Учреждение образования
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ»
Кафедра интеллектуальных информационных технологий

Отчет по лабораторной работе №1 по курсу “Естественные языковые интерфейсы интеллектуальных систем” на тему: “Разработка автоматизированной системы формирования словаря естественного языка”

Выполнили студены группы 021701: Кулак П.О.
Седеневский А.М.
Малаев А.А.

Проверил: Крапивин Ю.Б.

Минск 2023

Цель работы:
Освоить принципы разработки прикладных сервисных программ для решения задачи автоматического лексического и лексико-грамматического анализа текста естественного языка.
Вариант задания: 6
Расширение файла: doc или docx
Язык текста: русский.
Задание - 2:
Список слов, упорядоченный по алфавиту и включающий только лексемы с дополнительно оформленными записями для образования словоформ. В этих записях должна храниться следующая информация: основа слова; часть речи; окончания слова, соотнесенные с соответствующей морфологической информацией: род, падеж, число и т.п. При работе с таким словарем должны быть обеспечены средства генерации той или иной словоформы в соответствии с введенными «правилами».

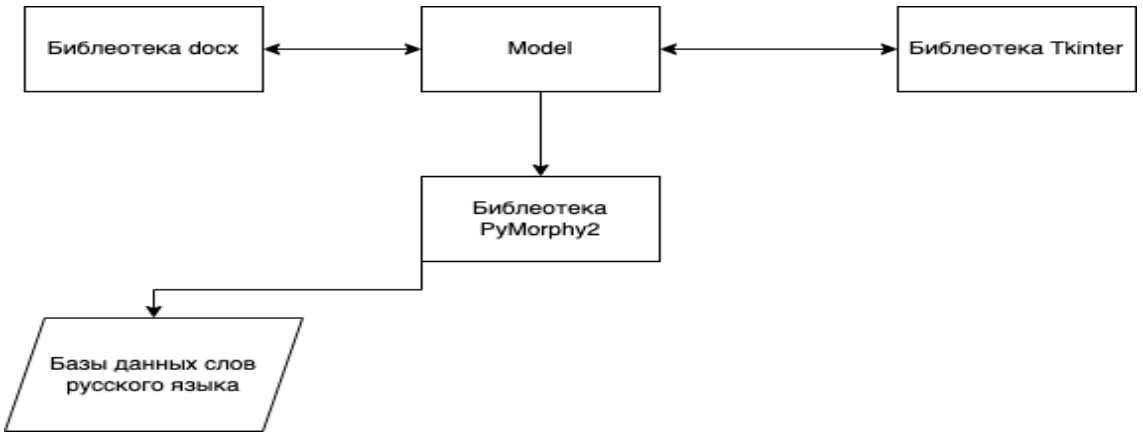
Результат работы:

Описание системы:

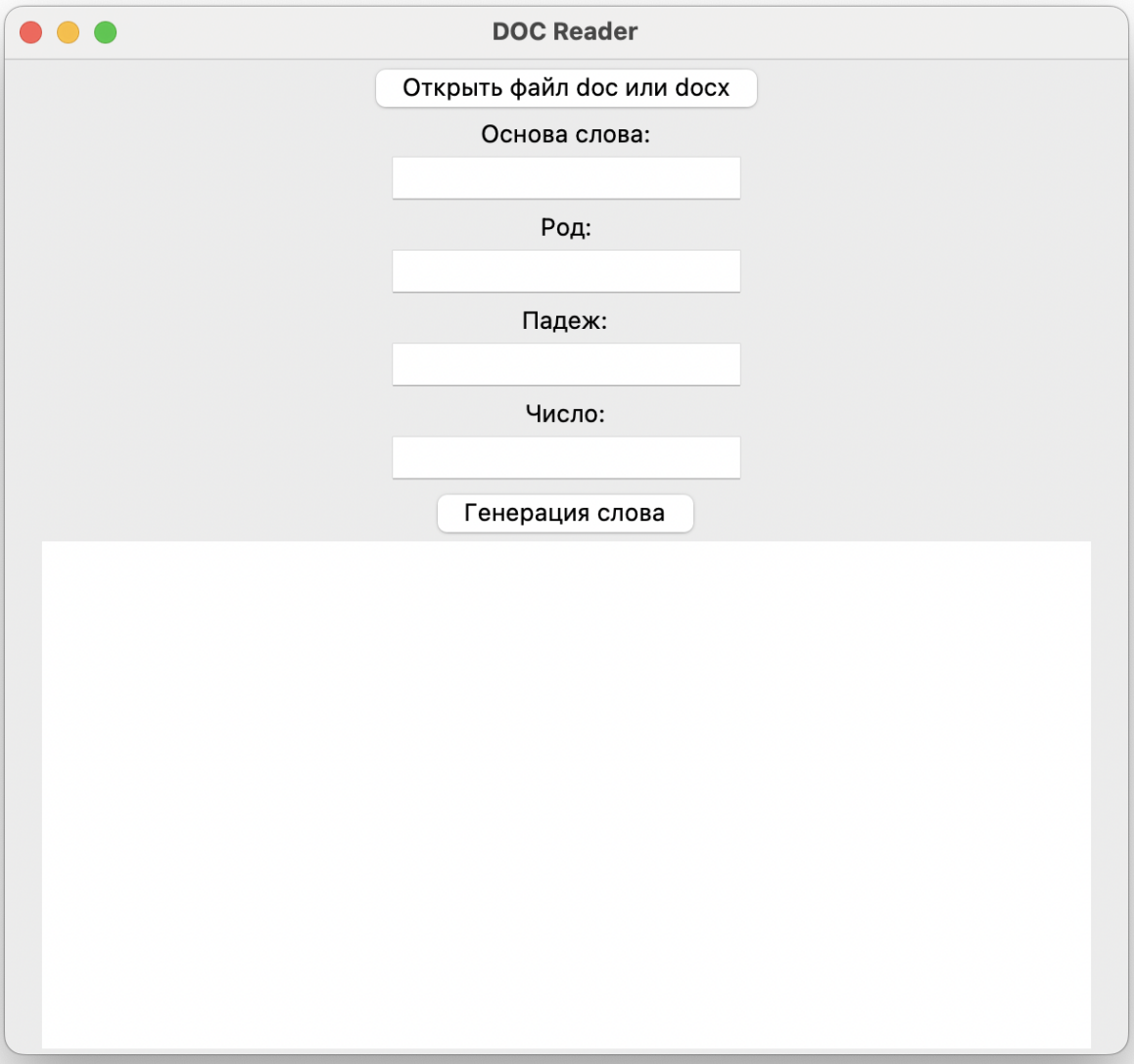
- на входе – естественно-языковой текст;
- на выходе – упорядоченный по алфавиту перечень извлеченных из данного текста лексем естественного языка с дополнительной информацией о форме, в которой данные лексемы использовались в тексте;
- возможность ручного добавления текста из файла типа doc или docx;
- возможность ручного ввода, удаления, редактирования текста
- возможность редактирования полученного словаря, поиска по нему

Использованные технологии:

- Язык программирования - Python
- Библиотека для создания графического интерфейса - Tkinter
- Библиотека для анализа естественно-языкового текста (Русс. яз.) - PyMorphy2
- Библиотека для обработки Doc-файлов - docx



Интерфейс:



Описание интерфейса:

- Кнопка “Открыть файл doc или docx”

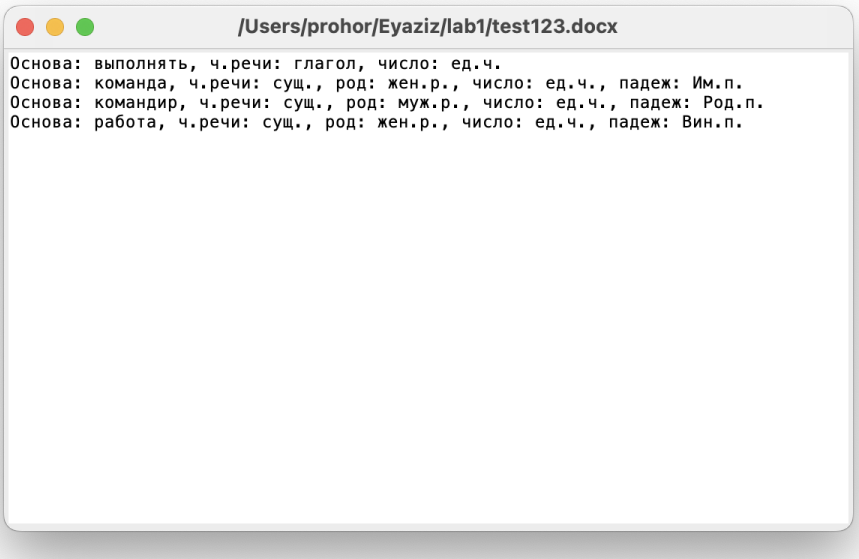
  Кнопка позволяет загрузить файл типа doc или docx, из которого будет взят текст для создания словаря и анализа и будет создан словарь и показан в новом окне;
- Поле ввода естественно-языкового текста. Имеет 2 состояния:
  - Пустое;
  - Заполненное
    - Заполнение вручную (ввод с клавиатуры);

