

**Учреждение образования  
«Белорусский государственный университет  
информатики и радиоэлектроники»**

Кафедра интеллектуальных информационных технологий

## «Разработка системы автоматического реферирования документов»

Проверил: Крапивин Юрий Борисович

1

## Содержание

1. Цель работы и вариант .....	3
2. Информация о текстовой коллекции документов .....	4
3. Описание системы, данных и алгоритмов .....	5
3.1. Описание структуры системы .....	5
3.2. Описание типов данных .....	5
3.3. Описание алгоритмов .....	6
3.4. Результат тестирования системы .....	7
4. Использование библиотек .....	9
5. Вывод.....	10

## 1. Цель работы и вариант

Цель работы - освоить на практике основные принципы автоматического реферирования документов.

Вариант 4:

4	Русский, Немецкий	Sentence extraction+ ML	Научные статьи по computer science, Сочинения по литературе
---	----------------------	----------------------------	--

## **2. Информация о текстовой коллекции документов**

Были использованы тексты из газет и классических произведений на соответствующем языке. В проекте можно найти следующие файлы:

- ⑩ computerScience.txt — содержит тренировочный текст на немецком языке;
- ⑩ essayLiterature.txt — содержит тренировочный текст на русском языке;

### 3. Описание системы, данных и алгоритмов

#### 3.1. Описание структуры системы

Система представляет собой веб-приложение, написанное с использованием Flask, присутствуют классы, ответственные за интерфейс и логику за ним. Разбор текста происходит с применением метода Sentence extraction и библиотеки nltk и встроенным ML. Система реализована на языке python.

#### 3.2. Описание типов данных

В данной системе мы использовали такие типы данных как строка, массив, ассоциативный массив.

#### 3.3. Описание алгоритмов

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  $S_i$  вычисляется произведением значений функций приведенных ниже.

Функции, характеризующие положение предложения в документе  $Posd(S_i)$  и положение в абзаце  $Posp(S_i)$ :

$$Posd(S_i) = 1 - \frac{BD(S_i)}{|D|}$$

$$Posp(S_i) = 1 - \frac{BP(S_i)}{|P|}.$$

где

$|D|$  - число символов в документе D, содержащем предложение  $S_i$ ;

$BD(S_i)$  – количество символов до  $S_i$  в D( $S_i$ );

$|P|$  - количество символов в абзаце P, содержащем предложение  $S_i$ ;

$BP(S_i)$  – количество символов до  $S_i$  в абзаце.

Модифицированная TFIDF функция:

$$Score(S_i) = \sum_{t \in S_i} tf(t, S_i) \cdot w(t, D).$$

$tf(t, S_i)$  - частота термина t в предложении  $S_i$ ;

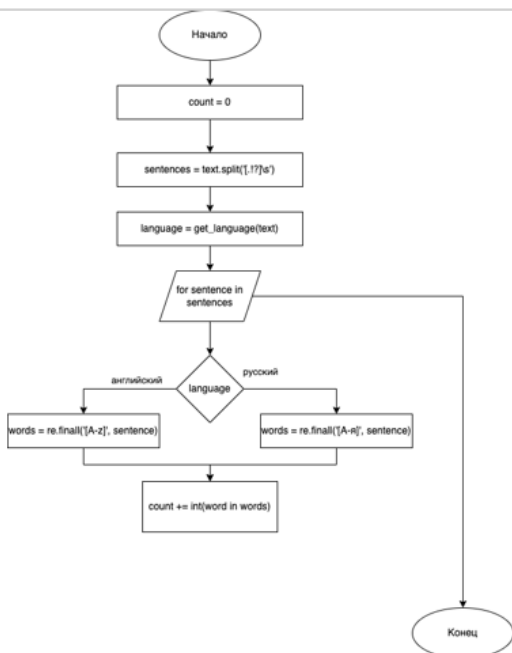
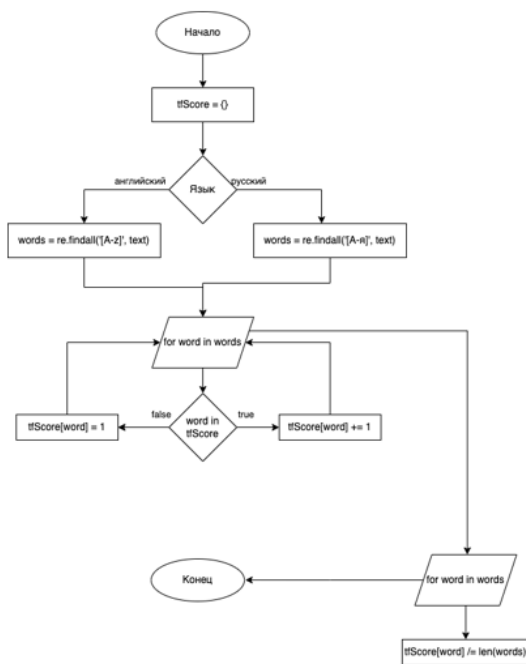
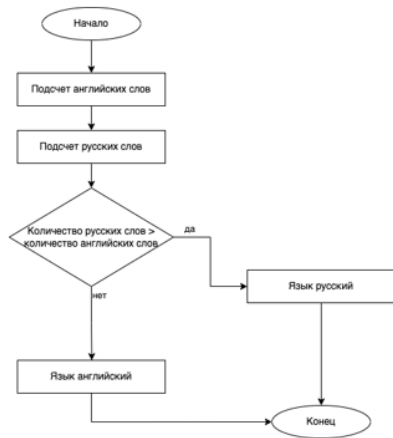
$$w(t, D) = 0.5 \left( 1 + \frac{tf(t, D)}{tf_{max}(D)} \right) \cdot \log \left( \frac{|DB|}{df(t)} \right).$$

$tf(t, D)$  - частота термина t в документе D;

$df(t)$  - количество документов, с термином t;

$tf_{max}(D)$  - максимальная частота термина в документе D;

$|DB|$  - количество документов.



