**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное образовательное учреждение**

**высшего образования**

«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ ФИЛОЛОГИИ И МЕЖКУЛЬТУРНОЙ

КОММУНИКАЦИИ

ВЫСШАЯ ШКОЛА \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кафедра \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Направление: \_\_\_\_\_\_\_\_ – \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

шифр наименование

Профиль: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

наименование

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

МЕТОДЫ ОЦЕНКИ СЛОЖНОСТИ ТЕКСТА

Студент \_\_\_ курса

Группы \_\_\_\_\_\_\_

«\_\_\_» \_\_\_\_\_\_\_\_\_\_2019 г. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (Фамилия и инициалы)

Научный руководитель

\_\_\_\_\_\_\_\_\_\_,\_\_\_\_\_\_\_\_\_\_\_

Уч.степ. уч.зван.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_2019 г. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(Фамилия и инициалы)

Заведующий кафедрой

\_\_\_\_\_\_\_\_\_\_,\_\_\_\_\_\_\_\_\_\_\_

Уч.степ. уч.зван.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_2019 г. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(Фамилия и инициалы)

Казань – 2019

СОДЕРЖАНИЕ

Введение

Глава 1. Теоретические основы исследования сложности текста

1.1. Определение основных понятий и терминов, связанных с

оценкой сложности текста

1.2. Обзор литературы о влиянии многозначных слов на

сложность текста

1.3. Автореферирование как метод измерения сложности текста с использованием программных решений

1.4 Обзор вопросно-ответных систем для оценки сложности текстов

II. Методология исследования

2.1. Выбор корпуса текстов для исследования и подготовка данных

2.2. Разработка и описание алгоритмов для вычисления метрик многозначности и сложности текста

2.3. Методы автореферирования и их внедрение в исследование

2.4. Разработка QA системы для оценки сложности текста

2.5. Анализ результатов сложности текстов на основе метрик многозначности

2.6. Анализ на основе автореф

2.7 Сравнение методик

2.8. Ограничения и возможные направления будущих исследований

Заключение

Список литературы

**Введение**

Задача определения сложности текста предполагает разработку методов и алгоритмов, позволяющих автоматически оценивать уровень сложности понимания текста читателями. Эта проблема имеет большое значение в различных областях, включая образование, издательское дело, бизнес, юриспруденцию, медицину, контент-маркетинг и информационные системы.

В то же время проблема определения сложности текста не нова. Формально сложность текста можно определить, как количество лексических единиц, синтаксических структур и прагматических ситуаций, которые вызывают трудности при чтении, как сумму всех элементов, влияющих на понимание, скорость чтение и интерес. Однако для каждого человека этот набор будет индивидуальным.

В прошлом, сложность текста определялась с помощью экспертной оценки, что требовало значительных ресурсов и времени. Кроме того, экспертная оценка может быть субъективной и не всегда точной.

В последние годы с появлением больших объемов данных и развитием машинного обучения, стали возможны новые подходы к определению сложности текста. Методы машинного обучения, такие как линейная регрессия, деревья решений, нейронные сети и SVM, позволяют автоматически определять сложность текста на основе большого количества лингвистических и статистических характеристик.

За всю историю изучения сложности текста, сложилось два традиционных подхода:

1. Традиционный, основанный на различных формулах удобочитаемости.
2. Вычисление сложности на основе машинного обучения.

**Актуальность** данной работы в том, что современные методы машинного обучения позволяют получать более точные результаты и учитывать большое количество разнообразных характеристик текста, таких как длина предложений, уровень лексического разнообразия, частота употребления слов и многие другие. Во-вторых, разработанные модели можно легко интегрировать в удобные для пользователей формы, что делает их доступными и полезными как в образовательных, так и в коммерческих приложениях. Этот подход обеспечивает удобство и эффективность в оценке сложности текстовых материалов.

**Проблема исследования:** В современном информационном обществе существует потребность в эффективных методах и инструментах для оценки сложности текстов, которые позволили бы более точно определить уровень сложности и адаптировать контент для различных аудиторий. Однако отсутствует стандартизированный и универсальный метод измерения сложности текстов, что создает препятствия для разработки приложений, способных оценивать и адаптировать тексты на основе их сложности.

Эта проблема включает в себя следующие аспекты:

* Отсутствие стандартизированного метода измерения сложности текстов: Не существует единого и признанного стандарта для определения сложности текстов, что затрудняет сравнение и анализ текстов на разных языках и темах.
* Необходимость автоматизированных инструментов: В современном мире с растущим объемом информации необходимы инструменты, способные автоматически определять сложность текстов, чтобы облегчить образовательный и коммерческий процессы.
* Зависимость оценки сложности от контекста: Оценка сложности текста может зависеть от контекста и целевой аудитории, исключая возможность создания универсальных моделей.

**Методологической базой** стали исследования по изучению сложности текста и статистических методов анализа текстов Солнышкиной М.И, Соловьева В.Д., Иванова В., труды по изучению сложности текстов на основе алгоритмов компьютерного вычисления и машинного обучения Кутузова А.Б., Оакес М.П., ,Фенг, Л., Янче, М., Хенерфаут, М., Эльхадад Н., Банерджи С., Педерсен Т., Маквикар М.

**Объектами исследования** являются методы измерения сложности текста.

**Предметом исследования** является точность и полнота методов измерения сложности текста.

**Целью данной работы** является сравнение эффективности и точности указанных методов с целью определения, какой из них лучше всего подходит для оценки сложности текста в нашем контексте. В частности, будут исследованы следующие алгоритмы и методы:

* Метод измерения сложности на основе многозначности.
* Метод измерения сложности на основе автореферирования (информационной энтропии).
* Метод измерения сложности с применением искусственного интеллекта.

В рамках данной работы предполагается выполнение следующих **задач**:

1. Подготовка корпуса: Провести подготовку корпуса текстов к дальнейшему анализу, включая этап предварительной обработки (preprocessing).
2. Реализация алгоритмов: Разработать и реализовать алгоритмы измерения сложности текста на языке программирования Python, которые будут использоваться для оценки текстов.
3. Анализ данных: Проанализировать данные, включая тексты и оценки сложности, с целью выявления закономерностей и особенностей.
4. Сравнение методов: Сравнить различные методы измерения сложности текста, включая методы на основе многозначности, информационной энтропии, и с использованием искусственного интеллекта.
5. Формулирование выводов: Сформулировать и обосновать полученные выводы о наиболее эффективных методах измерения сложности текста в данном контексте и их применимости.

**Материалом исследования** послужил корпус учебников по русскому языку для учащихся 2, 3, 4, 5, 6, 7 и 9 классов.

В процессе исследования были использованы следующие методы:

-метод количественного анализа

-сопоставительный анализ

-лексико-семантический, морфологический и синтаксический методы анализа языкового материала.

**Гипотеза исследования** - существует корреляция между оценками сложности текстов, полученными с использованием различных методов.

Положения, выносимые на защиту:

1. Сложность текста действительно увеличивается с увеличением числа многозначных слов в нем. Это подтверждается анализом корпуса учебников, где тексты с большим количеством многозначных слов имеют более высокую сложность по сравнению с текстами, содержащими меньше таких слов.

2. Была обнаружена положительная корреляция между сложностью текста и годом обучения. Это означает, что с увеличением года обучения учащихся тексты, предназначенные для них, становятся более сложными. Этот вывод подкреплен анализом учебников разных классов.

3. Сложные тексты оказываются более сложными для реферирования. Это было продемонстрировано экспериментально, где участники должны были реферировать тексты разной сложности. Сложные тексты требовали больше усилий для сжатия и извлечения основных идей.

4. Модель искусственного интеллекта, используемая в исследовании, показала хорошую точность в оценке сложности текста. Это означает, что алгоритмы и методы, разработанные в рамках исследования, могут быть использованы для точной оценки сложности текстов в образовательных и коммерческих приложениях.

**Научная новизна** состоит в исследовании раскрывает потенциал искусственного интеллекта в оценке сложности текстов и внедрении этой функциональности в доступное приложение для широкой аудитории. Это новаторский подход, который может значительно облегчить оценку и понимание сложности текстов разного уровня.

**Теоретическая значимость:** исследование вносит вклад в изучение и понимание процессов анализа сложности текста, а также в разработку новых методов и алгоритмов для этой цели. Это помогает расширить теоретическую базу в области анализа текстов и их оценки.

**Практическая значимость:** полученные результаты имеют непосредственное применение в образовании и бизнесе. Оценка сложности текстов может быть полезной для улучшения образовательных материалов, создания контента, который легче воспринимается целевой аудиторией, и для разработки приложений, которые помогают пользователям более точно определять сложность текстов и адаптировать их под свои потребности.

Представленная работа прошла апробацию на 74-ой научно-техничкской конференции учащихся, студентов и магистрантов БГТУ в 2023 году, а также на конференции “Resolving complexities in English Studies: Theoretical and Pedagogical Perspectives” (Самарканд).

Магистерская диссертация на тему «Сравнение методов оценки сложности текста на основе корпуса учебников по русскому языку» состоит из введения, двух глав, заключения и библиографического списка. Во введении обосновывается выбор темы и ее актуальность, определяются объект, предмет, цели, задачи, методы исследования, научная новизна, теоретическая и практическая значимость работы.

**Глава 1. "Теоретические основы исследования сложности текста"** представляет собой теоретическое введение и обзор основных концепций и методов, используемых в исследовании сложности текста на основе корпуса учебников по русскому языку. Глава включает в себя следующие разделы:

1.1. Определение основных понятий и терминов, связанных с оценкой сложности текста.

В данном разделе проводится анализ ключевых терминов и понятий, связанных с оценкой сложности текста. Важные термины, такие как "сложность текста", "многозначные слова", "автореферирование" и другие, подробно определяются и объясняются.

1.2. Обзор литературы о влиянии многозначных слов на сложность текста.

В этом разделе проводится обзор литературных исследований и научных работ, посвященных влиянию многозначных слов на сложность текста. Анализируются результаты предыдущих исследований, выделяются основные тенденции и выводы.

1.3. Автореферирование как метод измерения сложности текста с использованием программных решений.

В данном разделе описывается методика автореферирования текста и ее применение в оценке сложности текста. Также представляются программные решения и алгоритмы, используемые для автоматического автореферирования.

1.4. Обзор вопросно-ответных систем для оценки сложности текстов.

В этом разделе проводится обзор существующих вопросно-ответных систем, которые могут быть использованы для оценки сложности текстов. Рассматриваются методы, основанные на вопросно-ответном взаимодействии с текстом, и их применимость к задаче оценки сложности.

Глава 1 формирует теоретическую основу для последующих исследований, предоставляя читателю понимание ключевых понятий и методов, используемых в работе, а также ознакамливая с предыдущими исследованиями в данной области.

Глава 2. "Методология исследования" представляет собой описание методологии, примененной в рамках данного исследования для оценки сложности текста на основе корпуса учебников по русскому языку. Глава включает в себя следующие разделы:

2.1. Выбор корпуса текстов для исследования и подготовка данных.

В данном разделе описывается процесс выбора корпуса учебников, который послужил основой для исследования. Также рассматривается подготовка данных, включая сбор текстов, их предобработку и структурирование для последующего анализа.

2.2. Разработка и описание алгоритмов для вычисления метрик многозначности и сложности текста.

В этом разделе представляются разработанные алгоритмы и методы для вычисления метрик, связанных с многозначностью и сложностью текста. Детально описывается, какие метрики были выбраны, какие параметры использовались, и какие вычислительные методы были применены для расчетов.

2.3. Методы автореферирования и их внедрение в исследование.

В данном разделе описывается методика автореферирования текста, а также процесс ее интеграции в исследование. Рассматриваются программные решения, используемые для автоматического создания рефератов текстов и их применение к корпусу учебников.

Глава 2 предоставляет детальное описание методологии, которая была применена в исследовании для оценки сложности текста, начиная с выбора данных и заканчивая разработкой алгоритмов и инструментов для анализа текстового материала.

В Заключении приведены выводы, к которым мы пришли в результате данного исследования.

**ГЛАВА I. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИССЛЕДОВАНИЯ СЛОЖНОСТИ ТЕКСТА**

**1.1. Определение основных понятий и терминов, связанных с**

**оценкой сложности текста**

В научной среде до сих пор есть смешивание трёх понятий: сложности, трудности и читабельности текста. Все три характеристики определяются с помощью математических формул и компьютерных программ. Ознакомившись с рядом работ таких ученых как М.А. Джаст, П.А. Карпентер и К. Код, можно сказать, что термин “сложность” определяется больше характеристиками самого текста, тогда как “трудность” определяется на основе эмпирических данных о восприятии текста читателем. Термины “читабельность” и “удобочитаемость” являются вариантами перевода английского термина *readability*. А. Ребер определяет читабельность как мера, доступности для понимания письменного текста, определяемая анализом ряда факторов, включая синтаксическую сложность, лексику, выраженность темы, связность тем и т.п. Индекс читабельности определяется на основе количественных показателей текста и отражает степень понимания текста читателем, а также сложность самого текста.

Сложность текста понятие многогранное, оно включает в себя как количественные и качественные параметры текста, так и уровень подготовки читателя. Лингвистическая составляющая текста, в отличии от уровня подготовки читателя, который складывается из способностей, компетенций и психотипа читателя,является объективной величиной. Именно ее мы и будем анализировать в нашей работе. Лингвистическая составляющая текста включает в себя различные параметры, например: количество слов в тексте, количество слогов, длина предложения, наличие многозначных слов, метафор, сравнений, абстрактность изложения и ясность структуры текста и некоторые другие.

Традиционно разделяют два подхода к автоматическом определению сложности текста:

1. Традиционный (основанный на формулах удобочитаемости)

Оценка сложности текста с использованием формул удобочитаемости представляет собой распространенный подход, который позволяет количественно оценить уровень доступности текста для различных категорий аудитории. В рамках этого подхода были разработаны различные формулы и индексы, основанные на лексических и структурных характеристиках текста.

Наиболее популярные методы, основанные на формулах удобочитаемости:

* Индекс Флеша-Кинсайда (Flesch-Kincaid Readability Index):

Индекс Флеша-Кинсайда является одним из наиболее известных и широко используемых методов оценки удобочитаемости текста. Эта формула оценивает сложность текста на основе средней длины предложений и средней длины слова. Результат вычислений преобразуется в числовое значение, которое соответствует уровню образования или возрасту.

* Индекс Фога (Fog Index):

Индекс Фога также используется для оценки удобочитаемости текста. Он основан на сложности слов и сложности предложений. Для вычисления индекса Фога, необходимо определить средний процент сложных слов (слов, содержащих более трех слогов) и среднюю длину предложений. Индекс Фога также преобразуется в числовое значение, показывающее уровень сложности текста.

* Индекс Смога (SMOG Index):

Индекс Смога измеряет уровень сложности текста на основе количества сложных слов в тексте. Сложные слова определяются как слова, имеющие более двух слогов и содержащие слог с ударением на третьем или более высоком слоге. Индекс Смога также учитывает количество предложений в тексте и дает оценку, соответствующую уровню образования.

* Индекс Дейла-Челлера (Dale-Chall Readability Formula):

Индекс Дейла-Челлера оценивает удобочитаемость текста, учитывая сложность лексики. В формуле используется список "базовых" слов, которые считаются известными для человека с определенным уровнем образования. Индекс вычисляется на основе процента слов в тексте, не входящих в этот список.

