

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

**«Дальневосточный федеральный университет»**

(ДВФУ)

**ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ**

**(ШКОЛА)**

**Департамент программной инженерии и искусственного интеллекта**

Зеленин Кирилл Вадимович

РАЗРАБОТКА ИНСТРУМЕНТАЛЬНОЙ СИСТЕМЫ ДЛЯ РАБОТЫ ПОЛИТОЛОГОВ: МОДУЛЬ «АНАЛИЗ ВНЕШНЕПОЛИТИЧЕСКИХ  
ДОКУМЕНТОВ».

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

по образовательной программе подготовки бакалавров по направлению

09.03.04 «Программная инженерия», профиль «Программная инженерия»

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |  |  | Студент группы Б9121-09.03.04 | | | | | | | | |
|  |  |  | | | | Зеленин К.В. | | | | |
|  | | | | | | |  |  | (подпись) | | | |  | | | | |
|  | | | | | | | | | | | | | | | | | |  |  | Руководитель |
|  | | | | | | |  |  | ассистент ДПИиИИ | | | | | | | | |
|  | | | | | | |  |  | ученая степень, должность | | | | | | | | |
|  | | | | | | |  |  |  | | | | |  | | Чусова А.Е. | |
|  | | | | | | |  |  | (подпись) | | | | |  | | (ФИО) | |
|  | | | | | | | | | | | | | | | | | |  |  | Защищен с оценкой |
| Регистрационный № | | | | | | |  |  |  | | | | | | | | |
|  | | | | | | |  |  |  | | | | | | Зеленин К.В. | | |
|  | | | |  |  | |  |  | « |  | » |  |  | | | | 2025 г. |
| (подпись) | | | |  | (ФИО) | |  |  |  | | | | | | | | |
| « |  | » |  | | | 2025 г. |  |  |  | | | | | | | | |

г. Владивосток

2025

Оглавление

[Введение 3](#_Toc189396807)

[1.1 Общее аспекты анализа документов в политологии 4](#_Toc189396808)

[1.2 Методы анализа документов 5](#_Toc189396809)

[1.3 Анализ существующих средств 10](#_Toc189396810)

[Требования к программному средству 15](#_Toc189396811)

[2. Формальная постановка 17](#_Toc189396812)

[2.1 Предварительная обработка текста 17](#_Toc189396813)

[2.2 Поиск вхождений ключевых слов 19](#_Toc189396814)

[2.3 Анализ контекста вхождений 21](#_Toc189396815)

[2.4 Формирование и использование словарей 23](#_Toc189396816)

[2.5 Графические представление результата анализа 26](#_Toc189396817)

[3 Проектирование системы 29](#_Toc189396818)

[3.1 Требования к системе 29](#_Toc189396819)

[3.1.1 Диаграммы прецедентов 30](#_Toc189396820)

[3.1.2 Функции системы 33](#_Toc189396821)

[3.2 Архитектура системы 44](#_Toc189396822)

[Список использованных источников 49](#_Toc189396823)

# Введение

Политика - это сфера деятельности, которая занимает центральное место в функционировании любого государства. Политологи же занимаются аналитикой текущих и будущих действий государств, анализирует их взаимодействия между собой, и прогнозируют будущее политические события и развития международных отношений. Исследование политических процессов, принятие решений и анализ политических документов являются важными аспектами политологических исследований. Именно анализ внешнеполитических документов и выступлений представителей государств позволяет политологом замечать тренды в изменении политики государства, а также делать прогнозы дальнейших действий государства.

Внешнеполитический документ - документ, который публикуется государственными институтами и содержит информацию о текущем и будущем направлении внешней политики какого-либо государства.

Но зачастую внешнеполитические документы и выступления крайне объёмны из-за чего их анализ достаточно затруднителен. С целью облегчения анализа внешнеполитических документов и выступлений имеет смысл использовать современные методы анализа текстов.   
Целью курсовой работы является: Сравнение и выбор самого оптимально метода анализа текста, поиск и анализ существующих средств, доказательство актуальности разработки программного обеспечения работы с документами для политологов.

