошибками.
«не зачтено»	Выполнено не более 53 % заданий предложенного теста, на поставленные вопросы ответ отсутствует или неполный, допущены существенные ошибки в теоретическом материале (терминах, понятиях).

Самостоятельная работа (разработка кода). Перечень формируемых компетенций: ИУК-4.2, ИУК-4.3, ИПК-1.2 , ИПК-2.2

1. Запишите звучащий файл. Воспользуетесь готовой моделью whisper (small) и преобразуйте записанный звучащий файл в текст

```
result = model.transcribe("/content/example.mp3")
print(result["text"])
```

2. Основываясь на технологиях TTS преобразуйте письменный текст в устную речь

```
import numpy as np
from logmmse import logmmse

enhanced = logmmse(np.array(audio[0]), sample_rate, output_file=None,
initial_noise=1, window_size=160, noise_threshold=0.15)
display(Audio(enhanced, rate=sample_rate))
```

Контрольная работа (разработка кода): Перечень формируемых компетенций: ИУК-4.2, ИУК-4.3, ИПК-1.2 , ИПК-2.2

1. Скачайте файл-корпус (csv) с отзывами на товары

2. Дообучите модель генерации gpt-3 таким образом, чтобы на выходе модели генерировались отзыв.

```
# Trainer object instead of deffault learning pytorch loop
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    data_collator=data_collator
)
trainer.train()
generate('хорошая куртка', max_length=250)
```

Критерии оценивания самостоятельных и контрольных работ основываются на проверке алгоритма, структуры и корректности кода и оцениваются по следующим параметрам:

Критерии оценивания самостоятельной и контрольных работ:

Оценка	Критерии
	1. Наличие или отсудив ошибок в коде 2. Полнота решения практических задач 3. Своевременность выполнения; 4. Умение работать с собственным материалом 5. Совмещать базовые технологии ИИ со своей предметной областью
«зачтено»	Основные требования к коду выполнены. Код работает без ошибок, возможны синтаксические недочеты, не везде присутствуют комментарии;
«не зачтено»	Задача не решена, обнаруживается существенное непонимание проблемы

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

Зачет по дисциплине принимается на основе достижения рубежных показателей в рейтинге (не ниже 55 баллов), при выполнении практических заданий (ИПК-1.2, ИПК-2.2), тестов (ИУК-4.1, ИУК-4.2, ИУК-4.3), посещения занятий.

Рейтинг, баллы

1 – присутствие на лекции

1 – присутствие на занятии

1-3 – работа на занятии

1-36 – подготовка к занятию и работа на практическом занятии (в т.ч. выполнение домашнего задания)

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

Первая часть. Решение тестовых задач (УК-4, ПК-1, ИУК-4.1):

Тест №1

УК-4

1. Какой метод используется для определения языка текста?
   - a) Морфологический анализ
   - b) Синтаксический анализ
   + c) Языковая модель

2. Что такое языковая модель?
   - a) Анализ синтаксической структуры предложения
   - b) Определение частей речи слова
   + c) Вероятностная модель последовательности слов

3. Какие методы используются для обработки текстов на естественных языках?
   - a) Морфологический анализ
   - b) Синтаксический анализ
   - c) Семантический анализ
   + d) Все вышеперечисленные

4. Что такое частотный анализ слов?
   + a) Анализ распределения слов по частоте в тексте
   - b) Анализ синтаксической структуры предложения
   - c) Определение частей речи слова
   - d) Построение семантической модели текста

ИУК-4.1

5. Какой метод используется для определения схожести документов?
   - a) Морфологический анализ
   - b) Синтаксический анализ
   + c) Модель TF-IDF

6. Что такое модель TF-IDF?
   - a) Анализ синтаксической структуры предложения
   - b) Определение частей речи слова
   + c) Модель, учитывающая частотность и обратную частотность слова в тексте

7. Какой метод используется для определения ключевых слов в тексте?
   - a) Морфологический анализ
   - b) Синтаксический анализ
   + c) Извлечение ключевых слов

8. Что такое извлечение ключевых слов?
   + a) Процесс определения наиболее значимых слов в тексте
   - b) Анализ синтаксической структуры предложения

- c) Определение частей речи слова
- d) Построение семантической модели текста

9. Какой метод используется для определения семантической близости между словами?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Векторные представления слов

10. Что такое семантическая близость слов?
- a) Анализ синтаксической структуры предложения
- b) Определение частей речи слова
+ c) Степень сходства по значению между словами

11. Какой метод используется для определения семантической связи между словами?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Семантический анализ

12. Что такое семантический анализ текста?
- a) Анализ синтаксической структуры предложения
- b) Определение частей речи слова
+ c) Анализ значения слов и выражений в тексте

13. Какой метод используется для определения синонимов и антонимов слова?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Тезаурус

14. Что такое тезаурус?
+ a) Словарь, содержащий синонимы и антонимы слов
- b) Анализ синтаксической структуры предложения
- c) Определение частей речи слова
- d) Построение семантической модели текста

15. Какой метод используется для определения смысла слова в контексте?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Векторные представления слов