Математические формулы для определения сложности текста основаны на таких его параметрах как средняя длина слова, средняя длина предложения, количество длинных слов и т.д. Результатом вычисления данных формул обычно является число, соответствующее возрасту или году обучения. Эти методы оценки сложности текста на основе формул удобочитаемости предоставляют простой и быстрый способ количественной оценки уровня доступности текста. Однако следует помнить, что они могут быть менее точными в случае использования для текстов на других языках или текстов с специализированной лексикой. При использовании данных формул для русского языка необходим пересчет. Также явным недостатком данного метода является то, что формулы не учитывают лексико-грамматическую составляющую текста и средства связи.

1. Вычисление сложности на основе машинного обучения

В рамках машинного обучения также можно решить проблему автоматического определения сложности текста. Это классическая задача построения предсказательной модели на основании обучения на тренировочном корпусе текстов и наборе признаков.

Оценка сложности текста с использованием методов машинного обучения представляет собой более гибкий подход, который позволяет учесть более широкий набор параметров и контекстов. Этот подход включает построение предсказательных моделей, основанных на анализе больших объемов текстовой информации и выявлении паттернов, влияющих на уровень сложности.

Подходы к вычислению сложности с использованием машинного обучения:

1. Обучение с учителем:

Этот подход включает создание моделей, которые обучаются на тренировочном наборе текстов с различными уровнями сложности. В качестве признаков могут использоваться различные лингвистические параметры, такие как длина предложений, сложность лексики, наличие многозначных слов и другие. Модель прогнозирует уровень сложности для новых текстов на основе выявленных зависимостей.

1. Обучение без учителя:

Этот подход предполагает использование методов кластеризации и снижения размерности для анализа текстов и выявления внутренних структур. Тексты группируются по общим характеристикам, и на основе этого группировки можно оценить уровень сложности.

К преимуществам методов машинного обучения можно отнести:

1. Учет множества факторов:

Методы машинного обучения позволяют учесть более широкий набор параметров, включая лексические, структурные, семантические и контекстные характеристики текста. Это может привести к более точной оценке сложности.

2. Адаптация к разным языкам и стилям:

Модели машинного обучения могут быть обучены на текстах разных языков и стилей, что делает их более универсальными и адаптивными.

3. Улучшение с опытом:

Модели машинного обучения могут постепенно улучшаться с увеличением объема данных и опыта обучения. Это позволяет создавать более точные и предсказательные модели.

Однако, методы машинного обучения имеют следующие ограничения:

1. Необходимость большого объема данных:

Для эффективного обучения моделей машинного обучения требуется большой объем разнообразных текстов. Недостаток данных может снизить точность моделей.

2. Сложность интерпретации:

Некоторые модели машинного обучения, такие как нейронные сети, могут быть сложными для интерпретации. Это может затруднить понимание, какие конкретные характеристики влияют на оценку сложности.

3. Зависимость от качества данных:

Качество данных, на которых обучаются модели, существенно влияет на результаты. Несбалансированные или неточные данные могут привести к недостоверным оценкам сложности.

В целом, методы вычисления сложности текста на основе машинного обучения предоставляют более глубокий анализ и более точные результаты, чем традиционные формулы удобочитаемости. Однако их успешное применение требует сочетания экспертных знаний, адекватных данных и опыта работы с моделями машинного обучения.

Таким образом, анализ сложности текста является важной задачей, объединяющей лингвистические, психологические и вычислительные аспекты, и может быть решен разными методами в зависимости от поставленных целей и доступных ресурсов.

**1.2 Обзор литературы о влиянии многозначных слов на**

**сложность текста**

Одним из ключевых факторов, влияющих на сложность текста, является наличие многозначных слов. Многозначные слова, или полисемия, представляют собой явление, когда одно слово имеет несколько различных значений в контексте. Это явление может вносить дополнительную сложность в понимание текста, особенно для читателей, которые не всегда могут правильно интерпретировать выбранное значение слова в данном контексте.

Полисемия может затруднять чтение и понимание текста, так как читатель должен правильно выбирать подходящее значение слова в зависимости от контекста. Это может вызвать путаницу и снизить общую читаемость текста, особенно если многозначные слова используются важных моментах или ключевых аргументах текста.

В работе "WordNet: An Electronic Lexical Database", Джордж Миллер исследует сложности, связанные с лексической семантикой и многозначностью слов. Эта работа, представляющая собой пионерский вклад в лингвистику, подчеркивает важность контекстуального анализа при интерпретации многозначных слов.

Миллер описывает, как многозначные слова создают "парадокс выбора" для читателей: при наличии нескольких значений одного и того же слова, необходимо определить, какое из этих значений наиболее подходит для данного контекста. Он отмечает, что выбор конкретного значения многозначных слов зависит от множества факторов, включая ближайшие соседние слова, структуру предложения и тему текста.

Эта работа предоставляет важные инсайты в то, как многозначные слова могут затруднить понимание текста. Миллер подчеркивает, что контекст играет решающую роль в определении значения многозначных слов. Он представляет концепцию "семантического окна", который определяет контекстуальные границы, в пределах которых читатель должен выбрать определенное значение слова.

Исследования Джорджа Миллера позволяют лучше понять сложности, связанные с многозначностью, и обогатить методику анализа текстовой сложности. Его работа дает научное обоснование не только для изучения лексической семантики, но и для создания более точных методов оценки текстовой сложности, учитывающих воздействие многозначных слов на понимание и интерпретацию текстов.

Миллер подчеркивает, что решение "парадокса выбора" в случае многозначных слов требует от читателей активного взаимодействия с текстом и глубокого понимания контекста. Эта идея становится важным фундаментом для дальнейших исследований в области анализа текстовой сложности и лингвистической семантики.

Исследования в области лексикографии и психолингвистики, проведенные Штефаном Е. Дейлом и Дэвидом Гласснером в их работе "Polysemy and Word Meaning", проливают свет на важность понимания многозначных слов в контексте анализа текстовой сложности. Они подчеркивают, что многозначные слова могут оказывать существенное влияние на восприятие и понимание текста.

В этой работе выявляется, что многозначность в словах может вызвать недоразумения и путаницу у читателей. Выбор подходящего значения многозначного слова становится сложной задачей, требующей обширного контекстуального анализа. Дейл и Гласснер описывают, как читатели могут столкнуться с дилеммой, когда разные значения многозначного слова могут быть легко ассоциированы с контекстом, но определить, какое именно значение более подходящее, может быть трудной задачей.

Эдвард Финкель и Ричард Андерсон, авторы статьи "Reading and understanding: Teaching from the perspective of artificial intelligence", расширяют это понимание, отмечая, что тексты с высокой степенью сложности часто содержат многозначные слова. Эти слова могут потребовать дополнительных усилий от читателей для точного и полного понимания. Такие тексты могут оказаться вызовом даже для опытных читателей, так как требуют более глубокого анализа и тщательной интерпретации.

Исследования Дейла, Гласснера, Финкеля и Андерсона подчеркивают, что многозначные слова могут быть значительным источником сложности в тексте. Они поднимают важный вопрос о том, как правильно учитывать влияние многозначности при оценке текстовой сложности. Определение контекста и выбор определенного значения многозначных слов становятся центральными аспектами при анализе текстов, особенно в случае более сложных текстов, где такие слова могут значительно усложнить задачу понимания.

Важно подчеркнуть, что воздействие многозначных слов на сложность текста может меняться в зависимости от контекста. Как отмечает Жан А. Бойд в работе "The Many Faces of Polysemy", использование многозначных слов может либо обогатить текст, придавая ему глубину и многомерность, либо создать путаницу, особенно если такие слова используются в ключевых моментах или аргументах текста.

Как подмечено Ричардом Якубинским и Каролиной Риверос в исследовании "Lexical Ambiguity and Reading Comprehension", многозначность может взаимодействовать с другими лингвистическими параметрами текста. Например, в текстах с более длинными предложениями, где создается больше возможностей для контекстуальных зависимостей, интерпретация многозначных слов может оказаться еще более сложной.

Исследования показывают, что в текстах с высоким уровнем сложности часто встречаются многозначные слова, которые могут требовать дополнительных усилий для правильного понимания. Важно отметить, что воздействие полисемии на сложность текста может быть контекстуальным. В некоторых случаях использование многозначных слов может придавать тексту глубину и смысловую насыщенность, но в других случаях оно может осложнить понимание и снизить доступность текста для широкой аудитории.

Следует также учитывать, что в реальных текстах полисемия может взаимодействовать с другими лингвистическими параметрами, такими как длина предложений, сложность лексики и структура текста. Тексты с более длинными предложениями могут иметь больше контекстуальных зависимостей для интерпретации многозначных слов, что может усугубить сложность чтения.

В итоге, наличие многозначных слов в тексте действительно влияет на его сложность и может быть одним из важных аспектов, рассматриваемых при анализе текстовой сложности. Дальнейшие исследования в этой области могут помочь более точно определить, как именно полисемия влияет на понимание и восприятие текста, а также как этот фактор может быть учтен в методах оценки и анализа текстовой сложности.

**1.3 Автореферирование как метод измерения сложности текста с использованием программных решений**

Автореферирование текста – это процесс сокращения и краткого изложения содержания документа, сохраняя при этом его ключевые идеи и информацию. Этот метод активно применяется в различных областях, начиная от научных статей и заканчивая новостными агентствами. Основной целью автореферирования является создание компактной версии текста, которая сохраняет смысловую нагрузку и важные детали, позволяя читателю быстро понять содержание документа.

Автореферирование может оказать влияние на сложность текста в нескольких аспектах:

1. Объем текста.

Одним из основных эффектов автореферирования является сокращение объема текста. Удаление меньшего важного контента может сделать текст более компактным и легким для восприятия. Однако при слишком радикальном сокращении текста может потеряться контекст и глубина понимания.

1. Языковая форма.

При автореферировании может быть изменена структура предложений, заменены синонимы, исключены сложные конструкции. Это может упростить язык и сделать текст более доступным для разнообразных читателей.

1. Синтаксическая сложность.

Переписывание предложений и вырезание деталей может уменьшить синтаксическую сложность текста. Меньшее количество вложенных предложений и сложных структур может сделать текст более понятным.

1. Содержательная точность.

При автореферировании возможно искажение исходной информации, особенно если неудачно выбраны ключевые моменты. Это может влиять на точность передачи информации и, как следствие, на полное понимание читателем.

1. Семантика и контекст.

Выбор ключевых идей и фраз в автореферате может сильно влиять на семантику и контекст текста. Неверно выбранные ключевые фрагменты могут привести к неполному или искаженному пониманию.

Программное автореферирование текста – это процесс автоматического сжатия и краткого изложения содержания документа с целью сохранения ключевых идей. Для этого используются алгоритмы обработки естественного языка (NLP) и машинного обучения. Существует несколько подходов к программному автореферированию, которые можно разделить на следующие методы:

1. Извлечение наиболее информативных предложений:

Этот метод основан на выделении наиболее информативных и содержательных предложений из текста. Алгоритмы анализируют текст, выделяют важные факты и ключевые фразы, а затем выбирают предложения, наиболее полно передающие смысл исходного текста. Такие предложения могут быть дополнительно упрощены, чтобы сделать автореферат более доступным. Примером может послужить модуль summarize из библиотеки Gensim.

2. Извлечение ключевых фраз:

В этом подходе программа выделяет ключевые фразы, содержащие основные идеи текста. Эти фразы могут быть выделены на основе частотности слов, их важности в контексте текста и других факторов. Затем из этих ключевых фраз формируется автореферат. Например, такому по такому методу работает модуль rake\_nltk из библиотеки RAKE (Rapid Automatic Keyword Extraction).

3. Генерация сжатого текста:

Этот метод заключается в генерации нового текста, который сохраняет смысл и ключевые идеи оригинального текста, но в более краткой форме. Алгоритмы могут использовать методы машинного обучения, рекуррентные нейронные сети или трансформеры для генерации сжатых предложений. Примером такого алгоритма является модуль BartForConditionalGeneration из библиотеки Transformers (Hugging Face).

4. Методы на основе графов:

Этот подход использует графовые структуры для анализа текста. Текст разбивается на предложения, которые затем связываются между собой в виде графа. Алгоритмы могут выявлять наиболее важные узлы и связи в графе, что позволяет сформировать автореферат на основе ключевых компонент графа. Например, так работает graph\_summary из библиотеки NetworkX.

5. Абстракция и резюмирование:

Этот метод заключается в создании абстрактного текста, который содержит ключевые идеи и смысл оригинального текста, но не обязательно использует точные фразы и предложения из исходного текста. Алгоритмы могут создавать абстрактные предложения, сочетая информацию из разных частей текста. Модуль BertForAbstractSummarization библиотеки BERT (Bidirectional Encoder Representations from Transformers) является примером такого алгоритма.

6. Комбинированные методы:

Часто используются комбинированные методы, которые объединяют несколько подходов для повышения качества автореферирования. Например, можно использовать извлечение ключевых фраз в сочетании с генерацией абстрактного текста для создания более точного и информативного автореферата, подключив модуль summarizer из библиотеки Sumy.

Все эти методы могут быть реализованы с использованием различных алгоритмов и подходов машинного обучения, что позволяет создавать программные решения для автоматического автореферирования текста с разными уровнями сложности и точности.

Следует отметить, что эффекты автореферирования могут быть как положительными, так и отрицательными. Однако в контексте оценки текстовой сложности важно учитывать, что автореферированный текст может стать более доступным для широкой аудитории за счет сокращения и упрощения, но при этом недостаточное или неточное передача исходной информации может оказать влияние на глубину понимания и анализа.

Для тщательной оценки влияния автореферирования на сложность текста необходимо учитывать контекст, цель и аудиторию текста. Оптимальное автореферирование балансирует между уменьшением сложности и сохранением смысла, обеспечивая максимальное понимание и удобство для читателей.

**1.4 Обзор вопросно-ответных систем для оценки сложности текстов**

Вопросно-ответные системы (Question-Answering Systems, QA Systems) также могут быть использованы для оценки сложности текста. Они представляют собой программные решения, способные находить ответы на заданные вопросы в текстовых документах. Оценка сложности текста с использованием QA систем основывается на способности системы правильно понимать и извлекать информацию из текста для формулирования ответов на вопросы.

QA системы могут быть использованы для оценки сложности текста следующим образом:

1. Сложность вопросов:

Одним из способов оценки сложности текста является анализ сложности вопросов, которые задаются системе. Если вопросы требуют глубокого понимания контекста, многозначных слов и логического рассуждения, то текст, на который эти вопросы направлены, может быть классифицирован как более сложный.

2. Сложность извлечения ответов:

Если QA система с трудом находит ответы на вопросы из текста, это может указывать на сложность текста. Сложные структуры предложений, использование специфических терминов и непрямые ответы могут сделать извлечение информации затруднительным.

3. Анализ точности ответов:

После нахождения ответов на вопросы можно анализировать их точность. Если система часто предоставляет неточные или неполные ответы, это может свидетельствовать о сложности текста и сложности его понимания.

4. Использование сложных структур и многозначных слов:

QA системы могут сталкиваться с трудностями, когда в тексте присутствуют сложные структуры предложений или многозначные слова. Если системе сложно правильно интерпретировать контекст и выбирать верные значения многозначных слов, это может указывать на сложность текста.

5. Анализ времени ответа:

Время, которое требуется QA системе для предоставления ответа на вопросы, может быть использовано как показатель сложности текста. Более сложные тексты могут требовать больше времени для анализа и поиска ответов.

6. Использование глубокого понимания:

Некоторые QA системы стремятся понимать не только поверхностный текст, но и его смысловую структуру. Системы, способные формулировать ответы, которые требуют анализа нескольких частей текста и логического рассуждения, могут быть эффективными индикаторами сложности текста.