# Общее аспекты анализа документов в политологии

Политология как наука часто занимается анализом трендов и акцентов внешней политики каких-либо государств для того, чтобы проще прогнозировать будущее действия государства. Лучшим способом оценки ценностей, целей и акцентов государства является анализ внешнеполитических документов и выступлений представителей государства, при этом в большинстве своём они объёмны и содержат много ненужной информации. Потому необходимо проводить тщательный анализ документа или выступления на предмет упоминаний каких-либо тем. Имеет смысл проводить анализ и последующие сравнение нескольких выступлений или документов, опубликованных с каким-либо промежутком времени. Такой анализ покажет на сколько какая-то тема стала более или менее важной для государства через время.

Таким образом такой анализ помог бы политологу сделать свои прогнозы и аналитику более обоснованной, так как они будут основываться на конкретных данных исходящих от государственных лиц.  
 Потому требуется провести сравнение методов анализа текстов, и выбрать такой, что сможет проводить как количественный и качественный анализ по конкретному термину в нескольких текстах.

# 1.2 Методы анализа документов

На сегодняшний день в политологии распространены следующие методы анализа текстов:

1)Традиционный анализ документа

Традиционный анализ документа концентрируется на смысловой интерпретации источника конкретным исследователем со своими взглядами и мировоззрением.

Делится на:

•Анализ контекста документа. Здесь оцениваются особенности политической ситуации, являющейся «средой» возникновения документа, интересы, цели, взгляды продуцентов документа. Важно учесть формального и фактического адресата документа, которые могут не совпадать. Например, в лоббистских кампаниях статья, опубликованная в СМИ и формально рассчитанная на самую широкую аудиторию, на деле может предназначаться узкому кругу лиц во властных структурах, принимающих решение по конкретному прагматическому вопросу.

• Внутренний содержательный анализ документа, предполагающий интерпретацию его смысла, логико-семантический анализ, анализ структуры, композиции, явно выраженных или скрытых акцентов. Также следует обращать внимание на те политические фигуры или организации, упоминание которых в тексте выполняет определенную функцию (например, ссылка на мнение председателя правительства в качестве аргумента, подтверждающего позицию автора). Таким образом выявляется и круг референтных для автора политических субъектов, что позволяет более глубоко понять его политическую позицию.[1]

2) Количественные методы:

а) Интент-анализ

Этот метод, заимствованный из области психологии, исключительно подходит для анализа конфликтных и кризисных политических ситуаций, также его можно использовать для интерпретации позиций нескольких участников политической дискуссии. Различают несколько вариантов его выполнения: денотатный, предикатный, комплексный и т. д. Суть метода заключается в определении интенций, так как в речи, особенно спонтанной, специально не подготовленной (а если и подготовленной, то прерывающейся спонтанными фразами), находят отражение мысли, восприятие, знания, эмоции, оценки, отношение политика к людям и событиям. С помощью интент-анализа исследователи делают попытку оценить, как в политической речи проявляются «внутренние» — личностные и когнитивные — состояния его автора.[3]

Метод состоит из последовательных этапов: выделение круга обсуждаемых тем и вопросов, определение связей между объектами, затем кодификации дескрипторов, далее проводится оценка групп объектов по нескольким интегральным измерениям, полученные значения усредняются и определяются интегральные значения каждого объекта по указанным параметрам.[2]

б) Контент-анализ

Самый распространенный метод, имеющий множество вариаций в различных методиках, позволяющий провести качественно-количественный анализ содержания текстовых массивов с целью последующей интерпретации выявленных числовых закономерностей. Заключается в оценке частотного распределения слов, словосочетаний словоформ и других единиц анализа (число их вариаций теоретически безгранично) относительно текста. Результатом является частота, относительный и удельный вес, вероятность встречаемости и пр., на основе чего делается качественный или количественный вывод в зависимости от выдвинутой гипотезы.[2]

3) Качественные методы:

а) Дискурс-анализ

Совокупность методик и техник интерпретации текстов или высказываний как продуктов речевой деятельности, осуществляемой в конкретных общественно-политических обстоятельствах и культурно-исторических условиях. Этот метод ориентирован, прежде всего, на изучение лингвистического уровня в структуре социальной коммуникации как доминирующего на протяжении определенного исторического периода развития общества и культуры. Сам метод заключается в последовательности ряда операций: фиксация изучаемого материала; выделение его формальных характеристик; обозначение контекста как коммуникативной ситуации; выбор направления и стратегии анализа; теоретическое дифференцирование и структурирование этапов исследования; определение техники и средств анализа при использовании конкретной модели исследования; дефиниции единиц анализа; проверка системы категорий в теории и на эмпирическом материале; осуществление основных этапов исследования; фиксация результатов исследования, их обобщение, истолкование и структурирование. Дискурс-анализ позволяет выделить не только существенные характеристики социальной коммуникации, но и второстепенные, содержательные и формальные показатели. [2]

