**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОАПНИЯ   
РОССИЙСКОЙ ФЕДЕРАЦИИ**

**Федеральное государственное автономное образовательное учреждение**

**высшего образования**

**ДАЛЬНЕВОСТОЧНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

**Кафедра прикладной математики, механики, управления и программного обеспечения**

Отчет по лабораторной работе

по дисциплине «Компьютерная лингвистика»

по образовательной программе подготовки бакалавров по направлению 02.03.03 "Математическое обеспечение и администрирование информационных систем"

Выполнили студенты гр. Б8404

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/ Р.Д. Рындыч

(подпись)

Доцент кафедры ПММУиПО,

к.ф.-м.н., доцент \_\_\_\_\_\_\_/Л.И. Прудникова

(подпись)

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2019 г.

г. Владивосток

2019 г.

Оглавление

[Введение 3](#_Toc6487327)

[Квазиреферирование 4](#_Toc6487328)

[Статистические методы 5](#_Toc6487329)

[Структура программы автореферирования 5](#_Toc6487330)

[StringSlicing 5](#_Toc6487331)

[WordStorage 5](#_Toc6487332)

[Abstractor 6](#_Toc6487333)

[Mystem 6](#_Toc6487334)

[Командная строка 6](#_Toc6487335)

[Альтернативный вариант 8](#_Toc6487336)

[Интерфейс 9](#_Toc6487337)

# Введение

Искусство реферирования как извлечения наиболее существенной информации из исходно- го текста и ее последующей презентации во вто- ричном (реферативном) тексте стало неотъемлемой частью современной жизни. Тем не менее реферирование имеет многовековую историю, уходящую корнями во времена Шумерской цивилизации, когда писцы создавали краткие записи на глиняных табличках. Далее, в эллинистический период и в эпоху распада Римской империи на пергаменте и свитках также делались краткие тезисы и описания документов. Тогда (предположительно) и возник термин abstractus. Первые реферативные журналы (РЖ) в Европе появились в XVII–XVIII вв., т.е. в эпоху, которая известна как «Эпоха открытий». Такие журналы позволяли своевременно информировать ученых о научных достижениях и наиболее значимых публикациях. Первыми журналами, публикующими рефераты, были «Philosophical Trans- actions», издающийся с 1665 г. Лондонским королевским обществом, итальянский журнал «Giornali dei letterati» (1668–1680), немецкий журнал «Acta eruditorum» (1682). В XIX в. стали выходить в свет журналы, которые могут быть отнесены уже к собственно РЖ: «Pharmazeutisches Zentralblatt», переименованный впоследствии в «ChemischesZentralblatt», «Fachzeitschrift für die gesarnte Papier, Pappen und Papierstoffindustrie», «Stahl und Eisen», «Science Abstracts».

Задачи автоматического анализа и синтеза неоднофразовых текстов связаны с решением следующих проблем:

1. Предложения текста могут содержать элементы, значение которых определяется элементами других предложений.
2. Для определения общего смысла текста существенен не только смысл отдельных предложений, но и отношения между ними.
3. Текст имеет такие неотъемлемые характеристики, как тема и композиционная структура, что существенно при определении смысла не только отдельных фрагментов, но и всего текста в целом.

В этой работе использовался метод квазиреферирования.

# Квазиреферирование

Квазиреферирование сводится к извлечению из документов некоторых минимальных фрагментов текста, в максимальной степени связанных с основными положениями документа. Создание квазиреферата, таким образом, представляет собой соединение выбранных фрагментов.

Выделяют три главных направления, применяемые для квазиреферирования:

-индикаторные методы, основанные на оценке элементов текста исходя из наличия в них специальных слов и словосочетаний – так называемых маркеров важности («в заключение», «было отмечено, что...» и пр.), характеризующих их смысловую значимость;

-позиционные методы, опирающиеся на предположение о том, что информативность элемента текста находится в зависимости от его позиции в документе;

-статистические методы, основанные на оценке информативности различных фрагментов текста по частотным показателям, таким, как расположение этого блока в оригинале, частота появления терминов в тексте, частота использования терминов в ключевых предложениях и т.д.

# Статистические методы

В данной работе применены статистические методы

Определение веса фрагментов исходного текста в статистических методах выполняется по алгоритмам, разработанным еще в 1960–70-е годы и ставшим уже традиционными. Общий вес текстового блока на этом этапе вычисляется по формуле:

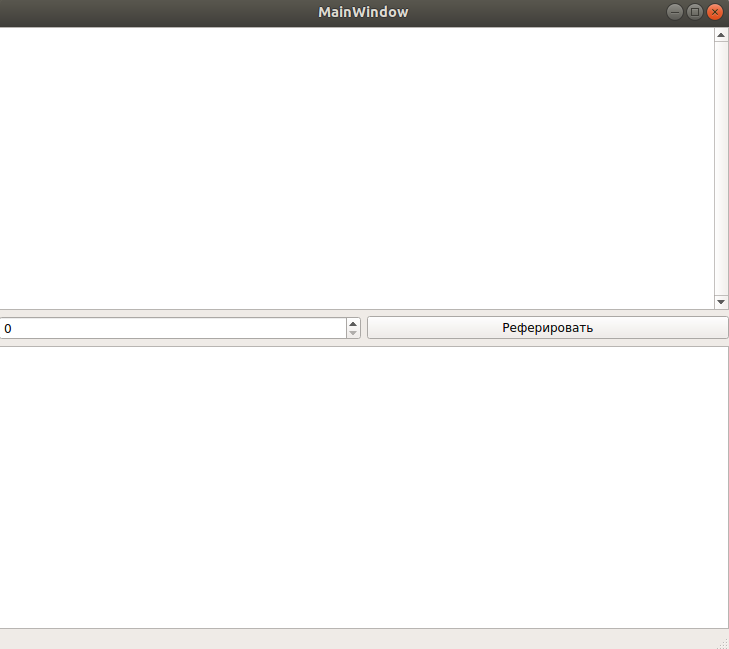
W= L+ K+ S,

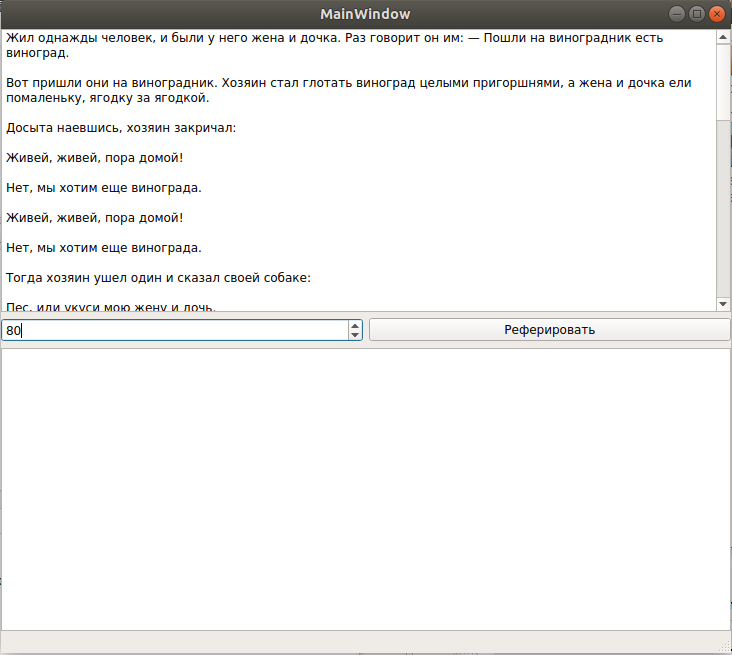
где L определяется расположением блока в исходном тексте и зависит от того, где появляется данный фрагмент — в начале, в середине или в конце, а также используется ли он в ключевых разделах текста, например, в заключении К учитывает использование блока в «резюмирующих» конструкциях типа «в заключение», «в данной статье», «результатом является» и т.п.. Статистический вес текстового блока S вычисляется как нормированная по длине этого блока сумма весов входящих в него терминов – слов и словосочетаний. Блоки с наивысшими весовыми коэффициентами и будут включены в текст квазиреферата.

# Интерфейс

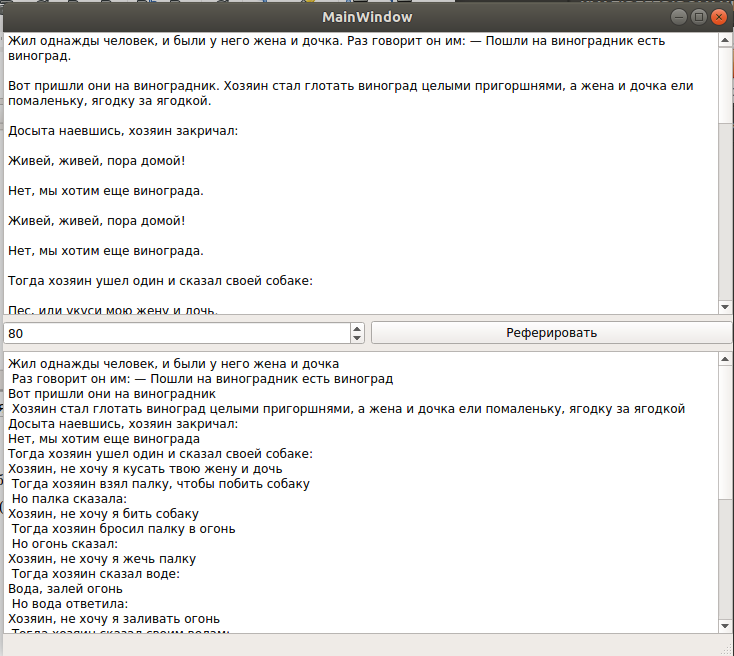
При запуске программы пользователю открывается окно ввода текста и коэфициента сжатия. (рис.1)

Пользователю предлагается ввести текст и степень сжатия в диапазоне от 0 до 99.(Рис 2)

  
Рисунок 1: Главное окно

  
Рисунок 2: Окно с заполненными полями

После нажатия кнопки «Реферировать» будет выполнена программа, результат работы которой будет отображен в нижнем большом окне(Рис 3)

  
Рисунок 3: Результат работы программы