- Таблица с возможностью скролла для представления словаря. Имеет поля:
    - Основа - базовая часть слова, на которую могут накладываться грамматические изменения;
    - Часть речи - категория слов, обозначающая их грамматическую роль в предложении;
    - Род - грамматическая категория, указывающая на соотношение существительного с мужским, женским или средним родом;
    - Число - грамматическая категория, указывающая на количество объектов, которые обозначаются словом;
    - Падеж - грамматическая категория, определяющая синтаксическую роль слова в предложении.
- Кнопка “Генерация слова” позволяет использовать введенные пользователем параметры для генерации слова согласно этим параметрам;

Демонстрация работы дополнительного функционала:

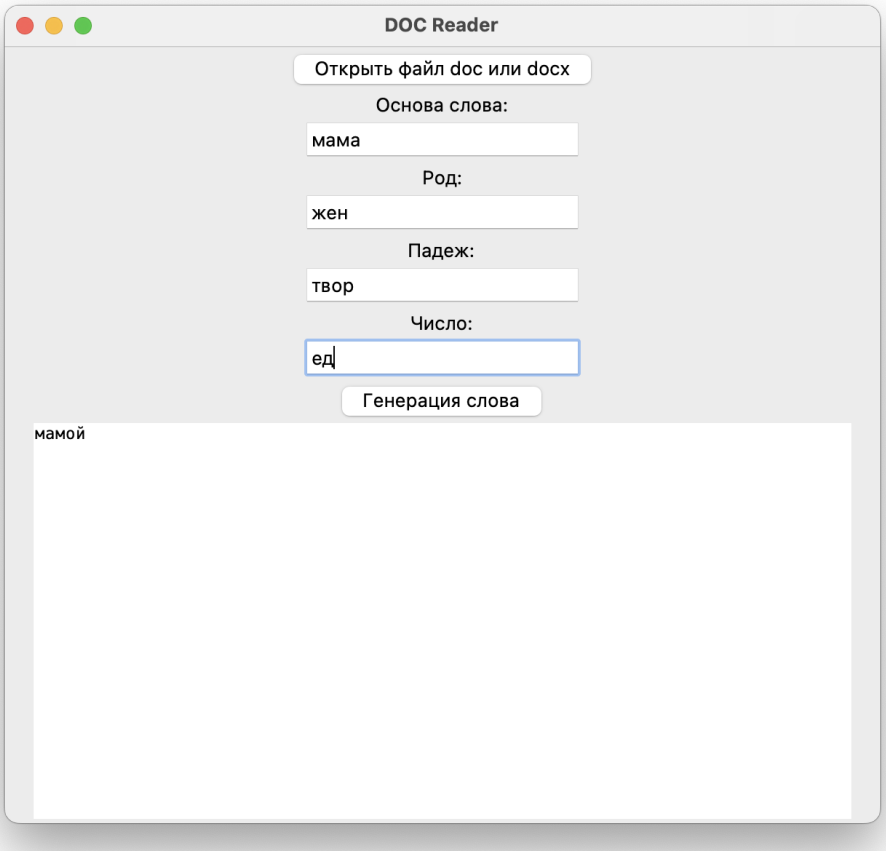
- Функционал редактирования:

    Нажать на кнопку “Открыть файл doc или docx”



Откроется новое окно с составленным словарем из текста.

- При заполненных полях формы, нажатие кнопки «Генерация слова» сгенерирует слово по заданным параметрам.



Описание алгоритмов:

Две основные операции, которые умеет делать морфологический анализатор - разбор слов текста файла и запись их в словарь с полной информацией о них, генерация нового слова по заданным параметрам.

Описание быстродействия, оптимизации и правильности:

- словарь со всеми сопутствующими данными занимает около 15Мб оперативной памяти;
- скорость морфологического разбора слов - от 10-20 тыс. слов/сек до >100тыс. слов/сек
- скорость морфологического разбора предложений и слов в них - 532 прдл./сек для тестовой выборки
- правильность сегментации:
    - токенизация - 9 ошибок на 1000 слов, 2.9 сек на обработку датасета

- сегментация на предложения - 52 ошибки на 1000 предложений, 6.1 сек на обработку датасета.