б) Когнитивное картирование

Один из качественных методов анализа текстовых данных, который позволяет измерить психологические параметры политических лидеров или других политических субъектов. Данный метод был разработан в рамках когнитивной психологии для поиска в структуре мышления отдельного человека устойчивых конструкций познания и понимания окружающей реальности. Операционный код традиционно определяется как система когнитивных конструкций, которая используется политическим актором при анализе им конкретных политических ситуаций относительно постоянно, что и делает возможным ее выявление.[2]

Когнитивное картирование позволяет выявить казуальную, причинную структуру политического текста, что используется для анализа кризисных политических ситуаций, а также для моделирования мышления политика. Фактически когнитивное картирование позволяет определить факторы, которые учитывают политики при принятии решений. Стратегия выбора факторов и задает возможную типологию политического мышления.[4]

С формальной стороны когнитивная карта представляет собой граф, в узлах которого находятся описания некоторых важных событий, си­туаций, а отношения, связывающие узлы (стрелки или дуги), отражают причинные связи между событиями, влияние событий друг на друга.[2]

Рассмотрим представленные методы и найдём наиболее подходящий для решения поставленной задачи.

Традиционный метод анализа слишком субъективен и трудозатратен для того использовать при объективном анализе больших объёмов текстов. Также он в основном опирается на проводящего этот анализ исследователя из-за он не подходит для решения поставленной перед нами задачи.

Интент-анализ в основном фокусируется на определения цели и темы, которую автор текста доносит в тексте, а нашем случае такой метод не подходит, так как для нашей задачи требуются метод, который бы анализировал именно, что объективной содержание текста.

Когнитивное картирование же не почти не как не опирается на количественный анализ, в основном фокусируясь на психологическом анализе текста и автора, из-за чего этот метод также не подходит для решения данной задачи, потому что в ней необходим именно количественный и качественный анализ текста.

Контент-анализ же является методом, который объединяет как количественный, так и качественный подходы к анализу текста. Он позволяет рассматривать тексты с различных точек зрения, анализировать их содержание, структуру, стилевые особенности, а также учитывать контекст и цели, которые автор преследовал при написании текста. Такой комплексный подход позволяет получить более полное представление о текстовой информации и принять более обоснованные решения на основе проведенного анализа. Именно поэтому в данной работе отдаётся предпочтение данному методу.

# 1.3 Анализ существующих средств

1. “Ваал-мини”

Система ВААЛ, позволяет прогнозировать эффект неосознаваемого воздействия текстов на массовую аудиторию, анализировать тексты с точки зрения такого воздействия, составлять тексты с заданным вектором воздействия, выявлять личностно-психологические качества авторов текста и делать многое другое.

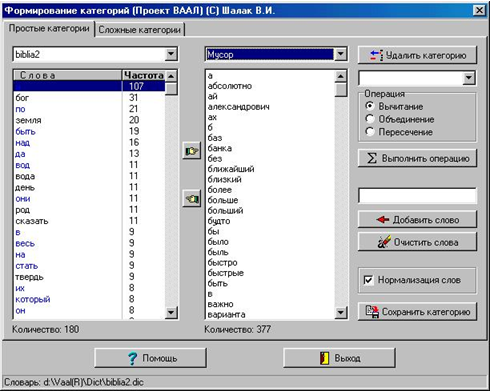
Программа обладает широким функционалом в том числе и для контент-анализа, способна составлять словарь текста, выделять нужные и ненужные слова, и подсчитывать их количество в тексте. Система ВААЛ включает в себя ряд модулей, позволяющих проводить полноценный контент-анализ текстов. Некоторые из этих модулей уникальны, так как реализуют методы, не представленные ни в одной из коммерческих систем для компьютерного контент-анализа.[5]