ПК-1

16. Что такое морфологический анализ текста?
+ a) Анализ грамматической структуры слов
- b) Анализ синтаксической структуры предложения
- c) Определение частей речи слова
- d) Построение семантической модели текста

17. Какой метод используется для определения грамматической структуры предложения?
- a) Морфологический анализ
+ b) Синтаксический анализ
- c) Семантический анализ

18. Что такое синтаксический анализ текста?
- a) Анализ грамматической структуры слов

+ b) Анализ структуры предложения и связей между словами
- c) Определение частей речи слова
- d) Построение семантической модели текста

19. Какой метод используется для определения структуры предложения?
- a) Морфологический анализ
+ b) Синтаксический анализ
- c) Семантический анализ

ИПК-2.2

20. Что такое семантический анализ предложения?
- a) Анализ грамматической структуры слов
- b) Анализ структуры предложения и связей между словами
+ c) Анализ значения предложения в контексте

21. Какой метод используется для определения связей между предложениями в тексте?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Семантический анализ

22. Что такое частотный анализ словосочетаний?
- a) Анализ распределения словосочетаний по частоте в тексте
- b) Анализ синтаксической структуры предложения
- c) Определение частей речи словосочетания
+ d) Анализ частотности встречаемости слов в определенных контекстах

23. Какой метод используется для извлечения информации из текста?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Извлечение информации

24. Что такое извлечение информации?
+ a) Процесс извлечения структурированной информации из текста
- b) Анализ синтаксической структуры предложения
- c) Определение частей речи слова
- d) Построение семантической модели текста

25 Какой метод используется для автоматического ответа на вопросы?
- a) Морфологический анализ
- b) Синтаксический анализ
+ c) Вопросно-ответные системы

26. Что такое вопросно-ответная система?
- a) Анализ синтаксической структуры предложения
- b) Определение частей речи слова
+ c) Система, способная отвечать на вопросы по тексту

27. Какой метод используется для определения семантической роли слов в предложении?
- a) Морфологический анализ
+ b) Синтаксический анализ
- c) Семантический анализ

Вторая часть. Решение практических задач (ИУК-4.2, ИУК-4.3, ИПК-1.2, ИПК-2.2):

Задача 1 ИУК-4.2

Создайте корпус текстов, содержащий факты. Требования: объем не менее 10 000 токенов

Задача 2 ИУК-4.2

В созданном корпусе найдите сочетания по частям речи, согласованных по роду числу и падежу. Для этого необходимо создать контекстно свободные грамматики в программе Томита парсер. Название факта должно отображать выбранные тип сочетаний частей речи. Например, в случае, если выбрано сочетание существительное и прилагательное, тогда факт должен называться «СущПрил» в config.proto. Сами грамматики прописаны в файле .cxx.

Задача 3 ИУК-4.3

В проограмме Tomita parser создайте грамматики, извлекающие цепочки фактов результата футбольного матча.

Пример грамматик файла .cxx:

#encoding "utf-8" // сообщаем парсеру о том, в какой кодировке написана грамматика

#GRAMMAR_ROOT S // указываем корневой нетерминал грамматики

National -> Noun<gram="geo"> | Noun<c-agr[1]> Noun<gram="geo", c-agr[1]>;

Club -> Noun<h-reg1, quoted>;

Club -> Word<h-reg1, l-quoted> Word* Word<r-quoted>;

Team -> Club | National;

Result -> Verb<kwtype="result_verb"> interp(ResultFact.Result::norm="nom,sg");

Score -> AnyWord<wff=/[0-9]:[0-9]/> interp(ResultFact.Score);

S -> Team interp(ResultFact.FirstTeam) AnyWord* Result AnyWord* Team interp(ResultFact.SecondTeam) AnyWord* Score;

S -> Team interp(ResultFact.FirstTeam) AnyWord* Team interp(ResultFact.SecondTeam) AnyWord* Result AnyWord* Score;

S -> Team interp(ResultFact.FirstTeam) AnyWord* Team interp(ResultFact.SecondTeam) AnyWord* Score;

//вывод в таблицу

Извлеките из текста цепочки (фрейм) фактов: команда1 – команда2 – результат – счет

Текст:

Сборная Бразилии в товарищеском матче победила Японию – 3:1. По голу у южноамериканцев забили Неймар, Марселу и Габриэл Жезус.

Хозяева одержали победу со счётом 2:0. «Трёхцветные» открыли счёт на 18-й минуте, когда точным ударом отметился Антуан Гризманн.

Польша и Уругвай сыграли вничью в товарищеском матче - 0:0. В еще одном поединке Бельгия и Мексика также не смогли выявить победителя - 3:3.

Напомним, 2 ноября «Марсель» на выезде играл против португальской «Виктории Гимарайнш» (0:1). Во время разминки перед матчем в адрес Эвра звучали оскорбления от болельщиков «Марселя».

Сборная России по футболу проиграла команде Аргентины в товарищеском матче (0:1). Единственный гол в матче забил форвард «Манчестер Сити» Серхио Агуэро.

Сегодня футболисты саратовского «Сокола» уступили в Красногорске (Московская область) «Зоркому» 1:5.