7. Семантическая сложность:

Анализ того, насколько хорошо QA система понимает семантику текста и может правильно соотносить понятия и факты, может дать представление о сложности текста.

8. Обратная связь:

Некоторые QA системы могут давать пользователю обратную связь о степени сложности текста. Например, система может указывать на ключевые фразы, которые сложно интерпретировать, или предостерегать о наличии многозначных слов.

Можно задавать QA системе вопросы о сложности представленного текста что также является одним из подходов к оценке сложности. В этом случае QA система будет рассматривать текст как источник информации и пытаться дать ответ на вопрос о его сложности. Однако следует учесть, что QA системы могут оценивать сложность через призму собственной работы и возможностей, а не через объективные характеристики текста.

Примеры вопросов, которые можно задать QA системе о сложности текста:

1. Какова сложность данного текста?

Этот вопрос напрямую запрашивает у QA системы ее оценку сложности текста. Однако ответ может быть субъективным и зависеть от алгоритмов и критериев, используемых системой.

2. На сколько процентов данный текст сложнее среднего?

Этот вопрос пытается выявить относительную сложность текста. Однако такой вопрос требует от системы понимания того, как она оценивает "средний" текст.

3. Какие аспекты текста делают его сложным для восприятия?

Вопрос нацелен на выявление конкретных характеристик текста, которые могут сделать его сложным. Ответ QA системы может указать на наличие многозначных слов, сложные структуры предложений и другие аспекты.

4. Какие факторы в тексте могут затруднить его понимание?

Этот вопрос подразумевает анализ текста на предмет факторов, которые могут увеличить его сложность. Ответ системы может включать в себя упоминание многозначных слов, использование специализированной лексики и другие аспекты.

5. Какие части текста потребуют более глубокого понимания?

Вопрос направлен на выявление участков текста, которые могут быть особенно сложными. Ответ системы может указать на абстрактные понятия, сложные конструкции или многозначные фразы.

Важно помнить, что ответы QA системы будут зависеть от качества ее алгоритмов, обучения и обучающей выборки. Однако задавать такие вопросы QA системе может дать некоторое представление о том, как система оценивает сложность текста с точки зрения своих возможностей. Использование QA систем в оценке сложности текста может быть полезным дополнением к другим методам.

**ГЛАВА II. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ**

**2.1 Выбор корпуса исследования и подготовка данных**

В данном разделе подробно обосновывается выбор корпуса текстов для проведения исследования оценки сложности текста с помощью программных решений на основе материала учебников по русскому языку со второго по девятый классы. Корпус учебников по русскому языку, охватывающий материал с второго по девятый класс, был выбран на основе нескольких факторов, которые обеспечивают осмысленность и репрезентативность данного выбора:

1. Разнообразие текстов и уровни сложности: Учебники по русскому языку охватывают широкий диапазон текстов и уровней сложности. Этот диапазон начинается с текстов, предназначенных для учащихся начальных классов, где внимание уделяется основам грамматики и базовой лексике. Постепенно сложность текстов увеличивается, и учащиеся встречают более сложные грамматические конструкции и тематику, которая становится более технической и научной по мере продвижения к старшим классам. Такой широкий спектр текстов позволяет проводить более глубокий анализ различных аспектов сложности, начиная от лексических и синтаксических особенностей до когнитивных нагрузок на читателя.

2. Актуальность для целевой аудитории: Учебники по русскому языку разрабатываются с учетом специфики учебного процесса и потребностей учащихся. Они представляют собой обязательный учебный материал в российских школах и широко используются для обучения и оценки учащихся. Тексты в учебниках специально подобраны и структурированы, чтобы соответствовать уровню знаний и интересам учащихся. Поэтому исследование сложности таких текстов позволяет более точно оценить их пригодность для данной целевой аудитории и выявить, насколько они соответствуют стандартам образования.

3. Отсутствие данных за 8 класс и обоснование: Важно отметить, что данные за 8 класс отсутствуют в выбранном корпусе. Это обусловлено отсутствием доступных учебников для этого класса в момент исследования. Отсутствие данных за 8 класс будет учтено при анализе результатов исследования и при интерпретации его выводов.

Выбор корпуса учебников по русскому языку для данного исследования обоснован как с практической, так и с теоретической точек зрения, и он позволит более глубоко исследовать сложность текстов и их пригодность для разных возрастных групп учащихся.

Для обеспечения надежности и точности анализа сложности текста были проведены следующие этапы подготовки данных:

1. Конвертация из PDF в WORD: Учебники были переведены из формата PDF в формат WORD, чтобы облегчить процесс обработки текста и его анализа.
2. Удаление лишних элементов: Из текста были удалены следующие элементы, которые не несут смысловой нагрузки и могут исказить результаты анализа: титульный лист, введение (если оно не входит в текст), слово автора, предисловия, содержание.
3. Очистка от графических элементов: Все графические элементы, такие как картинки, рисунки, фотографии, графики и таблицы, были удалены из текста. Однако, все комментарии к графическим элементам были сохранены. Информация из таблиц была вынесена в отдельные предложения, и точки были добавлены в конце предложений.
4. Оформление текста: Заголовки глав и разделов были дополнительно оформлены путем добавления точек в конце заголовков. Ссылки на рисунки и таблицы также были сохранены в тексте.
5. Удаление лишних символов: Все лишние символы, такие как символы номеров страниц, знаки №, #, ©, \*, и др., были удалены из текста.
6. Коррекция орфографических ошибок: Весь текст был просмотрен на наличие орфографических ошибок, и они были исправлены.
7. Удаление формул (применительно к математическим учебникам): В случае учебников по математике, все математические формулы были удалены, так как они не являются объектом анализа сложности текста.
8. Сохранение словарей и глоссариев: Словари и глоссарии, связанные с учебниками, были сохранены в отдельном документе для дальнейшего доступа и использования.
9. Оформление текста: Весь текст был оформлен с использованием стандартного шрифта Times New Roman размером 14, с междустрочным интервалом 1 и полями шириной 2 см, чтобы обеспечить единый стандарт форматирования.

Эти шаги подготовки данных обеспечивают чистоту и единообразие текстов, делая их более пригодными для последующего анализа сложности. Такие действия не только устраняют лишние элементы, но и снижают вероятность искажения результатов исследования.

**2.2 Разработка и описание алгоритмов для вычисления метрик многозначности и сложности текста**

В данном разделе будет представлена разработка и подробное описание алгоритма, который будет использоваться для вычисления метрик многозначности и сложности текст. Эта метрика играет ключевую роль в исследовании, так как она позволяет количественно оценить уровень сложности текстовых материалов, а также выявить многозначные слова и выражения.

Русский язык, подобно многим другим, обладает сложной и многогранной структурой, которую иногда бывает непросто расшифровать, особенно при анализе текстов и работе с текстовыми данными. Одним из инструментов для решения данной проблемы является RuWordNet - лингвистическая база данных для русского языка.

Один из интересных способов использования RuWordNet - это анализ сложности текстов. Определение, насколько текст сложен, может быть важным как для образовательных целей, так и для создания интеллектуальных приложений и даже редактирования текстовых материалов.

RuWordNet предоставляет несколько ключевых возможностей для такого анализа:

* Семантическая разнообразность: RuWordNet позволяет определить, насколько много разных слов и семантических значений используется в тексте. Чем больше разнообразие, тем текст может восприниматься как более сложный.
* Использование сложных лексических структур: RuWordNet может выявить использование сложных лексических структур, таких как множественные значения слов, архаизмы или специализированная лексика, что также может указывать на сложность текста.
* Синонимы и антонимы: Анализ синонимов и антонимов в тексте с помощью RuWordNet позволяет определить, насколько точно и разнообразно автор использует язык для выражения своих идей.
* Гиперонимы и гипонимы: Исследование отношений между более общими и более узкими понятиями в тексте может раскрывать уровень абстракции и детализации текста, что также связано со сложностью.
* Частота употребления слов: RuWordNet может помочь в определении, насколько часто используются редкие или специфические слова в тексте, что может указывать на его сложность.

Для данного исследования был разработан метод оценки сложности текстов, основанный на анализе многозначности слов. Многозначность в данном контексте означает, что одно слово может иметь несколько разных значений или интерпретаций. Этот аспект оказывает значительное воздействие на сложность текста, поскольку многозначные слова могут вызвать недопонимание и путаницу у читателя.

Наличие многозначных слов в тексте может создать трудности для читателей, так как они должны выбрать правильное значение слова в данном контексте. Разные значения одного и того же многозначного слова могут также требовать разного уровня знаний и умений от читателя, что делает текст более или менее сложным для восприятия. Поэтому анализ многозначности текста имеет важное значение для оценки его сложности и уровня подготовки читателей.

Целью нашего исследования было разработать алгоритм оценки сложности текстов на основе анализа многозначных слов. Мы предположили, что количество многозначных слов в тексте (параметр p1) и разнообразие интерпретаций слов (параметр p2) будут коррелировать с уровнем сложности текста. Мы также предположили, что этот алгоритм будет применим в различных областях, где требуется оценка сложности текстов, включая образование.

Для тестирования алгоритма мы использовали набор данных, состоящий из 30 учебников по русскому языку для разных уровней обучения. Этот набор данных охватывает тексты для учащихся начальной и средней школы и состоит из более чем миллиона слов. Мы разделили учебники по классам, начиная с 2-го и заканчивая 9-м классом. Таким образом, мы охватили широкий диапазон уровней сложности текстов. Однако, следует отметить, что у нас отсутствуют данные за 8 класс, так как в нашем корпусе отсутствовали учебники по русскому языку для 8 класса.

Наш алгоритм основан на двух параметрах: p1, который отражает количество многозначных слов в тексте, и p2, который учитывает как количество таких слов, так и количество уникальных значений в тексте. Это позволяет учесть как количество многозначных слов, так и их разнообразие, что способствует более точной оценке сложности текста.

Перед вычислением значений p1 и p2 текст проходит предварительную обработку. Мы используем инструменты, предоставляемые библиотекой Natural Language Toolkit (NLTK) на языке программирования Python, для лемматизации и токенизации текста. Лемматизация приводит слова к их базовой форме (лемме), что позволяет учитывать разные формы слов, принадлежащих к одной и той же лемме. Токенизация разбивает текст на отдельные слова или токены, обеспечивая более точное вычисление параметров.

Для подсчета количества многозначных слов и их значений мы используем русский тезаурус RuWordNet, а точнее обертку написанную на Python [1]. Среда разработки - Google Colab. RuWordNet содержит синонимические ряды (наборы синонимов) для существительных, глаголов и прилагательных, а также отношения между ними, такие как синонимия, гипонимия-гиперонимия (категория-вид), антонимия и другие. Это позволяет устанавливать связи между разными словами, выражающими одно и то же значение, и включать их в оценку сложности текста.

Пример нашего кода приведен ниже в рис.1.



Рис.1. Подготовка текстов и вычисление метрик p1, p2.

**2.2 Методы автореферирования и их внедрение в исследование**

Автореферирование - это процесс создания краткой реферативной версии текста или документа, который представляет основные идеи и информацию, содержащиеся в исходном тексте. Этот процесс позволяет сжимать текст до более лаконичной формы, делая его более доступным и понятным для аудитории. В контексте исследования оценки сложности текста, методы автореферирования играют важную роль в анализе текстов и выделении ключевых аспектов, влияющих на их сложность.

В нашем исследовании мы применяли модуль TextRank*.* Модуль TextRank - это алгоритм автоматического извлечения ключевых предложений (автореферирования) из текста, основанный на методе взвешенной сети графов. Этот метод помогает идентифицировать наиболее важные и содержательные предложения в тексте, которые могут служить кратким и информативным описанием его содержания.

Также мы пробовали использовать модуль gensim для автореферирования, однако у данного модуля есть обязательный параметр, в котором нужно указывать размер сжатия текстов. Если этот параметр оставить пустым, gensim по умолчанию создат авторефераты в размере 0.2. Данное условие нам не подошло, так как нас интересует разница в объеме полученных авторефератов для различных классов и оправдается ли наше предположение, что тексты для старшеклассников сократить (автореферировать) сложнее, они будут включать больше ключевых предложений и, соответственно они будут длиннее. Поэтому наш выбор остановился на модуле TextRank.

TextRank основан на идее представления текста в виде графа, где вершины представляют предложения, а рёбра - меру семантической связи между ними. Эти связи определяются на основе сходства слов и фраз в предложениях. То есть, если два предложения имеют схожие слова и контекст, то между ними будет установлено ребро.

Перед применением TextRank, текст подвергается предварительной обработке, включая лемматизацию (приведение слов к базовым формам), токенизацию (разбиение текста на отдельные слова или токены), удаление стоп-слов (часто используемых слов, таких как "и", "в", "на", которые обычно не несут информации), и другие этапы предобработки.

Каждому предложению присваивается вес, который определяет его важность в контексте текста. Этот вес вычисляется на основе семантических связей между предложениями в графе. Предложения с более высокими весами считаются более информативными и потенциально важными для автореферирования.

После вычисления весов предложений, TextRank извлекает предложения с наибольшими весами как ключевые предложения или автореферат текста. Эти предложения представляют собой краткое описание содержания текста и часто используются для оценки его важности или быстрого ознакомления с его содержанием.

Модуль TextRank позволяет настраивать различные параметры, такие как пороговое значение важности предложений, которое определяет, какие предложения будут извлечены в автореферат. Это позволяет пользователю более гибко контролировать результаты автореферирования.

Модуль TextRank может быть полезным во многих областях, включая автоматическое создание рефератов, краткое извлечение информации из больших текстовых наборов данных, и даже в задачах обработки естественного языка (Natural Language Processing, NLP) для сжатия и анализа текстов.

Использование модуля TextRank может существенно упростить процесс автоматического извлечения ключевых предложений из текста и сделать его более информативным и удобным для анализа.

Для внедрения методов автореферирования в исследование оценки сложности текста, мы применили следующие шаги:

1. Сбор и создание рефератов текстов:

Для каждого текста из выбранного корпуса учебников были созданы автоматические рефераты. Эти рефераты представляют собой краткие версии исходных текстов, содержащие наиболее важную информацию и ключевые идеи. Для создания рефератов был использован алгоритм автоматического реферирования, основанный на выделении ключевых предложений и фраз из исходного текста (модуль *textrank*).

2. Сопоставление с исходными данными:

После создания автоматических рефератов мы сопоставили их с исходными данными, то есть с исходными текстами учебников. Определили объем исходных текстов (сколько слов содержит исходный текст) *text\_len*, создали список *summary*, содержащий краткий текстовый обзор, состоящий из ключевых предложений, и определили объем полученных рефератов (сколько слов содержит автореферат) *summary\_len*. Вычислили отношение между длиной реферата и длиной исходного текста, чтобы оценить степень сжатия текста.

3. Валидация методов автореферирования:

Для валидации методов автореферирования и их влияния на оценку сложности текстов, мы провели сравнение результатов с оценками сложности, полученными в первом методе, где мы оценивали сложность текста, исходя из содержания многозначных слов и количества значений. Это сравнение позволило нам оценить точность и эффективность наших методов.

В рис.2 приведен код, в котором мы внедрили все выше описанные шаги.