Рисунок 1 – Пример работы программы Ваал-мини

1. **“Concordance”**

Программа с помощью, которой можно подсчитать количество слов и частотность для списка слов. Данный инструмент широко применяется при анализе текстовых данных, в науках, связанных с языками, лингвистике, для извлечения информации, лексикографии, переводе, а также в ряде других коммерческих областей и научных дисциплин. Позволяет выбрать вариант сортировки слов: поиск фраз, учет расстановки слов при поиске, а также использовать регулярные выражения для поиска. Просмотреть статистику по тексту можно по типу слов, процентному соотношению, количеству символов и предложений.[6]

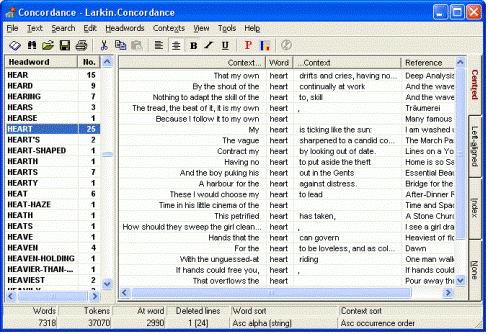


Рисунок 2 – Пример работы программы Concordance

1. “Simple Word Sorter”

Программа позволяет провести экспресс контент-анализ документа, идентифицировав наиболее часто встречающиеся слова, тем самым, определив, какие темы и микротемы доминируют в пространстве анализируемого текста. Не обладает широким функционалом, и потому понятна и проста в обращение для неопытного пользователя.[7]

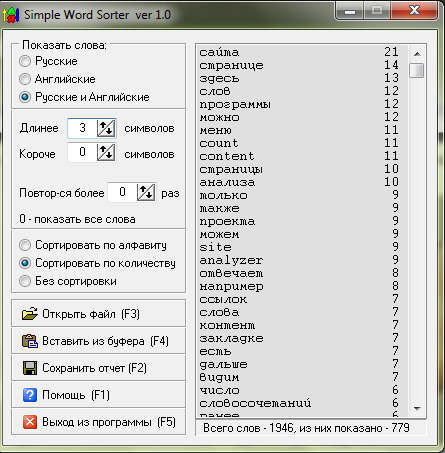


Рисунок 3 – Пример работы программы Simple Word Sorter

1. “Лекта”   
   Программа позволяет интерпретировать большие объемы текстовой информации, например, полную совокупность газетных статей за месяц или за год. Программа может выявить значимые темы, которые присутствуют в анализируемом массиве информации. Программа также реализует метод контент-анализа, который позволяет определить частоту встречаемости в текстах слов, словосочетаний или сюжетов.

Учитывает как отдельные слова, так и словосочетания. Рассчитывает частоту встречаемости выделенных семантических единиц в тексте. Позволяет дробить текст на однородные смысловые фрагменты. Итоговый результат: матрица факторных коэффициентов и факторных нагрузок, характеризующая содержание выделенных сюжетных линий в совокупности анализируемых текстов и матрицу, содержащую иерархию фрагментов текстов по степени их соответствия каждому их выделенных сюжетных линий.[8]

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

Рисунок 4 – Пример работы программы “Лекта”

Проведём сравнение представленных средств и выделим плюсы и минусы каждого, а также соответствие требованиям поставленной задачи.

Таблица 1 – Сравнение существующих средств

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Ваал-мини | Concordence | Simple Word Sorter | Лекта |
| Поддержка английского языка | + | + | + | - |
| Количественный анализ вхождения слова в текст | + | + | + | + |
| Демонстрация контекста слова | - | + | - | - |
| Эмоциональная оценка контекста | -  но анализирует эмоциональный окрас всего текста | - | - | - |
| Представления анализа в форме графика | - | - | - | + |
| Анализ сразу нескольких документов | + | - | - | + |

# Требования к программному средству

На основе поставленных в предыдущей главе задач можно сформулировать основные функции, которым разрабатываемое программное средство должно соответствовать, а именно:

1. Выделение основы слов текста;
2. Поиск всех вхождений слова, по которому введётся анализ, в тексте;
3. Анализ контекста найденных вхождений слова(то есть анализ тех слов, которые напрямую относятся к вхождению);
4. Формирование словарей, которые будут использованы для определения эмоционального окраса слова;
5. Графическое представление результата анализа нескольких тестов в виде таблиц и графиков.

