# Лабораторная работа №1. Разработка автоматизированной системы формирования словаря естественного языка

**Цель работы:**

Освоить принципы разработки прикладных сервисных программ для решения задачи автоматического лексического и лексико-грамматического анализа текста естественного языка.

**Задачи лабораторной работы:**

1. Познакомиться с назначением, структурой и функциональностью, предоставляемой базовым ЛП для решения задачи автоматического лексического и лексико-грамматического анализа ТЕЯ.
2. Закрепить навыки программирования при решении задач автоматической обработки ТЕЯ.

**Методические указания:**

Требуется спроектировать и программно реализовать структуры хранения данных, алгоритмы их обработки, необходимые в рамках следующих базовых требований к разрабатываемому приложению:

* входные данные – текст заданного естественного языка;
* выходные данные – перечень лексем с дополнительной информацией согласно задания;
* взаимодействие с пользователем посредствам графического интерфейса (интерфейс должен быть интуитивно-понятным и дружественным пользователю);
* наличие системы средств помощи пользователю;
* обеспечение возможности построения, сохранения, просмотра, редактирования, пополнения, фильтрации и поиска по заданному условию, документирования автоматически получаемого словаря либо заданной его части;
* поддержка различных форматов представления входных данных (TXT, RTF, PDF, DOC, DOCX).

Рекомендуется использовать функциональность стандартной, а также специализированных библиотек языка программирования Python для обработки естественного языка, например, nltk.

Вариант задания выбирается студентом самостоятельно и согласовывается с преподавателем. Средства разработки выбираются студентом самостоятельно. Защита лабораторной работы предполагает демонстрацию работоспособности всех реализованных функций в соответствии с требованиями.

**Требования к отчету:**

В отчете представить, в том числе графически, используя такие программные средства, как Microsoft Visio или Draw.io:

* структурно-функциональную схему разработанного приложения;
* описание структур хранения данных, алгоритмов их обработки, необходимых для реализации базовых требований к разработанной программе;
* оценку быстродействия приложения;
* выводы по работе и по перспективам использования приложения.

Отчет предоставить для проверки в электронном виде.

**Варианты заданий:**

**Задание 1.** Список слов, упорядоченный по алфавиту и включающий как лексемы, так и словоформы, с указанием частоты встречаемости каждой из форм. Для словоформ пользователю должна быть предоставлена возможность вводить дополнительную морфологическую информацию, а именно, отнесение слова к соответствующей части речи, указание рода, числа, падежа и т.п. При этом морфологическая информация может быть оформлена как отдельная неформатированная запись, т.е. это просто текст, который пользователь может оформлять произвольным образом.

**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».

**Задание 3.** Список слов, упорядоченный по алфавиту и включающий только лексемы с дополнительно оформленными записями для образования словосочетаний. В этих записях должны храниться слова (точнее, ссылки на эти слова), с которыми данное слово может сочетаться. При этом возможно обеспечение автоматизированного извлечения из исходных текстов типовых словосочетаний с их последующей обработкой. \*Можно добавить средства синтеза словосочетаний, но при этом в словарь необходимо будет добавить правила формирования словоформ (см. задание 2).

**Задание 4.** Список слов, упорядоченный по алфавиту и включающий только лексемы с дополнительно оформленными записями о месте и роли данного слова в составе предложения. К такой информации относится описание того, каким членом предложения может быть данное слово и в какой форме (падеж, число, время и т.п.). Например, если это существительное в именительном падеже, то оно может выступать в роли подлежащего; если это существительное в родительном падеже, то оно может быть дополнением; если это прилагательное, то оно может быть определением и т.п.

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Язык текста** | **Формат входного документа** | **Вариант задания** |
| 1 | Русский | TXT, RTF | Задание 1 |
| 2 | Русский | PDF | Задание 1 |
| 3 | Русский | DOC, DOCX | Задание 1 |
| 4 | Русский | TXT, RTF | Задание 2 |
| 5 | Русский | PDF | Задание 2 |
| 6 | Русский | DOC, DOCX | Задание 2 |
| 7 | Русский | TXT, RTF | Задание 3 |
| 8 | Русский | PDF | Задание 3 |
| 9 | Русский | DOC, DOCX | Задание 3 |
| 10 | Русский | TXT, RTF | Задание 4 |
| 11 | Русский | PDF | Задание 4 |
| 12 | Русский | DOC, DOCX | Задание 4 |
| 13 | Английский | TXT, RTF | Задание 1 |
| 14 | Английский | PDF | Задание 1 |
| 15 | Английский | DOC, DOCX | Задание 1 |
| 16 | Английский | TXT, RTF | Задание 2 |
| 17 | Английский | PDF | Задание 2 |
| 18 | Английский | DOC, DOCX | Задание 2 |
| 19 | Английский | TXT, RTF | Задание 3 |
| 20 | Английский | PDF | Задание 3 |
| 21 | Английский | DOC, DOCX | Задание 3 |
| 22 | Английский | TXT, RTF | Задание 4 |
| 23 | Английский | PDF | Задание 4 |
| 24 | Английский | DOC, DOCX | Задание 4 |