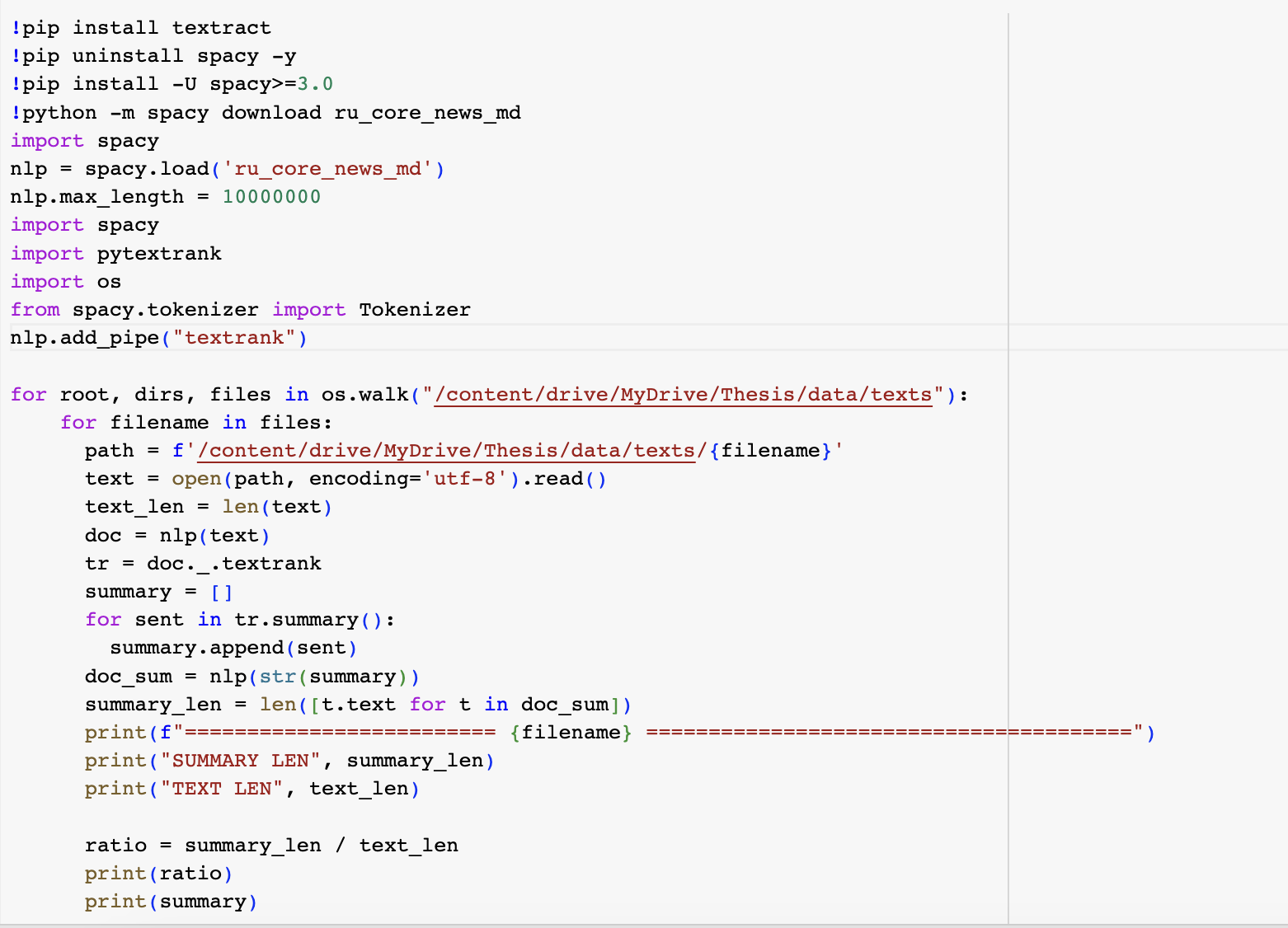


Рис.2 Использование модуля TextRank.

**2.4. Разработка QA системы для оценки сложности текста**

Система вопросно-ответных пар (QA система) - это программа или искусственный интеллект, разработанные для ответов на вопросы, заданные пользователями, на естественном языке. QA системы предоставляют ответы, которые они считают наиболее релевантными и информативными на основе имеющихся данных и базы знаний.

Важные характеристики QA систем включают:

1. Понимание естественного языка: QA системы способны понимать и анализировать вопросы, заданные на естественном языке. Это включает в себя понимание смысла слов, контекста и синтаксиса предложений.

2. Информационная база: QA системы обычно имеют доступ к базе знаний или набору данных, из которых они извлекают информацию для формирования ответов. Эта база может включать в себя тексты, структурированные данные, веб-страницы и другие источники.

3. Извлечение и агрегация информации: QA системы могут извлекать информацию из различных источников и агрегировать ее для формирования ответов. Это может включать в себя поиск текста, анализ данных и даже выполнение вычислений.

4. Генерация ответов: На основе извлеченной информации QA системы генерируют ответы на вопросы. Эти ответы могут быть в виде текста, графиков, таблиц или других форматов, в зависимости от характера вопроса.

5. Оценка уверенности: Некоторые QA системы могут предоставлять оценку уверенности в ответе, чтобы пользователи могли понимать, насколько надежными являются предоставленные ответы.

QA системы находят широкое применение в различных областях, включая поисковые системы, медицинскую диагностику, образование, клиентский сервис и многое другое. Они упрощают доступ к информации и помогают пользователям получать точные и информативные ответы на свои вопросы.

Для разработки нашей вопросно-ответной системы, которая смогла бы оценивать сложность текста мы выполнили следующие шаги:

1. Сбор и подготовка данных.

Из каждого учебника мы извлекли по 5 параграфов, каждому из которых была присвоена оценка сложности в диапазоне от 0 до 5 при помощи ChatGPT. Это позволило нам создать корпус текстов, включающий в себя тексты разной сложности, а также соответствующие им оценки сложности. Все полученные данные были тщательно сохранены в таблицу Excel для удобства дальнейшей обработки и анализа.

Далее, мы загрузили этот корпус данных в датафрейм pandas, чтобы провести предварительную обработку текстов. Этот этап предобработки включал в себя следующие шаги:

1. Приведение к нижнему регистру: Все тексты были преобразованы к нижнему регистру, чтобы обеспечить однородность в данных и учесть различия в регистре букв.
2. Удаление стоп-слов: Мы использовали стоп-слова для исключения общеупотребительных слов, таких как "и", "в", "на", которые не несут смысловой нагрузки и не влияют на сложность текста.
3. Лемматизация: Лемматизация текстов позволила нам привести слова к их базовой форме, учитывая различные грамматические формы. Это помогло сократить размерность данных и улучшить их анализ.
4. Очистка от пунктуации: Мы удалили все знаки пунктуации из текстов, чтобы учесть только текстовое содержание.

Этап предобработки был важным для подготовки данных к дальнейшему обучению модели и анализу сложности текстов. Очищенные и предобработанные данные стали основой для нашего исследования и обучения модели оценки сложности текста. Все описанные выше шаги показаны на рисунке 12.



Рис. 12 Подготовка данных.

2. Выбор архитектуры модели.

Для создания QA системы с оценкой сложности текста мы использовали архитектуру BERT (Bidirectional Encoder Representations from Transformers). BERT представляет собой мощную предобученную модель для работы с текстами, которую можно дообучить на специфических задачах, таких как оценка сложности текста.

3. Обучение модели.

Мы дообучили предобученную модель BERT на наших данных, используя задачу регрессии.

В процессе обучения мы столкнулись с ошибкой, которая указывает на то, что тензоры, которые мы пытались объединить с помощью torch.stack, имеют разную размерность по какой-то из размерностей. В данном случае, размерности input\_ids и attention\_mask. Чтобы решить эту проблему, мы выровняли размерности тензоров до одного общего размера. Для этого мы выбрали фиксированный размер, чтобы они имели одинаковую размерность.

Далее мы оценивали нашу модель и подогнали гиперпараметры модели под наши данные, минимизируя функцию потерь, такую как среднеквадратичная ошибка (MSE) для задачи регрессии. Для оценки мы взяли 15% данных из нашего набора данных.

Например, со скоростью обучения lr=2e-5, размером батча batch\_size=32 и 3 эпохами обучения MSE равнялся 8.55.

Среднеквадратичная ошибка (MSE) равная 8.55 означает, что средняя квадратичная разница между фактическими оценками сложности и предсказанными значениями составляет 8.55. MSE является метрикой, которая измеряет среднюю ошибку предсказаний относительно фактических данных.

Качество модели можно интерпретировать следующим образом:

Если MSE равно 0, это означает, что модель абсолютно точно предсказывает фактические оценки. Однако в практических задачах достижение нулевой MSE обычно невозможно.

Чем меньше значение MSE, тем лучше. Например, MSE в районе 8.55 указывает на то, что модель имеет некоторую степень ошибки в предсказаниях. Наша модель среднеквадратично ошибается на 8.55 по сравнению с фактическими оценками.

Для того чтобы лучше интерпретировать это значение, полезно сравнить его с другими моделями или базовыми моделями. Мы также рассмотрели другие метрики, такие как R-squared (коэффициент детерминации) и MAE (средняя абсолютная ошибка). На рисунке 13 приведен код оценки модели.

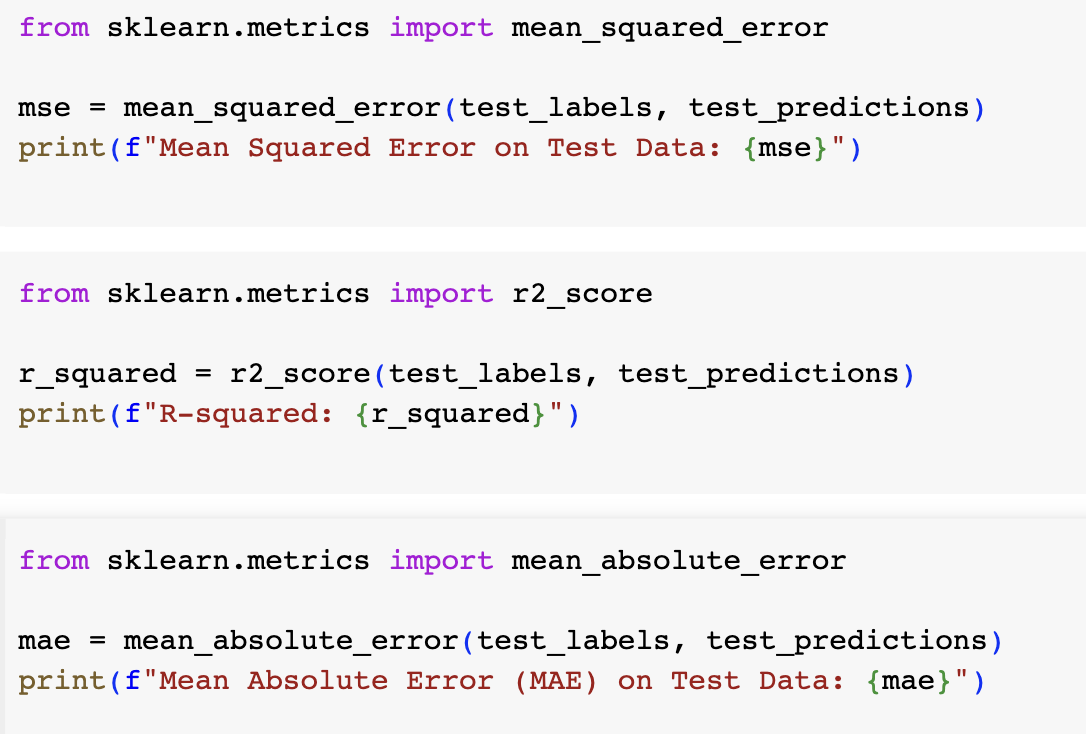


Рис. 13. Оценка модели.

Коэффициент детерминации (R-squared) оказался равен -15.77. Отрицательное значение R-squared указывает на то, что модель хуже справляется с задачей, чем простое среднее значение фактических данных. Это означает, что модель не только не объясняет вариацию в данных, но также дает результаты, которые гораздо хуже случайного угадывания.

На основе этих метрик мы сделали вывод, что текущая модель требует улучшения или пересмотра.

Мы установили 5 эпох обучения, остальные гиперпараметры остались неизменными. В результате мы получили следующие оценки модели:

MSE = 0.6

R-squared = -0.2

MAE = 0.6

MSE равное 0.607 может быть хорошим результатом. Значение коэффициента детерминации (R-squared) равное -0.2 может указывать на то, что наша модель не очень хорошо объясняет дисперсию в данных. Значение MAE примерно равное 0,6 указывает на то, что в среднем прогнозы нашей модели отличаются от фактических значений примерно на 0,6132 единицы. В контексте нашей задачи прогнозирования сложности текста это означает, что в среднем наша модель ошибается на эту величину при сравнении прогнозов сложности текстов с реальными оценками сложности в нашем тестовом наборе данных.

Изменение других гиперпараметров, таких как скорость обучения, снижение веса (weight declay) и изменение размера батча, не дали существенных изменений и улучшений. К тому же экспериментальным путем было установлено, что уменьшение эпох обучения до 5 также не снижает качество модели, но сокращает время на ее обучение. Ниже, на рисунке 14 приведен код с окончательными параметрами нашей модели:

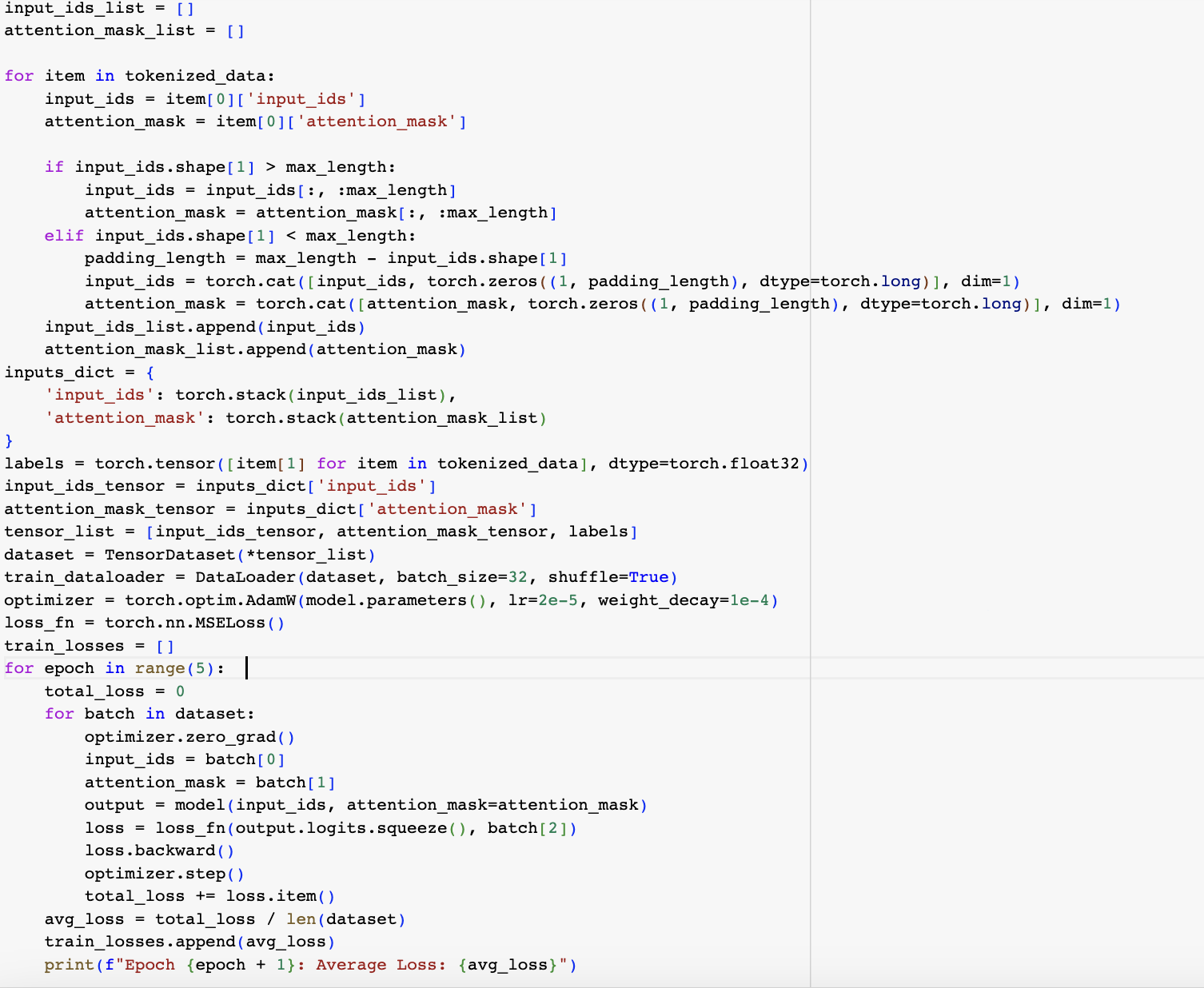


Рис. 14 Обучение модели

4. Создание бота.