Общий же функционал программного средства заключается в анализе текста по ключевому слову, при помощи словарей, где ключевое слово K - это слово, по которому проводиться анализ в тексте, а именно осуществляется поиск всех вхождений это слова, после чего происходит анализ контекста этих вхождений, словарём же называется множество пар значений <W,E>, где W – это слово, а E – число в диапазоне [-2,2], чем меньше число, тем негативнее эмоциональный окрас слова, и чем больше число, тем более позитивный окрас. Словари составляются экспертами предметной области – политологами.

Программное средство должно иметь следующий принцип работы, сначала слова текста и словаря приводятся к нормальной форме, то есть оставляется только основа слова, текст разбивается на массив предложений, а предложения разбиваются на массив слов. После происходит поиск всех вхождений заданного ключевого слова в текст(ключевое слово также приводится к нормальной форме), далее анализируется соседние к найденным вхождения слова, это может быть от двух до пяти слов влево и вправо от вхождения, далее будем называть этот набор слов контекстом вхождения. Слова из контекста вхождений будут сравниваться со словами, находящимися в словаре, и если будет обнаружены совпадения, то найденному вхождению будет присвоено определённое число, которое будет обозначать эмоциональную окраску контекста вхождения ключевого слова. Далее можно повторить данный процесс для нескольких текстов, и после получить сравнительный анализ, полученных на основе анализа данных, в виде графиков и диаграмм.

# 2. Формальная постановка

## 2.1 Предварительная обработка текста

Перед началом анализа текста на предмет вхождения в него определённых слов, необходимо провести предварительную обработку этого текста, а именно привести слова к нижнему регистру и нормальным(начальным) формам, то есть провести стемминг текста. Стемминг (нормализация) – это процесс нахождения основы и корня для заданного исходного слов[9]. Для нормализации слова будет использоваться алгоритм Snowball Stemmer из библиотеки NLTK [10].Данный алгоритм основывается на правилах, по которым слова переводятся из данной формы, в нормализированную форму. Для английского языка используется алгоритм Мартина Портера[11]. Для русского языка, так же используется модифицированный алгоритм Портера.

Но перед работой со словами, необходимо выделить их из текста, для этого текст будет разбит на массив предложений, а предложения будут разбиты на массив слов. Для этого также будет использована библиотека NLTK, а именно токенизатор предложений. Для токенизации предложений будет использоваться инструмент sent\_tokenize из библиотеки NLTK [12], а для токенизации слов будет использоваться инструмент word\_tokenize из той же библиотеки [13].Алгоритмы обоих используемых инструментов токенизации, основаны на использовании регулярных выражений.

По итогу данный функциональный компонент принимает на вход текст и словарь, а возвращает массив нормализованных слов из текста и массив нормализованных слов из словаря. Текст разбивается на предложения, предложения разбиваются на слова, а далее все слова текста и словаря на основе корпуса языка и правил преобразования нормализуются.

Изображение выглядит как текст, снимок экрана, Шрифт, графический дизайн

Автоматически созданное описание

Рисунок 5 – Блок-схема алгоритма нормализации слов текста

## 2.2 Поиск вхождений ключевых слов

Данный функциональный компонент ищет все совпадения ключевого слова в массиве, полученном в предыдущем функциональном модуле, а именно массив предложений текста вида T(, где это массив слов предложения вида , где – это последовательности символов русского и английского языка в диапазоне , значения n и m это количество предложений в тексте, и слов в i-ом предложении соответственно. Данный функциональный компонент на выходе возвращает все пары индексов вида [i, j], где i – это индекс предложения, который находится в диапазоне [1, … ,n], а j – это индекс слова в j-ом предложении, который находится в диапазоне [1, … ,m], где слово с индексами = K(ключевому слову).

