Министерство образования Республики Беларусь

Учреждение Образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Кафедра интеллектуальных информационных технологий

Лабораторная работа № 1

«РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ФОРМИРОВАНИЯ СЛОВАРЯ ЕСТЕСТВЕННОГО ЯЗЫКА»

|  |  |  |
| --- | --- | --- |
| Проверил: |  | Выполнил: |
| Крапивин Ю. Б. |  | Рахимбердиев Т. Ж  Казибаев С.  Моминов С |
|  |  |  |

Минск 2022

**1. Цель работы**

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

**2. Ход работы**

Вариант : Английский язык, формат PDF, задание .

**Решение**

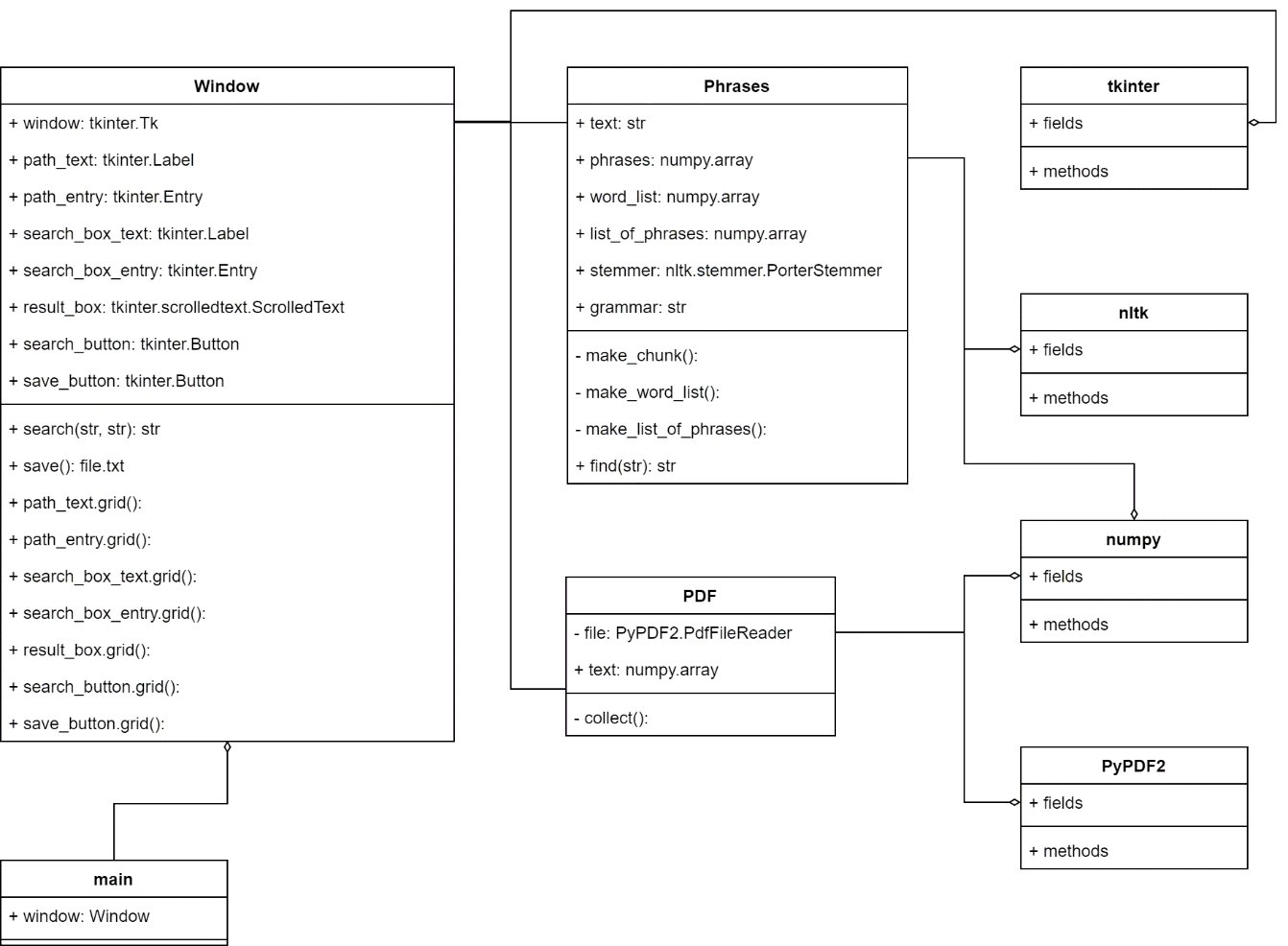


Рис 1. Диаграмма классов.

Для хранения данных были использованы массивы numpy, для извлечения данных из файлов .pdf библиотека PyPDF2, для обработки текста была использована библиотека nltk, для графического интерфейса библиотека tkinter.

Было написано два класса:

1. Класс для сбора текста из PDF файла – считает количество страниц в файле, с каждой страницы собирает текст и заносит в отдельную ячейку массива numpy. При передаче в качестве текстового аргумента собирает все ячейки в одну текстовую строку.
2. Класс для создания массива содержащего слова и словосочетания – хранит в себе переданный в него текст, в котором находит словосочетания по заданному образцу, которые сохраняет в массив numpy. После, класс на основе полученных словосочетаний составляет массив numpy, в котором хранятся все слова, используемые в найденных ранее словосочетаниях. Далее, на основе двух полученных массивов класс находит соответствия, обходя каждое слово в массиве слов и сопоставляя его с каждым имеющимся словосочетанием, если слово из второго массива присутствует в словосочетании, словосочетание вместо со словом записываются во временный массив numpy, при нахождении еще одного соответствия – оно будет добавлено во временный массив. При окончании обхода массива словосочетаний для одного слова из второго массива, имеющийся временный массив будет добавлен в качестве ячейки в третий массив numpy – массив, содержащий слова и словосочетания, в которых они используются. После добавления временного массива в третий массив, временный обнуляется. Алгоритм прекращает работы после обхода всех слов из второго массива. Также класс имеет метод для поиска по третьему массиву на основе переданных символов, которые дополнительно проходят stemming, результаты возвращаются в виде строки.

**Результаты**

Была спроектирована и реализована система, имеющая графический пользовательский интерфейс, обеспечивающая возможность построения, сохранения, просмотра, редактирования, пополнения, фильтрации и поиска по заданному условию, документирования автоматически получаемого словаря либо заданной его части и имеющая систему помощи пользователю

**3. Выводы**

Полученная система обрабатывает PDF файл длинной в 4 страницы и размером 21 КБ менее чем за секунду. Файл длинной в 262 страницы и размеров 681 КБ за 5 минут, 10 секунд (или 310 секунд). Помимо словосочетаний система может находить и другие конструкции на основе POS тэгов.