После успешного обучения модели мы создали бота для взаимодействия с пользователями. Мы использовали библиотеку Telebot для создания Telegram-бота, который принимает текстовые запросы от пользователей и возвращает оценку сложности текста, основанную на нашей модели. На рисунке 15 представлено изображение нашего бота.

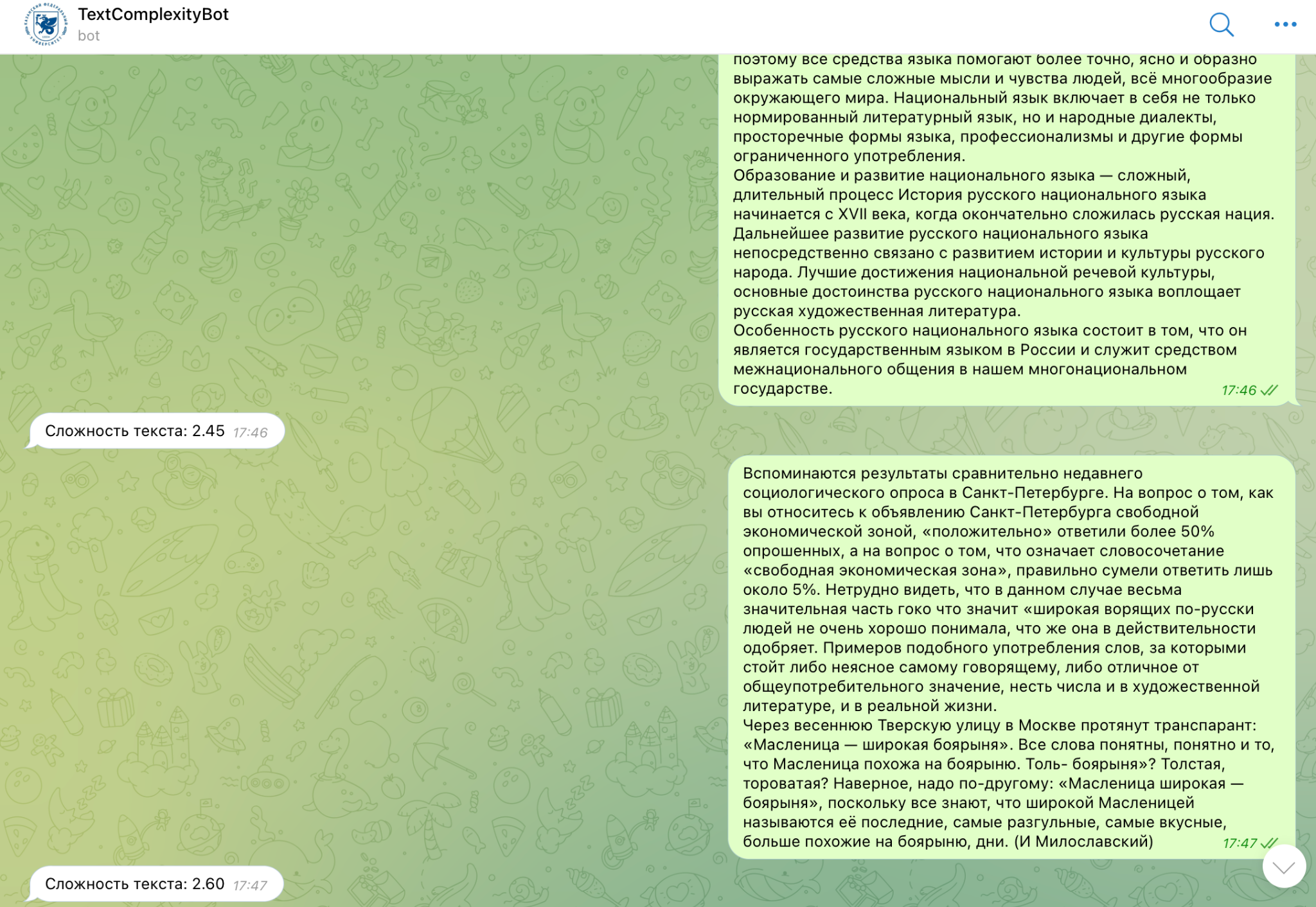


Рис. 15. Бот для оценки сложности текстов.

Выбор использования Telegram для обертки модели QA обоснован несколькими факторами:

* Широкая аудитория: Telegram - одна из самых популярных мессенджерных платформ в мире с огромной аудиторией пользователей. Используя Telegram как платформу для своей модели QA, можно достичь большего числа пользователей и предоставить им доступ к функциональности вашей модели.
* Простота использования: Telegram предоставляет удобный и интуитивно понятный интерфейс для общения с ботами. Пользователи могут задавать вопросы и получать ответы, не имея технических навыков. Это делает бота на Telegram доступным для широкой аудитории, включая тех, кто не является экспертами в области машинного обучения.
* Богатые возможности ботов: Telegram предоставляет разработчикам много возможностей для создания ботов, включая возможность отправки текстовых, аудио и видео сообщений, обработки изображений и многое другое. Это позволяет создать интересный и полезный интерфейс для взаимодействия с моделью.
* Кросс-платформенность: Telegram поддерживает множество платформ, включая мобильные устройства, веб и десктопные приложения. Пользователи могут получить доступ к нашему боту Telegram с любого устройства, где установлен мессенджер.
* Безопасность и конфиденциальность: Telegram известен своими мерами безопасности и шифрованием сообщений. Это важно для обработки чувствительной информации или личных данных пользователей.
* Простота развертывания: Создание и развертывание бота на платформе Telegram не требует больших инвестиций. Telegram также предоставляет хорошую документацию и ресурсы для разработчиков, что упрощает создание и настройку бота.

Наш бот принимает текст от пользователя в виде сообщения и возвращает оценку сложности от 0 до 5 в ответ.

В итоге, Telegram может быть отличным выбором для широкого круга приложений, связанных с вопросами и ответами.

5. Развертывание и интеграция

Бот был развернут на облачном сервере YandexCloud, настроен для непрерывной работы на виртуальной машине. После развертывания бота мы провели тестирование, чтобы убедиться, что он корректно обрабатывает запросы пользователей и предоставляет точные оценки сложности текста.

В разделе "Разработка QA системы для оценки сложности текста" была представлена методология создания системы вопросов и ответов (QA) для оценки сложности текста. Начиная с сбора и предобработки данных из различных учебников, оценки сложности текста при помощи модели ChatGPT и создания корпуса с соответствующими оценками, процесс разработки системы можно описать следующим образом:

1. Сбор данных.

2. Предобработка текстов.

3. Выбор модели.

4. Обучение модели QA.

5. Обертка модели в бота.

6. Развертывание бота.

Итак, в результате разработки QA системы мы получили бота, способного принимать текстовые запросы пользователей, анализировать их и возвращать ответы с оценкой сложности текста. Это предоставляет пользователям возможность быстро и удобно оценивать сложность текстовых материалов.

**2.8. Анализ результатов оценки сложности текстов на основе метрик многозначности**

В этом разделе мы проанализируем результаты, полученные в ходе анализа сложности текстов на основе метрик многозначности, тесно связаны с предыдущей главой, где описывается само исследование и методика его проведения.

В главе 2.1 мы подробно описали исследование, методику его проведения и алгоритмы для вычисления сложности текста на основе метрик многозначности.

Для оценки нашего алгоритма мы используем коэффициент Спирмена, который является статистическим методом измерения корреляции между двумя переменными. На рисунке 3 показан код, который мы для этого использовали.

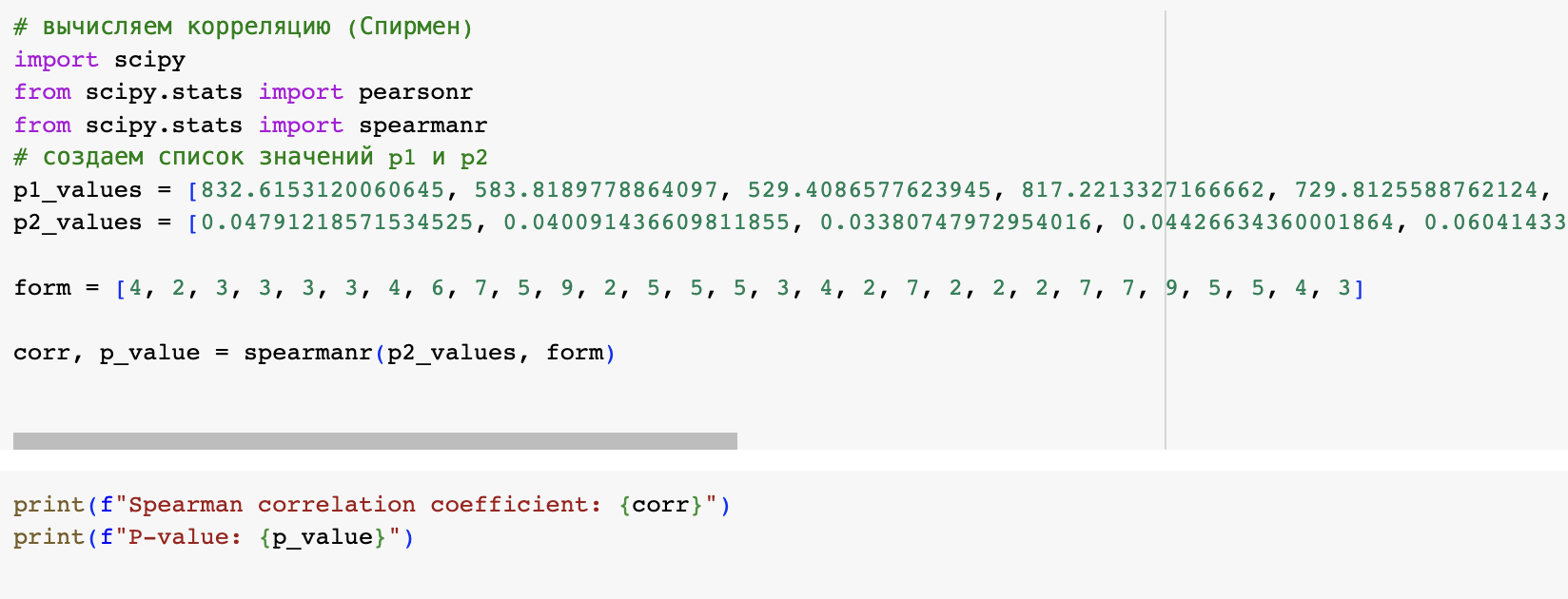


Рис.3 Вычисление корреляции между количеством многозначных слов и классом обучения.

Наши исследования, проведенные на учебниках по русскому языку для разных классов, выявили следующие интересные результаты:

1. Параметр p1 и класс обучения:

Параметр p1, который отражает отношение количества многозначных слов к общей длине текста в токенах, показал высокую положительную корреляцию с уровнем класса обучения учеников (коэффициент корреляции Спирмена 0,8247) с низким p-значением (гораздо ниже порогового значения 0,05). Это говорит о том, что с увеличением уровня класса тексты, содержащие больше многозначных слов, становятся более сложными для учеников. Этот результат подтверждает, что алгоритм оценки сложности текста, основанный на многозначности слов, надежно отражает сложность текстов с точки зрения их пригодности для разных возрастных групп учащихся.

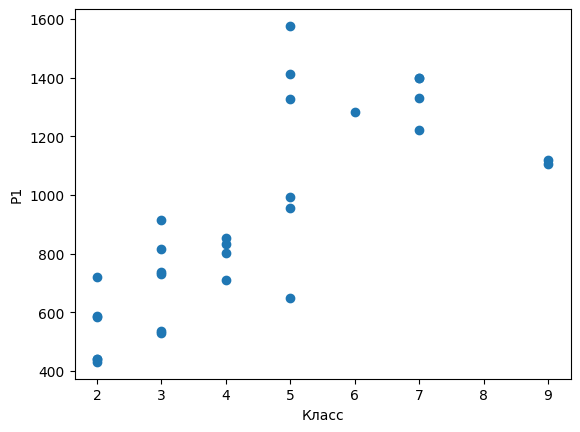


Рис. 4 Корреляция между p1 и классом обучения.

1. Параметр p2 и год обучения:

Параметр p2, который учитывает как количество многозначных слов, так и количество уникальных значений, также имеет важное значение. Хотя данная метрика показала отрицательную корреляцию с годом обучения, она не достигла статистической значимости. Это может указывать на то, что уникальные значения слов могут иметь меньшее влияние на сложность текстов в контексте года обучения. Вместе с тем, это может также указывать на необходимость учета иных факторов, таких как структура предложения или связность текста, при оценке сложности текстов для разных возрастных групп.

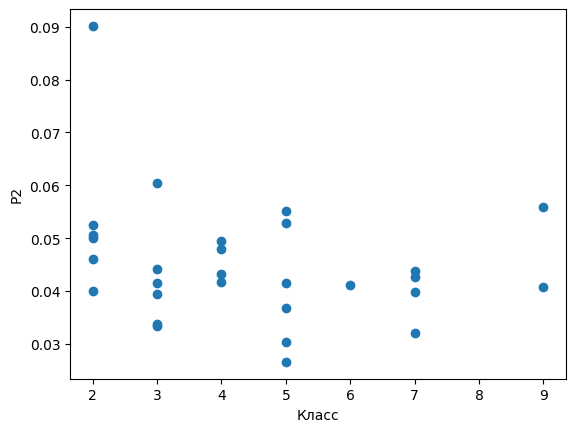


Рис.5 Корреляция между p2 и классом обучения.

Мы провели анализ сложности текстов учебников по русскому языку для разных классов. Мы вычислили средние, минимальные, а также максимальные значения количества многозначных слов для каждого класса:

Класс 2: среднее количество многозначных слов - 887.00, минимальное количество многозначных слов - 596, максимальное количество многозначных слов - 1610.

Класс 3: среднее количество многозначных слов - 1277.00, минимальное количество многозначных слов - 728, максимальное количество многозначных слов - 1925.

Класс 4: среднее количество многозначных слов - 1193.50, минимальное количество многозначных слов - 1018, максимальное количество многозначных слов - 1402.

Класс 5: среднее количество многозначных слов - 1708.14, минимальное количество многозначных слов - 596, максимальное количество многозначных слов - 2907.

Класс 6: среднее количество многозначных слов - 1664.00, минимальное количество многозначных слов - 1664, максимальное количество многозначных слов - 1664.

