Учреждение образования

“БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ ”

Кафедра интеллектуальный информационных технологий

Отчет по лабораторной работе №3

по курсу: “Естественно-языковой интерфейс интеллектуальных систем”

Выполнил студент группы 921702: Баранов А.Д.

Проверил: Крапивин Ю.Б.

Минск

2022

**Цель работы**: освоить на практике основные принципы автоматического реферирования документов.

**Вариант 18:** английский, испанский.

**Требования к разрабатываемой системе:**

* на входе – на входе – текстовые документы одинакового размера (например, 10 страниц формата А4), содержащие тексты из предметных областей на естественных языках согласно варианту подлежащие процедуре автоматического реферирования;
* на выходе – активная ссылка на исходный документ и построенный реферат документа, состоящий из 2-х разделов: классического реферата; реферата в виде списка ключевых слов;
* наличие средств сохранения в файл и распечатки полученной на выходе информации;
* интерфейс системы должен быть предельно простым и доступным для пользователей любого уровня, содержать понятный набор инструментов и средств, а также help-средства.

**Теоретические сведения**

**Весовые коэффициенты значимости терминов**

Функция автоматического реферирования – дать сжатое представление текстовой информации, позволяющее пользователю экономить время при поиске и отборе необходимой информации, т.е. «отсеивать» менее значимую информацию.

Можно выделить следующие типы рефератов:

- реферат в виде списка ключевых слов

- классический реферат

- структурированный реферат

- запросно-ориентированный реферат

**Реферат в виде ключевых слов** представляет собой список, возможно иерархический (в виде дерева) наиболее информативных слов и словосочетаний (именных групп) обрабатываемого документа. Такой реферат позволяет пользователю понять основные темы описанные в документе. Например: *лазер, лазерный луч, синий лазер, красный лазер, устройство* и т.д. Или в иерархическом виде*:*

*лазер*

*лазерный луч*

*синий лазер*

*красный лазер*

*устройство*

и т.д.

**Классический реферат** – это набор наиболее информативных предложений текста, возможно трансформированных (удаление вводных конструкций, замена анафорических местоимений и т.д. с целью улучшения связности реферата и уменьшения его объема).

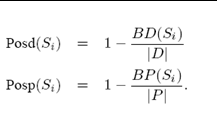
При построении классического реферата методом sentence extraction необходимо:

1. Вычислить веса слов документа. При этом слова из латинских букв, числа, стоп-слова - не учитываются. Базовый вес слова вычисляется по формуле TF\*IDF.
2. Вычислить веса предложений согласно формулам, приведенным ниже.
3. Осуществить генерацию реферата.

Этап генерации представляет собой выбор из исходного текста определенного количества предложений с наибольшим весом в той последовательности, в которой они идут в тексте. Рекомендуемый размер реферата 10 предложений.

Вес каждого предложения Si вычисляется произведением значений функций приведенных ниже.

Функции, характеризующие положение предложения в документе и положение в абзаце :



где

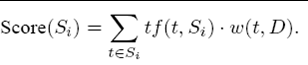
|D|- число символов в документе D, содержащем предложение Si;

BD(Si) – количество символов до Si в D(Si);

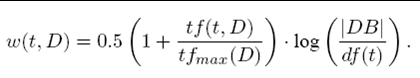
|P| - количество символов в абзаце P, содержащем предложение Si;

BP(Si) – количество символов до Si в абзаце.

**Модифицированная TFIDF функция:**



- частота термина t в предложении Si;



- частота термина t в документе D;

- количество документов, с термином t;

 - максимальная частота термина в документе D;

|DB| - количество документов.я

**Алгоритм на ML**

В алгоритме используются такие понятия, как подобие указанных структур графа.

Сначала объединяется весь текстовый корпус. Затем он разбивается на предложения. Векторы используются для построения векторов, куда вводится информация о количестве терминов в одном предложении. Из векторов строится матрица подобия, которая обрабатывается в графе. Затем рассчитывается рейтинг предложений по важности, и наиболее важные предложения включаются в окончательный реферат

**Реализация**

Система состоит из двух частей: серверной и клиентской. В серверной части находится код на языке Python, который реализует TFIDF и ML методы.

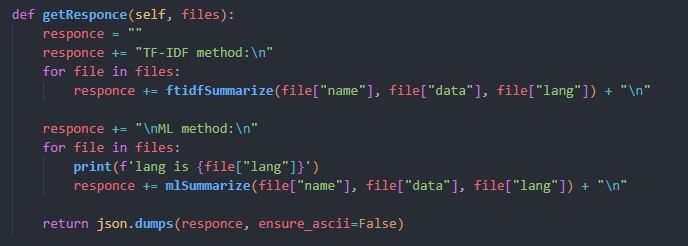


Рис.1 Запрос к серверу



Рис.2 Реализация TFIDF метода

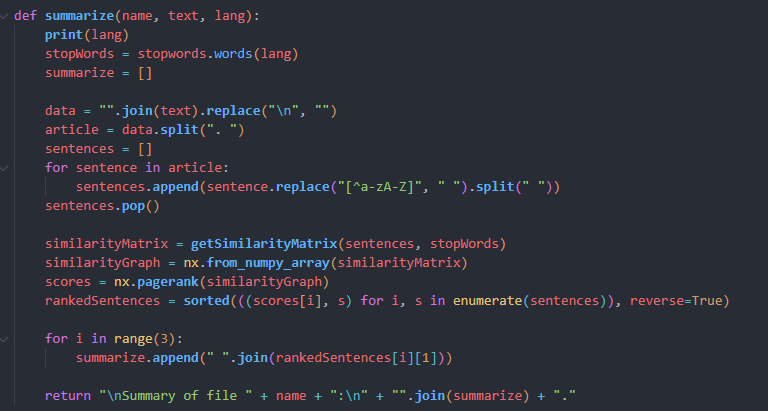


Рис.3 Реализация ML метода

**Пользовательский интерфейс:**

Приложения написано на языке TypeScript с использованием фреймворка Angular.

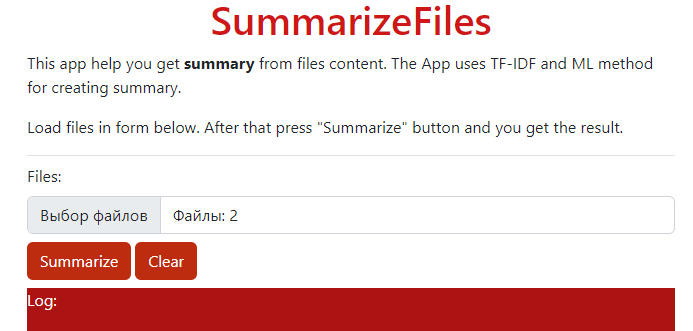


Рис. 4. Пользовательский интерфейс

Пользователь может выбрать несколько файлов для обработки.



Рис.8 Пример работы системы

**Вывод:**

Разработанная система позволяет формировать рефераты по предложенным текстам. Такие системы очень удобны для автоматической обработки текстов из множества источников: статей в интернет-газетах, публикаций в социальных сетях и так далее. Поскольку во многих источниках информация повторяется, то целесообразнее будет обработать текст для выделения важной информации. Системы автоматического абстрагирования текста сделают это за человека и дадут ему краткое эссе о самом главном.