Необходимо учесть, что ключевым словом при анализе политических документов зачастую является название темы, термина или наименование страны, и зачастую слова из выше перечисленных групп слов имеют множество синонимов, который в контексте политического документа будут обозначать одно и тоже (пример: Россия, Российская Федерация, РФ), потому K будет не одним словом, а массивом слов K, где это слова, являющиеся при этом синонимами, которые определяются пользователем, а индекс p – количество слов в массиве ключевых слов K. И следовательно слова текста будут сравниваться с каждым словом из массива K, и совпадением будет считаться случай, когда = , где K, а t – это индекс слова в массиве K. Все слова массива K должны быть разными, так как наличие двух одинаковых слов не даст каких-либо дополнительных данных при анализе, а только замедлит его процесс. При обнаружении совпадения слова из текста с словом из массива ключевых слов K, индекс t не учитывается при сохранении пары значений [i, j], так как слова массива K одинаковы по смыслу, и потому дополнительного уточнения с каким словом произошло совпадение не требуется.

## 2.3 Анализ контекста вхождений

Данный функциональный компонент определяет эмоциональный контекст вхождения слова, принимая на вход массив предложений и слов в них, а также все пары индексов, которые определяют места вхождений ключевого слова, на выходе же возвращается число, которое присваивается каждому вхождению, и которое обозначает эмоциональную оценку контекста вхождения. Эмоционально-оценочная лексика(слова) — это выразительная лексика, выражающая отношение пишущего или произносящего к чему-либо, как положительное, так и отрицательное, поэтому данную группу лексических единиц можно разделить на два больших разряда: слова с положительной характеристикой и слова с отрицательной (негативной) характеристикой.[14] Эмоциональная оценка контекста определяется на основе соседствующих к вхождению слов, в зависимости от заданных параметров, это может быть от двух до пяти слов влево и вправо по тексту от вхождения, данное число анализируемых слева и справа слов называется радиусом анализа и определяется пользователем.

Каждое слово, которое находится в контексте вхождения сравнивается со словами из словаря, в случае совпадения, контексту вхождения присваивается число соответствующее значению E этого слова, если в контексте вхождения будет обнаружено несколько совпадений, то найденному вхождению будет присвоено самое наименьшее из чисел, если все совпавшие из словаря слова имеют значение E < 0, и самое большое, если совпавшие слова имеют значение E > 0, в случае же, когда в контексте вхождения есть совпавшие слова, как с отрицательными, так и с положительными значениями, данному контексту вхождения будут присвоено значение 0, если же в контексте вхождения не будет не одного совпадения со словами из словаря, вхождению будет присвоен особое значение, которое будет означать отсутствие эмоциональной окраски у данного вхождения. Данный подход во многом схож с методом основанном на использовании банка позитивных и негативных слов, за исключением, что банк слов для анализа политических текстов, будет составляться экспертами-политологами[15].

Так как текст преобразован в массивы слов, то в случай анализа 3 соседних слов от вхождения это будут следующие слова(если слова с данными индексами будут существовать) , *,* ,, каждое из этих слов будет сравниваться со всеми словами из словаря, и если на пример слово , где W – это слово из словаря, то паре i, j будет присвоено дополнительное значение E, соответствующего слова из словаря.

Стоит заметить, что анализ контекста вхождения происходит только внутри одного предложения, если радиус анализа захватывает больше слов, чем есть в предложении, то анализ контекста вхождения проводиться только по тем словам, которые могут быть взяты из предложения.

## 2.4 Формирование и использование словарей

Как было описано ранее словарём называется массив значений <W,E>, где W – это слово, а E – число в диапазоне [-2,2], где слово с E = -2 будет считаться очень негативным, с E = -1 просто негативным, слово с E = 1 будет определять как позитивное, а слово с E = 2 же будет считаться как очень позитивное. Что же не было описано ранее так это факт того, что E не может быть равным нулю, так как предполагается, что число 0 в данном диапазоне будет обозначать слово с нейтральной эмоциональной окраской или без неё вовсе, и ради облегчения процесса формирования словаря, в нём находятся только слова имеющие какую-либо эмоциональную окраску отличную от нейтральной, слова же которых нет в словаре по умолчанию считаются как слова, имеющие значение эмоциональной окраски 0, такие слова никак не будут учитываться при анализе контекста вхождения слова, и следовательно добавление таких слов в словарь не требуется.