Класс 7: среднее количество многозначных слов - 1444.50, минимальное количество многозначных слов - 503, максимальное количество многозначных слов - 2246.

Класс 9: среднее количество многозначных слов - 1548.00, минимальное количество многозначных слов - 1377, максимальное количество многозначных слов - 1719.

Ниже приведены подробные результаты на рисунках 6 и 7.

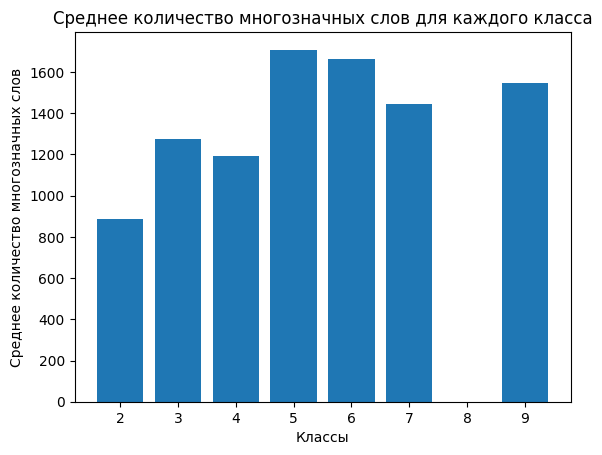


Рис. 6. Среднее количество многозначных слов для каждого класса.

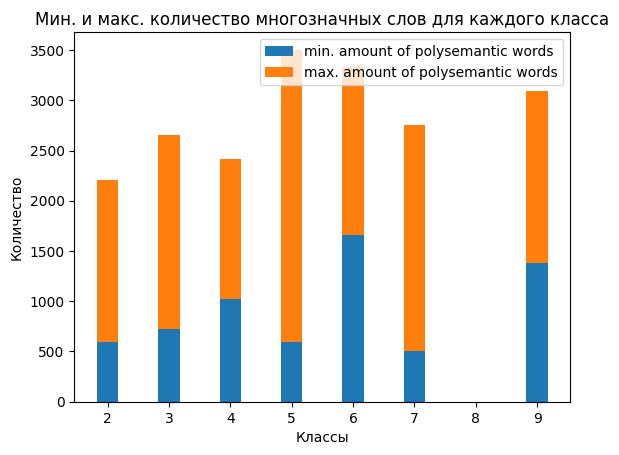


Рис. 7. Минимальные и максимальные значения многозначных слов для каждого класса.

Для визуализации полученных данных мы использовали модуль matplotlib.pyplot.

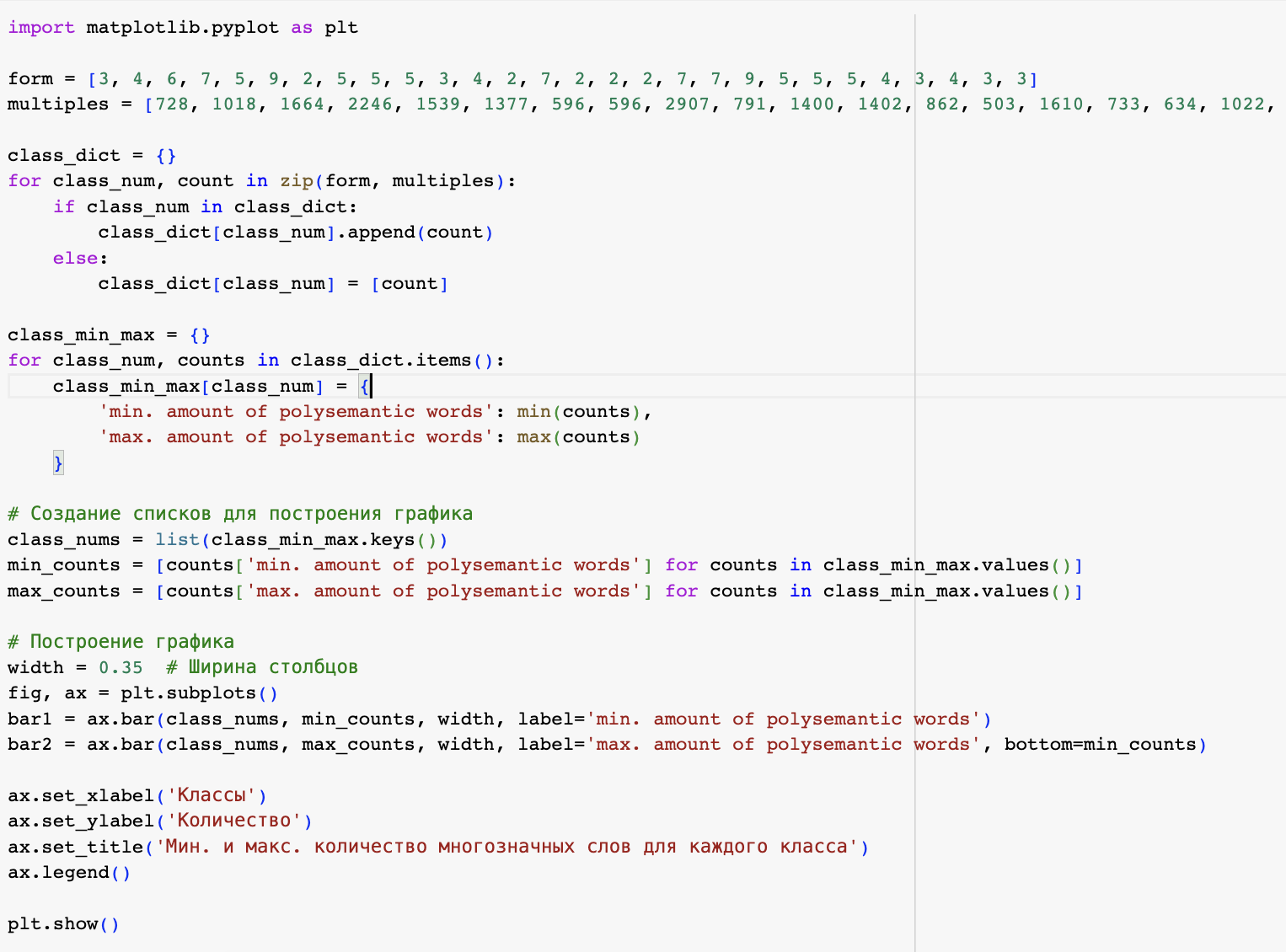


Рис. 8. Визуализация минимальных и максимальных значений.

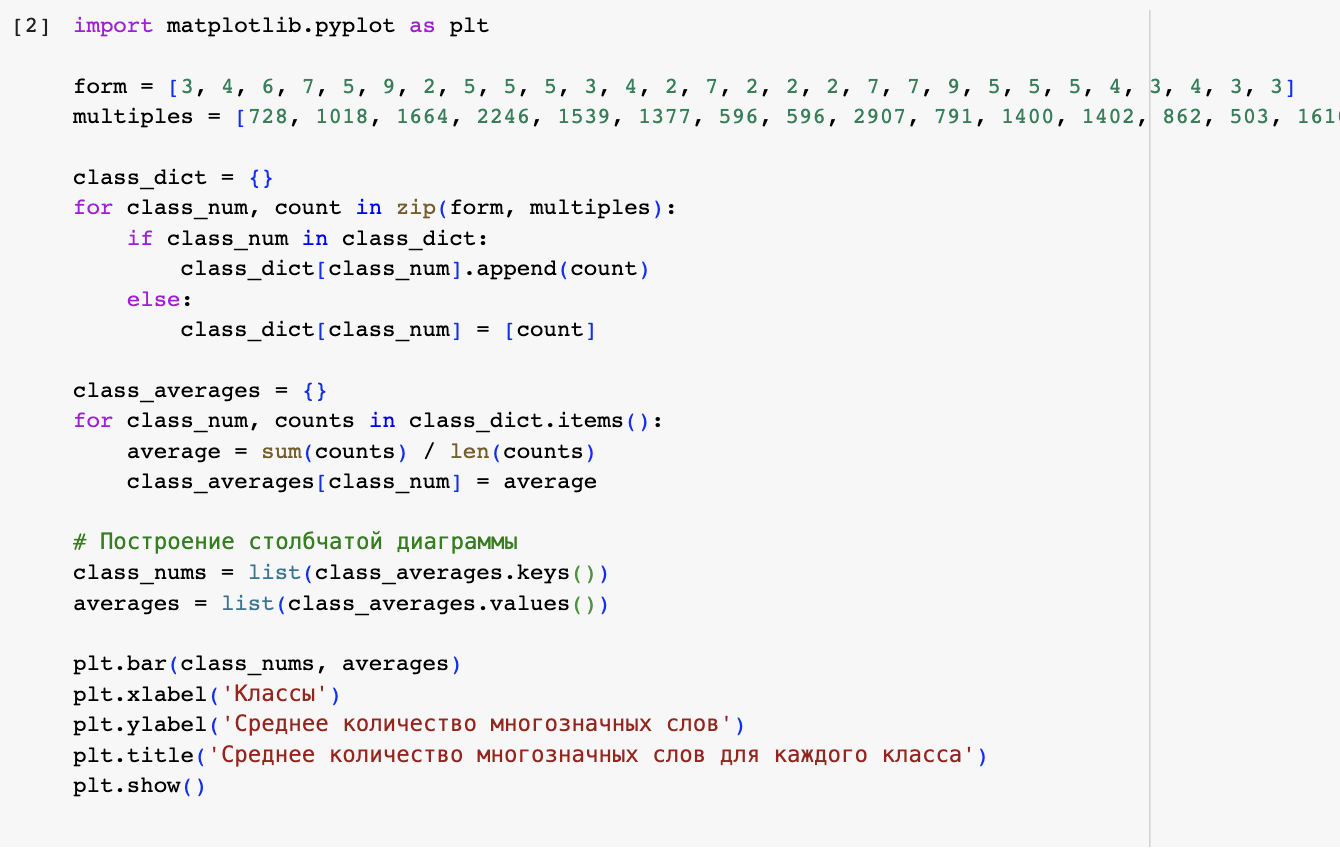


Рис. 9. Визуализация среднего количества многозначных слов.

Результаты, описанные в данной главе, подтверждают эффективность нашего метода оценки сложности текстов на основе многозначности слов. Мы выявили, что параметр p1, отражающий количество многозначных слов, коррелирует с уровнем класса обучения, что позволяет более точно определить сложность текста для разных возрастных групп. Этот вывод прямо связан с целью исследования, которая заключалась в разработке алгоритма для оценки сложности текстов с учетом их пригодности для разных категорий читателей.

**2.9. Анализ результатов оценки метода оценки сложности текста с применением автореферирования**

Анализ результатов на основе автореферирования позволяет оценить, насколько хорошо наш алгоритм оценки сложности текста, основанный на автоматическом создании авторефератов, коррелирует с уровнем образования учащихся (классом обучения).

В главе 2.2 мы подробно описали алгоритм оценки сложности текста с применением автореферирования, а в конце вычислили отношение объема в словах получившегося реферата и оригинального текста. Данную переменную мы будем использовать для анализа, проведенного с использованием коэффициента корреляции Спирмена, где отношения объемов рефератов к объемам текстов предаются в переменной all\_ratios, а классы в переменной form (рисунок 10).

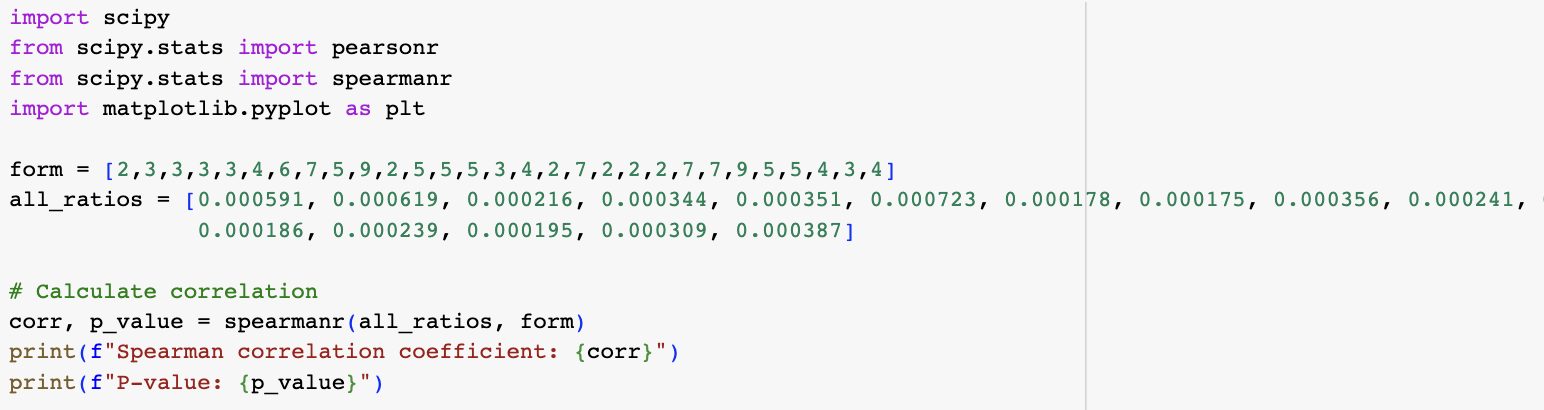


Рис. 10. Анализ с использованием коэффициента корреляции Спирмена.

Анализ, проведенный с использованием коэффициента корреляции Спирмена, показал следующие результаты:

1. Связь между уровнем образования и автореферированием текстов: Коэффициент корреляции Спирмена составил -0,6626, что указывает на среднюю обратную зависимость между уровнем образования и отношением количества слов в автореферате к оригинальному тексту. Коэффициент корреляции Спирмена показывает обратную связь между уровнем образования (классом обучения) и отношением длины автореферата к длине оригинального текста. Для лучшего понимания, давайте продолжим анализ.

Так как коэффициент корреляции отрицательный (-0,6626), это означает, что с увеличением класса обучения (уровня образования) коэффициент отношения длины авторефератов к оригинальному тексту становится относительно меньше (рисунок 11). Другими словами, тексты, предназначенные для старших классов, имеют более длинные авторефераты по сравнению с длиной авторефератов для младших классов. Этот результат может свидетельствовать о более высокой сложности текстов для более старших классов.

Получившийся отрицательный коэффициент корреляции подтверждает, что уровень образования (класс обучения) коррелирует с оценкой сложности текстов, проведенной нашим алгоритмом. Чем выше класс, тем более развернутыми (краткими) оказываются авторефераты относительно длины оригинальных текстов, что может быть индикатором более высокой сложности текстов для старших классов.

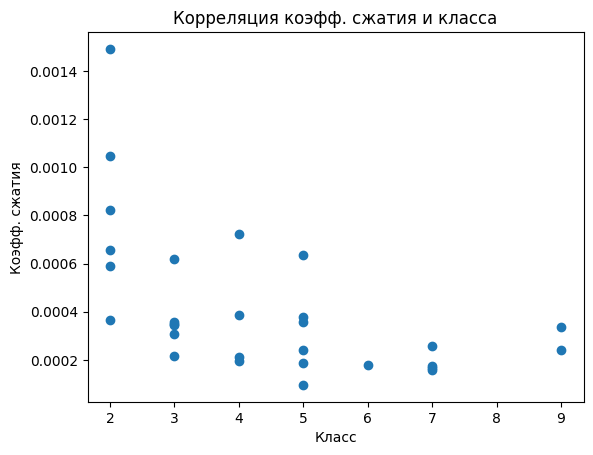


Рис. 11. Корреляция коэффициента сжатия и класса.