### 3.4. Результат тестирования системы

Пример взаимодействия с программой выглядит следующим образом

Реферирование

Результат

Информация:

Система включает следующие функции:

- автоматическое реферирование документов (необходимо перейти на вкладку "реферирование", которая расположена в верхнем меню навигации), для этого необходимо загрузить файл и нажать кнопку "начать"

- просмотр результатов (необходимо перейти на вкладку "результат", которая расположена в верхнем меню навигации). В данном разделе можно найти всю необходимую информацию касательно реферата, а также сохранить её

Результат

Информация

Выберите файл для реферирования:

Choose File

No file chosen

начать

Реферирование

Информация

Результат:

ESSAY ON TEXT:

und sechs zu fahren von ihnen würde auf den Kopf einer Stecknadel passen! "In den vergangenen zwanzig Jahren kostete ein Rolls Royce heute weniger als 3,00 US-Dollar und erhielt 3 Millionen Meilen Liefen Sie genug Kraft, um (das Schiff) die Königin Elizabeth II.

SUMMARIZE:

haben sich so schnell verändert, dass viele Menschen mit den Veränderungen nicht Schritt halten können, entwickelt in einem ähnlichen Tempo wie Änderungen in der Computertechnologie. In einem Tempo entwickelt, das dem des Computers entsprach?

KEY WORDS:

eine zeitung versuchte zu beschreiben, dass viele menschen wissen nicht, wie die autoindustrie aussehen würde, wenn sie es hätte entwickelt, das dem des computers entsprach

MLMethod:

Computer gibt es schon seit einigen Jahren und sechs zu fahren von ihnen würde auf den Kopf einer Stecknadel passen! " Diese Änderungen sind so schnell aufgetreten, dassViele Menschen wissen nicht, wie unser moderner Computer gestartet wurde. Einige deiner Eltern waren es wahrscheinlichum 1951, als der erste Computer von einer Geschäftsfirma gekauft wurde.

Скачать исходный текст

Скачать результат

Результат:

ESSAY ON TEXT:

И, конечно, ему нужна поддержка и сострадание близких.Вокруг было столько людей, но ни один из них даже не попытался понять и помочь бедному Грегору.Это произведение оказывает ошеломляющий эффект с самого его начала. Сам этот факт должен вызывать у людей недоверие и отращение, но так как автор рассказывает об этом как о совершенно обычном явлении, читателя интересует, что же будет дальше.В эпизоде, где мы видим родителей Грегора, трудно заметить хотя бы толику их сострадания.Во-вторых, сам Кафка не выражает как такового своего отношения к происходящему.Находясь в изоляции в собственной комнате, Грегор ужасно страдает.

SUMMARIZE:

Он сожалеет, что в нынешнем его облики он сохранил способность анализировать и чувствовать все, как человек. Несмотря на отношение семьи к его недугу, он продолжает с нежностью вспоминать о них, ведь был уверен, что родные просто не могут знать, что он все понимают. Как страшно и одиноко может быть человеку, который заперт в в собственню видоизменившемся теле? Грегор понимал, что от него одни проблемы, и его семье будет куда легче, если он умрет. К Грегору больше не относились как к человеку, к члену семьи.

KEY WORDS:

близкие люди от него отворачиваются, который теряется в собственных мыслях, в которую попал наш герой, где мы видим родителей грегора, что от него одни проблемы

MLMethod:

Но чтобы могло быть, если бы хотя бы один человек, понял ситуацию, в которую попал наш герой? Еще недавно он был совершенно обычным человеком со своими проблемами, моральными ценностями мировоззрением, а теперь всего лишь насекомое в заточении К Грегору больше не относились как к человеку, к члену семьи Итак, что же мы видим? Переполненного страхом человека, который теряется в собственных мыслях.

Скачать исходный текст

Скачать результат

## **4. Использование библиотек**

Для реализации системы использовалась библиотека nltk со встроенным Machine Learning, которая позволяет работать с естественным языком (в данном случае с русским и немецким).

Для выполнения лабораторной работы использовались следующие библиотеки:

1. nltk
2. nltk.corpus
3. math
4. gensim.summarization
5. rake\_nltk
6. string
7. nltk.corpus
8. nltk.cluster.util
9. numpy
10. networkx
11. sys



## **5. Вывод**

В данной лабораторной работе была реализована система для автоматического реферирования текстов в виде веб приложения с удобным пользовательским интерфейсом, поддерживаются два языка: немецкий и русский.