Но даже после составления словаря, в котором будут иметься все слова с эмоциональной окраской отличной от нейтральной, возникает другая проблема, разные политологи могут по-разному оценивать эмоциональную окраску каких-то конкретных слов, и так как учесть мнение каждого политолога, по эмоциональной оценке, какого-либо слова невозможно, было принято решение сделать словари модифицируемыми, сохраняемыми и переносимыми. В таком случае, если какой-то политолог не согласиться с эмоциональными оценками слов в стандартном словаре, который будет представлен программным средством по умолчанию, то он сможет изменить эмоциональные оценки каких-то слов или добавить новые слова, после редактирования политолог сможет сохранить новый словарь и пользоваться им для анализа текстов, имея при этом возможность отправить свои новый словарь другим людям, чтобы уже они могли им пользоваться или отредактировать.

В качестве же основы для стандартного словаря можно взять эмоциональные оценки английских слов из сентимент анализа библиотеки NLTK[16] и словаря AFINN(Affective Norms for English Words)[17],

Для более наглядного объяснения работы словарей представим показательный пример словаря, составленного неким экспертом-политологом:

Т а б л и ц а 2 – Пример исходного словаря

|  |  |
| --- | --- |
| W(слово словаря) | E(эмоциональная оценка слова) |
| справедливые | 2 |
| разумное | 1 |
| неоправданные | -1 |
| агрессия | -2 |
| агрессивность | -1 |

Данный пример показывает словарь, который ещё не был обработан функциональным модулем нормализации слов и потому все слова пока находятся в тех формах, в которых они были записаны экспертом-политологом.

После же начала работы с текстом, что сам анализируемый текст, что словарь нормализуются и приводятся к нижнему регистру, после всех изменений исходный словарь будет иметь уже следующий вид.

Т а б л и ц а 3 – Пример словаря обработанного ПО перед началом анализа текста

|  |  |
| --- | --- |
| W(слово словаря) | E(эмоциональная оценка слова) |
| справедливость | 2 |
| разум | 1 |
| неоправданность | -1 |
| агрессия | -2 |
| агрессия | -1 |

Как видно из примера, в ходе нормализации словаря могут возникнуть ситуации, когда из-за ошибки политолога в словаре оказалось два похожих однокоренных слова, которые после стемминга станут двумя одинаковыми словами, в таком случае при анализе текста будет учитываться только самое вышестоящие по списку слово, все повторы данного слова и их эмоциональные оценки будут игнорироваться.

## 2.5 Графические представление результата анализа

После анализа текста и получения информации обо всех вхождениях ключевого слова, а также эмоциональной окраски контекста этих вхождений необходимо представить полученные данные в удобном для анализа виде. По итогу анализа текста был получен массив значений [ i, j, E], где значения i и j определяют место вхождения, а E эмоциональный окрас контекста этого вхождения. Единственное из этих трёх значение, которое будет важно на этапе графического отображения результата это эмоциональная окраска вхождения, также на данном этапе подсчитывается общее количество вхождений и то сколько вхождений имело какую-то определённую оценку. При анализе одного документа эти данные лучше всего вывести в виде таблицы следующего вида:

|  |  |
| --- | --- |
| Общее количество упоминаний | 38 |
| Количество очень позитивных упоминаний | 0 |
| Количество позитивных упоминаний | 3 |
| Количество негативных упоминаний | 13 |
| Количество очень негативных упоминаний | 4 |

Т а б л и ц а 4 – Пример отображения полученных после анализа данных

Такая таблица показывает сколько было найдено вхождений ключевого слова, сколько из этих вхождений имеют E = 2, E = 1, E = -1, E = -2 соответственно.

Нужно учитывать, что общее количество вхождений не обязательно равно сумме всех вхождений с различными эмоциональными окрасками, это связано с тем, что не у каждого упоминания можно определить эмоциональную окраску.

Если же пользователь хочет сравнить результаты анализа нескольких документов, то такие данные удобнее отобразить в виде графика, но из-за большого количества различных показателей будет разумнее сделать несколько графиков, показывающих различные аспекты анализа, а именно:

1. График количества вхождений.

Данный график строиться исключительно на основе общего количества упоминаний по тексту, в качестве оси Y берётся количество вхождений ключевого слова, а в качестве оси X берутся равноудалённые точки, которые подписаны, как соответствующие отображаемому на оси Y количеству вхождений, документам. Такой график будет показывать, насколько советующий документ акцентируется на ключевом слове, и как данный акцент менялся из документа в документ.