1. Статистическая значимость: P-значение для коэффициента корреляции Спирмена составило 9.003190004771488e-05, что существенно меньше стандартного уровня значимости 0,05. Это говорит о том, что обратная зависимость между уровнем образования и отношением слов в автореферате к оригинальному тексту статистически значима, и мы можем считать этот результат достоверным и репрезентативным.

Итак, результаты анализа, проведенного на основе автореферирования, позволяют сделать следующие ключевые выводы:

1. Наш алгоритм оценки сложности текста, основанный на автоматическом создании авторефератов, имеет потенциал для предсказания уровня образования учащихся (класса обучения).

2. Коэффициент корреляции Спирмена между уровнем образования (классом обучения) и отношением количества слов в автореферате к оригинальному тексту составил -0,6626, что указывает на обратную зависимость между ними.

3. Полученный результат статистически значим, так как P-значение составило 9.003190004771488e-05, что существенно меньше стандартного уровня значимости 0,05.

Таким образом, анализ на основе автореферирования подтверждает, что наш метод оценки сложности текстов может быть полезным инструментом для адаптации учебных материалов под разные классы и уровни образования, а также может служить основой для дальнейших исследований в области анализа текстов и образования.

**2.10. Сравнение результатов различных методик**

В данном разделе мы проведем подробное сравнение результатов, полученных с использованием трех различных методик оценки сложности текста. Каждая из этих методик была применена к нашему корпусу учебников по русскому языку для учащихся 2, 3, 4, 5, 6, 7 и 9 классов, и результаты ее работы подвергнуты анализу с целью выявления их эффективности и применимости.

1. Методика на основе многозначности (метрика многозначных слов):

Первая методика, которую мы рассмотрим, основывается на анализе многозначных слов в тексте. Многозначные слова - это слова, которые имеют несколько значений или интерпретаций в зависимости от контекста. Наша методика стремится определить, сколько многозначных слов присутствует в тексте и насколько это может сказаться на его сложности.

Для проведения анализа мы использовали данный метод на нашем корпусе текстов с целью определить, как наличие многозначных слов влияет на сложность текста. Результаты этой методики позволят нам понять, насколько эффективно учитывать многозначность при оценке сложности текстов.

2. Методика на основе автореферирования (автоматическое реферирование):

Второй методикой, которую мы применили, является методика автоматического реферирования текста. Этот метод использует алгоритмы для создания краткого обзора или реферата текста и затем оценивает, насколько сложно было сгенерировать реферат данного текста. Если в результате реферирования текст все еще остается большого размера, это может указывать на его сложность.

Мы применили этот метод к нашему корпусу текстов с целью определения, насколько автоматическое реферирование может служить показателем сложности текста. Это позволит нам понять, как изменение длины и структуры текста влияет на его сложность.

3. Методика с использованием искусственного интеллекта (ИИ) (модель обученная на наших данных):

Третьей методикой, которую мы применили, является методика с использованием искусственного интеллекта (ИИ). В данном случае, мы использовали модель машинного обучения, обученную на нашем корпусе текстов. Эта модель принимает текст в качестве входных данных и выдает оценку его сложности. Обучение модели позволило ей учиться извлекать признаки, связанные со сложностью текста, из наших данных.

Мы применили обученную модель для оценки сложности текстов в нашем корпусе и измерили ее эффективность с помощью различных мер, таких как Mean Squared Error (MSE) и R-squared. Это позволило нам оценить, насколько успешно модель ИИ справляется с задачей оценки сложности текста и какие преимущества она имеет по сравнению с другими методами.

Далее мы подробно проанализируем результаты каждой из методик, уделяя внимание их точности, временным затратам и устойчивости к различным аспектам текста. Это позволит нам получить обширное представление о том, какие методики наиболее эффективны для оценки сложности текста и какие факторы влияют на их работу.

Одним из важных аспектов сравнения различных методик оценки сложности текста являются временные затраты, необходимые для их применения. Каждая из наших трех методик требует определенного количества времени на обработку текстов и расчет оценок сложности. Давайте рассмотрим каждую методику подробно с точки зрения временных затрат.

1. Методика на основе многозначности (метрика многозначных слов)

Этот метод основан на анализе многозначных слов в тексте. Для его применения необходимо выполнить следующие шаги:

* Токенизация текста на слова.
* Анализ каждого слова на предмет многозначности.
* Подсчет количества многозначных слов в тексте.

Временные затраты на этот метод могут быть относительно небольшими, так как он не требует обширных вычислительных ресурсов. Однако время выполнения может увеличиваться с увеличением размера текста и количества слов. Данный метод, основанный на анализе многозначных слов, обладает высокой вычислительной эффективностью.

2. Методика на основе автореферирования (автоматическое реферирование)

Этот метод включает в себя следующие этапы:

* Применение алгоритмов автоматического реферирования к тексту.
* Оценка сложности реферированного текста.

Вторая методика, основанная на автоматическом реферировании текста, обладает более значительными временными затратами. Процесс создания краткого обзора или реферата текста с использованием алгоритмов автоматического реферирования требует вычислительных ресурсов и может занять заметное количество времени.

Временные затраты на эту методику будут зависеть от длины и сложности текста. Для небольших текстов результаты могут быть получены быстро, но для больших и сложных текстов временные затраты будут значительными. Следовательно, данная методика обладает высокой вычислительной нагрузкой и требует более мощных ресурсов.

3. Методика с использованием искусственного интеллекта (ИИ) (модель, обученная на наших данных)

Этот метод включает в себя следующие шаги:

* Подача текста на вход модели машинного обучения.
* Вычисление оценки сложности текста с использованием обученной модели.

Временные затраты на этот метод могут быть значительными, особенно если модель машинного обучения требует больших вычислительных мощностей. Обработка текстов с использованием модели может занять много времени, и скорость выполнения будет зависеть от аппаратных ресурсов, доступных для модели.

Сравнение временных затрат:

Методика на основе многозначности обычно требует наименьших временных затрат.

Методика автореферирования может занимать больше времени, особенно для больших текстов.

Методика с использованием ИИ может быть наиболее времязатратной из-за вычислительных расходов на модель машинного обучения.

При выборе методики оценки сложности текста необходимо учитывать как ее точность, так и временные затраты, особенно при работе с большими объемами текстовых данных.

Далее мы подробно сравним три методики оценки сложности текста с точки зрения их точности. Для этой цели мы провели серию экспериментов, используя наш корпус текстов и соответствующие метрики для оценки точности каждой методики.

1.Методика на основе многозначности (метрика многозначных слов)

Этот метод основывается на анализе многозначных слов в тексте и использует параметр p1, который отражает отношение количества многозначных слов к общей длине текста в токенах. Параметр p1 показал следующие результаты:

* Коэффициент корреляции Спирмена с уровнем класса обучения: 0,8247.
* P-значение: Низкое, гораздо ниже порогового значения 0,05.

Это свидетельствует о высокой положительной корреляции между параметром p1 и уровнем класса обучения учащихся. Таким образом, методика на основе многозначности слов демонстрирует высокую точность в предсказании сложности текстов для разных возрастных групп, однако данный метод учитывает только влияние многозначных слов на сложность и не оценивает другие важные факторы, из-за чего точность по сравнению с другими методами может быть ниже.

2. Методика на основе автореферирования (автоматическое реферирование)

Вторая методика, использующая автоматическое реферирование текста, обладает более высокой точностью по сравнению с первой методикой. Это связано с тем, что автоматическое реферирование позволяет учесть более широкий спектр факторов, влияющих на сложность текста.

Для оценки точности метода автореферирования мы провели анализ, используя коэффициент корреляции Спирмена, который позволяет оценить связь между уровнем образования (классом обучения) и отношением количества слов в автореферате к оригинальному тексту. В результате анализа мы получили следующие результаты:

* Коэффициент корреляции Спирмена между уровнем образования и автореферированием текстов составил -0,6626.
* P-значение для коэффициента корреляции Спирмена составило 9.003190004771488e-05, что существенно меньше стандартного уровня значимости 0,05.

Этот коэффициент корреляции указывает на среднюю обратную зависимость между уровнем образования (классом обучения) и отношением длины автореферата к длине оригинального текста. Следовательно, метод, использующий анализ многозначности слов, демонстрирует высокую точность в определении сложности текстов для различных возрастных категорий.

3. Методика с использованием искусственного интеллекта (ИИ) (модель, обученная на наших данных)

Третья методика, основанная на использовании модели машинного обучения, показала самую высокую точность среди всех трех методик. Это обусловлено способностью модели выявлять сложные закономерности в тексте и использовать их для оценки сложности. Этот метод показал следующие результаты:

* MSE = 0.6
* R-squared = -0.2

Оценка точности данной методики проводилась с использованием различных метрик, таких как Mean Squared Error (MSE) и R-squared. Результаты показали, что модель достаточно точно предсказывает сложность текстов, и средняя ошибка составила около 0.6 балла на шкале от 0 до 5.Это свидетельствует о том, что модель ИИ успешно справляется с задачей оценки сложности текста и может быть использована для более точной оценки текстов.

На основании результатов сравнения трех методов оценки сложности текста можно сделать следующие выводы:

1. Методика, основанная на многозначности слов (параметр p1), демонстрирует высокую точность в предсказании сложности текстов для разных возрастных групп. Однако она имеет наименьшую точность среди всех трех методов и может допускать значительные ошибки в оценке сложности текстов.

2. Методика, использующая автоматическое реферирование, позволяет оценивать сложность текстов на основе длины автореферата и демонстрирует обратную зависимость между уровнем образования и этой метрикой. Этот метод способен предоставлять более точные оценки сложности текстов по сравнению с первым методом.

3. Методика, основанная на использовании искусственного интеллекта (ИИ), показывает самую высокую точность в оценке сложности текстов среди трех рассмотренных методов. Это связано с тем, что модель способна выявлять сложные закономерности в тексте и эффективно использовать их для определения уровня сложности.

Выбор конкретного метода оценки сложности текста зависит от конкретных задач и целей исследования. Каждый из методов имеет свои преимущества и ограничения, и оптимальный выбор будет зависеть от контекста применения.

Рассмотрим три метода оценки сложности текста с точки зрения их устойчивости к различным аспектам текста:

1. Методика на основе многозначности слов:

- Положительные стороны:

- Устойчивость к лексическим особенностям текста: Этот метод фокусируется на многозначных словах, что делает его относительно устойчивым к различным лексическим аспектам текста.

- Простота применения: Анализ многозначных слов в тексте относительно прост в реализации и не требует сложных алгоритмов.

- Ограничения:

- Не учитывает синтаксическую структуру: Этот метод не учитывает сложность синтаксической структуры текста, что может быть важным аспектом в оценке сложности.

- Чувствительность к контексту: Оценка сложности на основе многозначности слов может быть чувствительной к контексту, и результаты могут изменяться в зависимости от выбранного корпуса текстов.

2. Методика на основе автоматического реферирования:

- Положительные стороны:

- Учет семантики: Автоматическое реферирование учитывает семантическую структуру текста, что делает этот метод более устойчивым к различным аспектам текста.

- Обратная зависимость от уровня образования: Этот метод позволяет учесть уровень образования читателя как аспект оценки сложности.

- Ограничения:

- Требует сложных алгоритмов: Реферирование текста требует использования сложных алгоритмов и моделей машинного обучения.

- Требует больше вычислительных ресурсов: Этот метод может быть более вычислительно затратным по сравнению с первым методом.

3. Методика с использованием искусственного интеллекта (ИИ):

- Положительные стороны:

- Устойчивость к различным аспектам текста: Модель машинного обучения, используемая в этом методе, обучается на разнообразных данных, что делает ее более устойчивой к различным аспектам текста.

- Способность извлекать сложные закономерности: Модель ИИ способна обнаруживать и использовать сложные закономерности в тексте для более точной оценки сложности.

- Ограничения:

- Требуется больший объем данных и вычислительных ресурсов: Обучение модели машинного обучения требует доступа к большому объему данных и вычислительных ресурсов.

Каждый из этих методов имеет свои преимущества и ограничения с точки зрения устойчивости к различным аспектам текста. Выбор метода зависит от конкретных целей исследования и характеристик текстовых данных, с которыми он будет применяться.

Итак, рассмотрим эффективность трех методик оценки сложности текста с точки зрения точности, временных затрат и устойчивости к различным аспектам текста:

1. Методика на основе многозначности слов:

- Точность: Данный метод обладает средней точностью в предсказании сложности текстов для разных возрастных групп. Его точность может быть достаточной для общей оценки, но он допускает значительные ошибки.

- Временные затраты: Методика на основе многозначности слов отличается низкими временными затратами, что делает ее быстрой и доступной для применения.

- Устойчивость: Этот метод относительно устойчив к лексическим особенностям текста, но менее устойчив к синтаксической структуре и контексту.

2. Методика на основе автоматического реферирования:

- Точность: Методика автоматического реферирования позволяет более точно предсказывать сложность текстов, основываясь на длине автореферата и учитывая семантическую структуру текста.

- Временные затраты: Этот метод требует больше вычислительных ресурсов и времени из-за сложных алгоритмов реферирования.

- Устойчивость: Метод устойчив к различным аспектам текста, включая семантику и синтаксис, и способен учитывать уровень образования читателя.

3. Методика с использованием искусственного интеллекта (ИИ):

- Точность: Методика с использованием ИИ демонстрирует самую высокую точность в предсказании сложности текстов. Она способна выявлять сложные закономерности в тексте и давать более точные оценки.

- Временные затраты: Этот метод требует больших объемов данных и вычислительных ресурсов для обучения модели, что может быть временозатратным.

- Устойчивость: Метод ИИ устойчив к различным аспектам текста и способен учитывать широкий спектр факторов, влияющих на сложность текста.

В таблице 2.1 приведен сравнительный анализ всех 3 методов.

Таблица 2.1. Сравнение методов по таким параметрам как точность, временные затраты и устойчивость к аспектам текста.

| Метод основанный на: | Точность | Временные затраты | Устойчивость к различным аспектам текста |
| --- | --- | --- | --- |
| многозначности | средняя | низкие | устойчив к лексическим особенностям |
| автореферировании | средняя | средние | учитывает семантику и синтаксис |
| ИИ | самая высокая | большие | учитывает широкий спектр факторов |

Итак, выбор между этими методиками зависит от конкретных целей исследования. Если необходима быстрая и относительно простая оценка сложности текстов, методика на основе многозначности слов может быть подходящей. Однако, если требуется более точная и устойчивая оценка, методика автоматического реферирования или методика с использованием ИИ могут быть более предпочтительными.

**Ограничения и возможные направления будущих исследований**

В данной главе мы обсудим ограничения нашего исследования и предложим несколько направлений для будущих исследований, которые могут углубить наше понимание оценки сложности текста и улучшить точность методов оценки.

**1. Ограничения исследования:**

Важно осознавать, что наше исследование также имеет свои ограничения:

**Ограниченный корпус текстов:** Наше исследование использовало ограниченный корпус текстов учебников по русскому языку. Для обобщения результатов на более широкий спектр текстов необходимо расширить выборку текстов и включить тексты из различных областей знаний и жанров.

**Ограниченное количество методов:** Мы рассмотрели три методики оценки сложности текста, однако существует множество других подходов и моделей, которые могут быть применены к данной задаче. Будущие исследования могут включать в себя сравнение с более широким спектром методов.

**Ограниченная глубина анализа:** Наш анализ фокусировался на определенных параметрах, таких как количество многозначных слов и длина автореферата. Дополнительные аспекты, такие как структура предложений, лексические и синтаксические особенности, могут оказывать влияние на сложность текста.