Выводы:

Разработанный программный продукт является достаточно оптимизированным, имеет высокую степень правильности и быстродействия.
Использованный для проведения морфологического разбора инструментарий находится в стадии активного совершенствования, поэтому в будущем можно ожидать ещё большей производительности, быстродействия и правильности результатов.
Разработанный продукт можно назвать дружелюбным к пользователю - он даёт возможность как загрузить текст из файла типа doc или docx, а затем получить информацию о лексемах в этом тексте, так и сгенерировать любую лексему по заданным параметрам. Результаты представляются в понятном и расширенном виде, однако порог входа для использования результатов требует понимания граммем.
Инструмент может быть использован экспертами и учащимися, работающими с русским языком.

1.Синонимия: определение, виды, примеры, достоинства и недостатки.
Определение:
Синонимия - это лингвистическое явление, при котором имеются два или более слова, которые имеют сходный или почти одинаковый смысл, но различаются по звучанию, орфографии, морфологии или контексту.
Виды:
Синонимия может быть разделена на следующие виды:

1. Полная синонимия - это, когда слова имеют абсолютно одинаковый смысл в любых контекстах, например: автомобиль и машина.
2. Частичная синонимия - это, когда слова имеют общий смысл, но не полностью совпадают по значению, например: дом и здание.
3. Эквивалентность - это, когда слова на разных языках имеют одинаковый или почти одинаковый смысл, например: love и любовь.

Примеры:
Примеры полной синонимии: песня и песенка, автобус и автобусик, красивый и прекрасный.
Примеры частичной синонимии: машина и автомобиль, книга и том, ручка и перо.
Примеры эквивалентности: house и дом, mother и мать, dog и собака.
Достоинства:

1. Синонимы позволяют избегать повторов в речи или тексте.
2. Использование синонимов может улучшить стиль и выразительность речи или текста.
3. Синонимы могут помочь сделать текст более доступным для аудитории, которая может быть не знакома с некоторыми терминами.

Недостатки:

1. Некоторые синонимы могут быть не совсем точными или иметь небольшие оттенки значения, которые могут быть упущены.
2. Использование неправильного синонима может привести к непониманию или неверному толкованию текста.
3. В некоторых случаях синонимы могут быть недостаточно употребляются в обычной речи и использование их может показаться неестественным.

2. Автоматизация обработки текста: этап семантического анализа.

Этап семантического анализа в автоматизации обработки текста - это процесс выявления смысловых отношений между словами и фразами в тексте. Он является одним из ключевых этапов в обработке естественного языка и имеет решающее значение для многих задач обработки текста, таких как классификация текстов, извлечение информации и ответ на вопросы.

Семантический анализ включает в себя несколько подэтапов:

1. Морфологический анализ - процесс разбора текста на составляющие элементы (слова и морфемы) и определения их грамматических характеристик, таких как падеж, род, число, время и т.д.
2. Синтаксический анализ - процесс анализа структуры предложения и выявления связей между словами. Это включает в себя определение роли каждого слова в предложении и его зависимости от других слов.
3. Семантический анализ - процесс определения смысловых отношений между словами и фразами. Включает в себя определение значения каждого слова в контексте и выявление связей между словами, таких как синонимия, антонимия, гиперонимия, гипонимия и т.д.
4. Дискурсивный анализ - процесс анализа текста на уровне высказываний и определение связей между ними. Это включает в себя выявление темы текста, выделение ключевых идеи и определение отношений между ними.

На этапе семантического анализа используются различные методы и инструменты, такие как машинное обучение, нейронные сети, алгоритмы классификации, логический вывод и т.д. В результате семантического анализа текста может быть получена более точная и полная информация о его содержании, что может быть использовано для автоматической обработки текста и принятия решений на его основе.

3. Типовая структура базы знаний для решения задач автоматической обработки текста естественного языка.

Типовая структура базы знаний для решения задач автоматической обработки текста естественного языка может включать следующие элементы:

1. Словарь - набор слов и их грамматических характеристик, используемых в тексте. В словаре могут быть также указаны значения слов и их синонимы.
2. Грамматический анализатор - модуль, который определяет грамматические характеристики каждого слова в тексте, такие как падеж, род, число, время, залог и т.д. Это позволяет правильно понимать структуру предложений.
3. Семантический анализатор - модуль, который определяет смысловые отношения между словами в тексте. Он может использовать словарь и другие ресурсы для определения значения каждого слова в контексте и выявления связей между словами, таких как синонимия, антонимия, гиперонимия и т.д.
4. Модуль извлечения информации - модуль, который извлекает информацию из текста, связанную с задачей обработки текста, например, имена, даты, адреса, ключевые фразы и т.д.
5. Модуль классификации текста - модуль, который классифицирует текст по определенным категориям, например, новости, обзоры, научные статьи и т.д. Это может быть использовано для решения задач, связанных с автоматическим поисковым обращением.