Юношеская сборная России (U19) проиграла сверстникам из Румынии со счетом 1:2 в матче 1-го квалификационного раунда Евро-2018.

Глушаков дебютировал в сборной России 29 марта 2011 года в товарищеском матче с Катаром (1:1).

Россия победила Англию со счетом 5-0.

Россия продула Англии со счетом 5:0.

Сборная Норвегии проиграла соперникам из Англии со счетом 4-0.
Россия разошлась миром со Сборной Англии со счетом 0:0.

Задача 4 ИУК-4.3
Измените грамматики или газиттер (файл .gzt) таким образом, чтобы извлекались пропущенные цепочки фактов:
Переделайте код таким образом, чтобы извлекались факты из предложений:

1. Россия продула Англии со счетом 1:0.

2. Россия победила Англию со счетом 5-1.
Вывод html-фактов

ResultFact			
FirstTeam	**Result**	**SecondTeam**	**Score**
сборная Бразилии	победа	Япония	3:1
польша	ничья	Уругвай	0:0
Бельгия	ничья	Мексика	3:3
Марсель		Виктория	0:1
сборная России	поражение	Аргентина	0:1
Сокол	поражение	Красногорск	1:5
сборная России	поражение	Румыния	1:2
сборная России		Катар	1:1
россия	поражение	Англия	5:0
россия	ничья	Сборная Англии	0:0

Задача 5 ИПК-1.2
Как видно из таблицы из задачи 8, существует три состояния результата: поражение, ничья и победа. Какие еще есть состояния?
Ответ: отмена/перенос матча
Дополните газиттер таким образом, чтобы формальные грамматики учитывали четвертый результат.
Задача 6. ИПК-1.2, Техническое оформление результатов исследования (втоматизация оформления документа по ГОСТу в MS Word) 1. Создайте стили по ГОСТу и примените к тексту. 2. Создайте заголовки структуры диссертации и добавьте автоматическое содержание. 3. Добавьте изображения, используя перекрестные ссылки 4. Сделайте шаблон титульного листа диссертации
Задача 7. ИПК-2.2 Создайте алгоритм распознавания 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)
```

Задача 8. Обучите генеративную модель (трансформеры)

```
!pip install accelerate>=0.20.
 !pip install -U accelerate
!pip install -U transformers
import warnings
import os

os.environ["WANDB_DISABLED"] = "true"
warnings.filterwarnings("ignore")
import torch

# the if-condition is adapted to devices on M1/M2
device = 'cuda' if torch.cuda.is_available() else ('mps' if
torch.backends.mps.is_available() else 'cpu')
device
from transformers import GPT2LMHeadModel, GPT2Tokenizer
from transformers import TrainingArguments, Trainer
from transformers import DataCollatorForLanguageModeling, TextDataset
from transformers import AdamW, get_cosine_schedule_with_warmup
model_name_or_path = 'ai-forever/rugpt3small_based_on_gpt2'

# tokenizer based on GPT2 for text preprocessing
tokenizer = GPT2Tokenizer.from_pretrained(model_name_or_path)

# loading a pre-trained model based on GPT2
```

```
model = GPT2LMHeadModel.from_pretrained(model_name_or_path).to(device)
# sequence continuation generation function
# the parameters were selected during the experiments
def generate(prompt, do_sample=True, num_beams=2, temperature=1.8,
top_p=0.85, max_length=175):

    input_ids = tokenizer.encode(prompt, return_tensors="pt").to(device)

    model.eval()
    with torch.no_grad():
        out = model.generate(input_ids,
                             do_sample=do_sample,
                             num_beams=num_beams,
                             temperature=temperature,
                             top_p=top_p,
                             max_length=max_length,
                             )

    print(list(map(tokenizer.decode, out))[0])
# generation example before fine-tuning
generate('Люблю грозу в начале мая')
train_dataset = TextDataset(tokenizer=tokenizer,file_path='/content/data
(1).txt',
                            block_size=64)
data_collator = DataCollatorForLanguageModeling(tokenizer=tokenizer,
                                                mlm=False)
# training args for fine-tuning
training_args = TrainingArguments(
    output_dir = "./finetuned_model",
    overwrite_output_dir = True,
    num_train_epochs = 4,
    gradient_accumulation_steps = 2,
    fp16 = True,
    per_device_train_batch_size = 8,
    learning_rate = 0.0002,
    optim = 'adafactor',
    lr_scheduler_type = 'cosine'
)
# training args for fine-tuning
training_args = TrainingArguments(
    output_dir = "./finetuned_model",
    overwrite_output_dir = True,
    num_train_epochs = 4,
    gradient_accumulation_steps = 2,
    fp16 = True,
    per_device_train_batch_size = 8,
    learning_rate = 0.0002,
    optim = 'adafactor',
    lr_scheduler_type = 'cosine'
)
# Trainer object instead of deffault learning pytorch loop
```

```
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    data_collator=data_collator
)
# let's train!
trainer.train()
generate('Люблю грозу в начале мая', max_length=250)
```

Информация о разработчиках

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