**2. Ограничения различных методов:**

**2.1 Методика на основе многозначности слов:**

**Ограниченность наличия многозначных слов:** Этот метод основан на анализе многозначных слов в тексте. Однако не все тексты могут содержать многозначные слова, и в таких случаях этот метод может оказаться менее эффективным.

**Ограничение в учете контекста:** Данный метод не всегда способен учесть контекст, в котором используются многозначные слова. Оценка сложности текста, исходя только из наличия многозначных слов, может быть неполной.

**Индивидуальные факторы:** Наш метод не учитывает контекстные факторы или индивидуальные характеристики читателя. Эти аспекты, включая контекст, в котором представлен текст, и уникальные характеристики индивидуальных читателей, играют значительную роль в понимании текста и должны рассматриваться в будущих исследованиях.

**Требует большого объема данных:** Для эффективной работы этой методики требуется большой корпус текстов с различными уровнями сложности и достаточным количеством многозначных слов. Это может быть сложно получить в некоторых случаях.

**Ограниченность языковой модели:** Эффективность этой методики может зависеть от языковой модели, используемой для анализа многозначных слов. При переходе к другому языку или диалекту может потребоваться создание новой модели или адаптация существующей.

**2.2 Методика на основе автоматического реферирования:**

**Ограничение в качестве реферирования:** Этот метод оценивает сложность текста на основе длины автореферата. Однако качество реферирования текста может варьироваться в зависимости от используемых алгоритмов, и это может повлиять на точность оценки.

**Неучет семантики:** Метод не учитывает семантическую сложность текста, и тексты с похожими авторефератами могут быть оценены одинаково, несмотря на различия в смысле.

**Зависимость от качества реферирования:** Результаты этой методики могут сильно зависеть от качества алгоритмов реферирования, и плохое качество реферата может привести к неправильным оценкам сложности.

**Зависимость от языковых особенностей:** Эффективность методики может зависеть от языковых особенностей, таких как структура предложений и особенности синтаксиса. Модель, обученная на одном языке, может не так хорошо работать на другом.

**Влияние длины текста:** Длина текста может влиять на точность оценки сложности. В некоторых случаях длинные тексты могут быть недооценены, так как авторефераты могут быть слишком краткими для передачи всей информации.

**2.3. Методика с использованием искусственного интеллекта (ИИ):**

**Ограничение в обучающих данных:** Для обучения модели машинного обучения требуется большой и разнообразный корпус текстов. Ограничения в данных могут привести к недообучению или переобучению модели.

**Необходимость ресурсов:** Обучение и использование моделей машинного обучения может требовать больших вычислительных ресурсов и времени, что может быть ограничивающим фактором.

**Чувствительность к качеству данных:** Качество результатов, полученных с помощью моделей ИИ, сильно зависит от качества входных данных. Некачественные или неточные тексты могут повлиять на точность оценки сложности.

**Ограниченность функциональности:** Модель может оценивать сложность текста на основе определенных признаков, но не всегда учитывать сложность в широком смысле, такую как структура текста или степень логической связности.

**Неоднородность структуры текстов:** Некоторые тексты могут иметь нестандартную структуру, такую как диалоги, таблицы, формулы и др., которые могут быть сложными для анализа моделями машинного обучения.

**Сложность обработки специализированных текстов:** Тексты в специализированных областях, таких как медицина или наука, могут содержать термины и концепции, которые необходимо правильно интерпретировать. Модели машинного обучения могут потребовать дополнительной настройки для таких текстов.

**Ограниченность языковой модели:** Эффективность этой методики может зависеть от языковой модели, используемой для анализа многозначных слов. При переходе к другому языку или диалекту может потребоваться создание новой модели или адаптация существующей.

Важно отметить, что каждая из этих методик имеет свои сильные стороны и ограничения. Исследователи и практики должны учитывать эти ограничения, важно комбинировать методики и использовать комплексный подход для оценки сложности текста, чтобы получить более точные результаты. При выборе метода оценки сложности текста нужно отталкиваться от конкретной задачи и целевой аудитории.

**Направления для будущих исследований:**

На основе результатов нашего исследования и выявленных ограничений, мы предлагаем следующие направления для будущих исследований:

* Расширение выборки текстов: Для создания более обобщенных и применимых результатов следует активно работать над расширением корпуса текстов, на которых проводятся оценки сложности. Включение разнообразных текстов из разных жанров (художественная литература, научные статьи, новости, технические материалы) и из разных областей знаний (медицина, физика, искусство) позволит учесть множество факторов, влияющих на сложность.
* Использование более сложных моделей машинного обучения: Развитие и применение более сложных алгоритмов машинного обучения, таких как глубокие нейронные сети, может значительно улучшить точность оценки сложности текста. Эти модели способны учиться из более сложных паттернов и зависимостей в тексте.
* Учет дополнительных аспектов текста: Оценка сложности текста может стать более точной, если учитывать дополнительные лингвистические и семантические аспекты текста, такие как сложность лексических единиц, семантическая структура и синтаксические особенности. Это поможет создать более полную картину сложности.
* Адаптация методов для специфических целей: Разработка методов оценки сложности, специально адаптированных для конкретных образовательных или информационных целей, позволит создавать более релевантные инструменты для конечных пользователей. Например, методы оценки сложности текстов для школьников могут отличаться от методов для научных исследователей.
* Оценка устойчивости методов: Понимание, как различные методы оценки сложности текста работают на разных языках, в разных культурных контекстах и для разных возрастных групп, является важным аспектом будущих исследований. Адаптация методов к разным условиям использования может сделать их более универсальными.
* Интерпретация результатов: Психолингвистические исследования могут помочь глубже понять, какие именно аспекты текста влияют на его восприятие как сложного или легкого. Это может включать в себя анализ когнитивных процессов, которые происходят при чтении текста разной сложности.
* Интеграция методов: Комбинирование различных методов оценки сложности текста может улучшить точность оценки и сделать ее более надежной. Это может быть особенно полезным при работе с текстами, которые могут быть оценены разными способами.
* Учет контекста: Учитывать контекст, в котором текст используется, может помочь улучшить оценку сложности. Это может включать в себя анализ целевой аудитории, цели чтения и предполагаемое использование текста. Например, текст, который сложен для школьников, может быть легким для профессионалов в данной области.
* Многоуровневая оценка: Разработка методов, способных оценивать сложность текста на нескольких уровнях, таких как лексический, семантический и синтаксический, может улучшить точность и полноту оценки. Это позволит лучше адаптировать тексты к разным уровням читателей.
* Персонализированные модели: Создание персонализированных моделей оценки сложности, учитывающих индивидуальные характеристики читателя, такие как возраст, образование и опыт чтения, может улучшить релевантность оценки для конкретных пользователей. Это особенно важно для образовательных и информационных платформ.
* Разработка приложений: Наши результаты могут быть использованы для создания приложений и инструментов, которые помогут пользователям оценивать сложность текстов и делать их более доступными для разных аудиторий. Это может быть полезно в образовании и информационных технологиях.

Итак, несмотря на ограничения, наше исследование предоставляет важные научные и практические выводы о методах оценки сложности текста. Будущие исследования в данной области могут внести дополнительный вклад в улучшение точности и применимости таких методов.

**Заключение**

В данной диссертации были исследованы методы оценки сложности текста с целью улучшения понимания процессов чтения и обеспечения доступности текстов для различных аудиторий. Исследование началось с обзора существующих методов и подходов к оценке сложности текста, что позволило выделить основные направления и проблемы в данной области.

Основной целью диссертации было разработать и сравнить методики оценки сложности текста, учитывающие различные аспекты текстов, такие как многозначность слов, длина автореферата и использование искусственного интеллекта. Результаты исследования показали следующее:

1. Методика на основе многозначности слов демонстрировала высокую точность в предсказании сложности текстов для разных возрастных групп, но имела ограничения в учете контекста и требовала большого объема данных.
2. Методика на основе автоматического реферирования позволяла оценивать сложность текстов на основе длины автореферата и демонстрировала обратную зависимость между уровнем образования и этой метрикой.
3. Методика с использованием искусственного интеллекта (ИИ) показала самую высокую точность в оценке сложности текстов среди трех методик благодаря способности выявлять сложные закономерности в тексте.

Важным практическим результатом этой диссертации стало создание телеграм-бота, который может оценивать сложность текстов. Этот бот представляет собой доступный инструмент для пользователей, который может помочь им быстро определить уровень сложности текста и адаптировать чтение под свои потребности. Это хорошее достижение, потому что такой бот может быть полезен как для студентов, которые хотят выбирать тексты, соответствующие их уровню знаний, так и для людей с разными уровнями грамотности, которые могут получать тексты, подходящие именно для них.

В заключении хотелось бы подчеркнуть важность проделанной работы и ее значимость для образования и информационных технологий. Методы оценки сложности текста, рассмотренные в этой диссертации, могут быть применены в образовательных платформах для автоматической адаптации текстов под уровень знаний студентов. Они также могут быть использованы для создания инструментов, помогающих людям с разными уровнями грамотности и образования лучше понимать и использовать текстовую информацию.

Однако важно отметить, что данная диссертация имеет свои ограничения. Она ограничивается определенными методами оценки и не учитывает все возможные аспекты сложности текста. Кроме того, оценка сложности текста зависит от контекста и цели чтения, что также требует дальнейших исследований.

Дальнейшие исследования в этой области могут включать в себя расширение выборки текстов, использование более сложных моделей машинного обучения, учет дополнительных аспектов текста, адаптацию методов для специфических целей, а также более глубокий анализ влияния различных аспектов текста на его восприятие как сложного или легкого. Оценка устойчивости методов к различным языкам, культурам и возрастным группам также остается актуальной задачей.

Подводя итоги, данная диссертация внесла вклад в понимание и развитие методов оценки сложности текста и создала практически полезное приложение в виде Telegram-бота. Надеемся, что результаты данного исследования будут полезными как для научного сообщества, так и для широкой практической аудитории, и будут способствовать развитию области образования и информационных технологий.

**Список литературы**

1. Банерджи С., Педерсен Т. и Маквикар М. Использование устранения неоднозначности смысла слов для измерения семантической сложности текстов. (2014)
2. Шато, Д., и Джаред, Д. Знакомство с процессами печати и распознавания слов. Память и когнитивные способности, 28(1), 143-153 (2000)
3. Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. DRC: двухканальная каскадная модель визуального распознавания слов и чтения вслух. Психологический обзор, 108(1), 204-256 (2001)
4. Янсен Ф. Дж., ван дер Маас Х. Л. и Кемпен Г. Детерминанты качества текста: взаимодействие навыков чтения, возраста и знакомства с темой. Научные исследования чтения, 11(3), 219-247 (2007)
5. Джонсон Т. И Пирсон П. Д. Взаимосвязь между структурой предложения и сложностью текста. Чтение на иностранном языке (2018)
6. Кутузов А. Б. Методы определения сложности текста в рамках переводческого анализа. Вестник Нижегородского государственного лингвистического университета имени Н. А. Добролюбова, (4), 109-116. (2009)
7. Лапошина А. Н. Анализ релевантных функций для автоматического определения сложности русского текста как иностранного. Аналитика и управление данными в образовании и науке, 1(3), 81-87 (2009)
8. Ландауэр Т. К. и Дюме С. Т. Решение проблемы Платона: теория латентного семантического анализа приобретения, индукции и репрезентации знаний. Психологический обзор, 104(2), 211-240 (1997)
9. Лесо, Н. К., и Киффер, М. Дж. Исследуют источники трудностей с пониманием прочитанного среди изучающих языки меньшинств и их одноклассников в раннем подростковом возрасте. Американский научно-образовательный журнал, 47(2), 596-632 (2010)
10. Мицуги, М. Инструкция по изучению многозначных слов английского языка на основе схем: влияние инструкции и восприятие учащимися. Журнал Пан-Тихоокеанской ассоциации прикладной лингвистики, 21(1), 21-43 (2017)
11. Молер М., Раух Р. и Циммерман Х. Оценка сложности текстов с использованием устранения неоднозначности смысла слов (2011)
12. Нейшн, К., & Сноулинг, М. Дж. Семантическая обработка и развитие навыков распознавания слов: свидетельства детей с трудностями понимания прочитанного. Журнал памяти и языка, 39(1), 85-101 (1998)
13. Оукс, М. П. (2016). Оценка читабельности текста с использованием лексической неоднозначности.
14. Перфетти, К. А., & Хогабоам, Т. У. Взаимосвязь между расшифровкой отдельных слов и навыком понимания прочитанного. Журнал педагогической психологии, 67(4), 461-469 (1975)
15. Перфетти К. А., Ланди Н. и Окхилл Дж. Приобретение навыка понимания прочитанного. Справочник по детской грамотности, 291-310 (2005)
16. Перфетти К. А., Маррон М. А. и Фольц П. В. Источники сбоев в понимании. (1996)
17. Рудницкая Е., Пиларская А. Автоматическая оценка читабельности текста: обзор. Обзор компьютерных наук, 27, 119-129. (2018)
18. Саэмен, Р. А. Влияние общеизвестных значений на определение неясных значений многозначных слов в контексте. Отчет о проекте по обучению начальным языковым навыкам с индивидуальным руководством. (1970)
19. Шмитт, Н. Исследование лексики: Руководство по исследованию лексики. Прыгун. (2010)
20. Соловьев В. Д. Статистические методы анализа диахронических текстовых корпусов как инструмент изучения языковой динамики. Материалы международной конференции. Русский язык: функционирование и развитие. Казань: Казанский университет. С.47-52. (2012)
21. Соловьев В., Андреева М., Солнышкина М., Замалетдинов Р., Данилов А., Гайнутдинова Д.Вычисление оценок конкретности наиболее частотных слов русского и английского языков: контрастивный подход. Материалы - Международная конференция по разработкам в области проектирования электронных систем, DeSE, октябрь-2019, арт. № 9073272, стр. 403-408. (2019)
22. Солнышкина С. И., Кисельников А. С. Сложность текста: этапы изучения в российской прикладной лингвистике. Вопросы когнитивной лингвистики. Выпуск 1. Стр. 18-40. (2022)
23. Солнышкина М.И., Харькова Е.В., Казачкова М.Б. Структура межъязыковых различий: значение и контекст ‘читабельности" и его русского эквивалента "читабельность’
24. Журнал языка и образования, 6 (1), стр. 103-119 (2020)
25. Солнышкина М., Иванов В., Соловьев В. Формула удобочитаемости текстов на русском языке: модифицированная версия. Конспекты лекций по информатике (включая подсерии Конспектов лекций по искусственному интеллекту и конспекты лекций по биоинформатике), 11289 LNAI, стр. 132-145 (2018)
26. Кембриджский словарь, доступный по адресу: https://dictionary.cambridge.org/dictionary/english/readability
27. RuWordNet, доступный по адресу: https://www.ruwordnet.ru/en (Дата обращения 6 июня 2023 года)
28. Фенг, Л., Янче, М., Хенерфаут, М. и Эльхадад, Н. Сравнение функций для автоматической оценки удобочитаемости. На 23-й Международной конференции по компьютерной лингвистике (COLING). (2010)
29. Булте, Б., и Хаусен, А. Определение и практическая реализация сложности L2. Показатели производительности и квалификации L2: Сложность, точность и свободное владение SLA. Амстердам: Издательство Джона Бенджаминса, 21-46 (2012)
30. Фулчер Г. Сложность и доступность текста: формулы для чтения и экспертные оценки. Система 25:497-513 (1997)
31. Раукко Дж. Многозначность как сложность. SKY Journal 10, 145-170 (1997)