6. Модуль синтеза речи - модуль, который преобразует текст в речь. Это может быть полезно для создания голосовых интерфейсов и других приложений, которые требуют преобразования текста в речь.
7. База знаний - набор правил и знаний, используемых для решения конкретных задач в области обработки текста. База знаний может быть разработана на основе опыта экспертов и данных, собранных из различных источников.

Каждый из этих модулей может работать отдельно или вместе с другими модулями, образуя комплексную систему автоматической обработки текста естественного языка.

4. Уровни изучения текста, связь с разделами лингвистики.

Уровни изучения текста - это различные аспекты, на которые можно обратить внимание при анализе текста. Каждый уровень обычно связан с определенным разделом лингвистики. Рассмотрим основные уровни изучения текста и связанные с ними разделы лингвистики:

1. Фонетический уровень - на этом уровне анализируются звуки и звуковые характеристики, такие как интонация, темп речи и т.д. Связан с фонетикой.
2. Морфологический уровень - на этом уровне анализируются морфемы (минимальные значимые единицы слова) и их грамматические характеристики. Связан с морфологией.
3. Синтаксический уровень - на этом уровне анализируется структура предложений, связи между словами и их роли в предложении. Связан с синтаксисом.
4. Семантический уровень - на этом уровне анализируется значение слов и выражений, а также смысловые отношения между ними. Связан с семантикой.
5. Прагматический уровень - на этом уровне анализируется использование языка в конкретных коммуникативных ситуациях и целях. Связан с прагматикой.

Каждый уровень изучения текста важен для полного понимания и анализа текста. При этом важно учитывать, что эти уровни не являются отдельными и изолированными друг от друга, а представляют собой сложную и взаимосвязанную систему, где каждый уровень влияет на другие.

5. Прикладная лингвистика - это область знаний, которая занимается применением теоретических знаний лингвистики к практическим проблемам, связанным с использованием языка в реальных ситуациях. Она занимается анализом, описанием и практическим применением языковых знаний в различных областях деятельности.

Компьютерная лингвистика и прикладная лингвистика тесно связаны друг с другом, но имеют и некоторые различия:

- Компьютерная лингвистика фокусируется на разработке и применении технологий и методов обработки естественного языка, тогда как прикладная лингвистика применяет лингвистические знания для решения конкретных языковых проблем в различных сферах деятельности, таких как образование, межкультурная коммуникация, перевод и т.д.

- Компьютерная лингвистика использует технологии и методы искусственного интеллекта, машинного обучения и компьютерных алгоритмов для обработки естественного языка, тогда как прикладная лингвистика обычно использует более традиционные методы и технологии для решения языковых проблем.

- В области компьютерной лингвистики уделяется больше внимания разработке и улучшению технологий для автоматического анализа и генерации естественного языка, тогда как прикладная лингвистика более ориентирована на практические задачи, такие как создание учебных материалов для изучения языка, обеспечение эффективной коммуникации на рабочем месте и т.д.

Таким образом, прикладная лингвистика и компьютерная лингвистика имеют много общих точек соприкосновения, но каждая из них имеет свои специфические задачи и подходы к решению языковых проблем.

6. Компьютерная лингвистика возникла из необходимости обработки больших объемов текстовой информации и автоматической обработки естественного языка. Возникновение компьютерной лингвистики обусловлено следующими причинами:

- Развитие компьютерных технологий: С появлением компьютеров и программных средств для обработки текстов, стало возможным создание автоматических систем обработки естественного языка.

- Увеличение объема текстовой информации: С ростом объема текстовой информации стало необходимым разрабатывать методы автоматической обработки и анализа этой информации.

- Появление новых форм коммуникации: С развитием интернета и социальных сетей появились новые формы коммуникации, которые требуют автоматической обработки естественного языка, такие как обработка сообщений и комментариев в социальных сетях.

- Необходимость автоматизации процессов: В различных сферах деятельности, таких как банковское дело, медицина, право, наука и техника, стала необходима автоматизация процессов, в том числе и обработки текстовой информации.

- Развитие искусственного интеллекта: Компьютерная лингвистика является частью области искусственного интеллекта, которая занимается разработкой и применением методов и технологий для создания умных систем и решения сложных задач.

Таким образом, компьютерная лингвистика возникла из необходимости решения практических проблем, связанных с обработкой естественного языка, и является важной областью в области информационных технологий.