1. График изменения суммы эмоционального отношения

Данный график строиться на основе суммы эмоционального значения E всех упоминаний ключевого слова в тексте, в качестве оси Y сумма всех E вхождений в документе, а в качестве оси X берутся равноудалённые точки, которые подписаны, как соответствующие отображаемому на оси Y суммам эмоциональных значений вхождений, такой график будет показывать суммарное отношение к ключевому слову по документу, и будет показывать как менялось это отношение из документа в документ.

1. Диаграмма вхождений

Данная диаграмма показывает количество вхождений с эмоциональными оценками всех видов, такое отображение показывает различия в распределение упоминаний по эмоциональным оценкам между документами.

# 3 Проектирование системы

В данной главе рассматривается архитектура модуля инструментальной систем для работы политологов для анализа внешнеполитических документов. Описывается требования к системе, её функционал, классы, пользователи, ограничения и их взаимодействия.

## 3.1 Требования к системе

Требования к программной системе можно разделить на 2 типа функциональные, нефункциональные.

Функциональные требования (functional requirements) определяют функциональность ПО, которую разработчики должны построить, чтобы пользователи смогли выполнить свои задачи в рамках бизнес-требований [18].

Таблица 5 – Перечень функциональных требований разрабатываемой системы

|  |  |
| --- | --- |
| **№ п/п** | **Наименование функционального требования** |
| 1 | Анализ одного или нескольких документов на предмет вхождения необходимого термина |
| 2 | Анализ эмоционального контекста вхождения необходимого термина на основе словаря эмоциональных окрасов слов |
| 3 | Возможность создания и сохранения отличных от базового словарей эмоциональных окрасов слов |
| 4 | Формирования отчёта на осное результатов анализа нескольких документов в виде представительных таблиц и графиков |

В нефункциональных требованиях описаны цели и атрибуты качества. Атрибуты качества представляют собой дополнительное описание функций продукта, выраженное через описание его характеристик, важных для пользователей или разработчиков. К таким характеристикам относятся легкость и простота использования, легкость перемещения, целостность, эффективность и устойчивость к сбоям. Другие нефункциональные требования описывают внешние взаимодействия между системой и внешним миром, а также ограничения дизайна и реализации[18].

Таблица 6 – Перечень нефункциональных требований разрабатываемой системы

|  |  |
| --- | --- |
| **№ п/п** | **Наименование нефункционального требования** |
| 1 | Возможность формировать группы документов при анализе, а также выводить результаты анализа отдельно по группе. |
| 2 | Доступный и понятный интерфейс формирования словаря эмоциональных окрасов слов |
| 3 | Загрузка документов и словарей простым и очевидным способом, например drag-and-drop |
| 4 | Терминологическая доступность программного средства для политологов |

### 3.1.1 Диаграммы прецедентов

Диаграмма вариантов использования ~ use case diagram — модель анализа с указанием действующих лиц, которые могут взаимодействовать с системой для выполнения задач, и различные варианты использования, в которых может участвовать действующее лицо.[18]

Предполагается, что с системой будут взаимодействовать два класса пользователей, а именно:

* Эксперт-политолог
* Пользователь

В реальной ситуации один человек может действовать сразу как оба типа пользователей, данное разделение необходимо для дальнейшего описания, ожидаемого поведения и примеров использования программного средства каждым из классов пользователей.

Эксперт-политолог занимается, тем что формирует словари эмоциональных окрасов слов, сохраняет их в отдельные файлы, который далее может передавать пользователям или другим экспертам-политологам

Описанные действия описаны на рисунке 6

Изображение выглядит как текст, диаграмма, линия, Параллельный

Автоматически созданное описание

Рисунок 6 – Диаграмма прецедентов эксперта-политолога

Загрузка словаря будет происходить через интерфейс программы с открытием обзора файлов компьютера или через drag-and-drop. Далее открывается специальный интерфейс редактирование словарей, где можно изменять значения слов и добавлять новые, также в этом же редакторе можно создать полностью новый словарь. После полученный словарь можно сохранить в виде файла и передать его.

Пользователь же использует программу для анализа политических текстов, используя при этом доступные ему словари эмоциональных окрасов слов.

Описанные действия описаны на рисунке 7

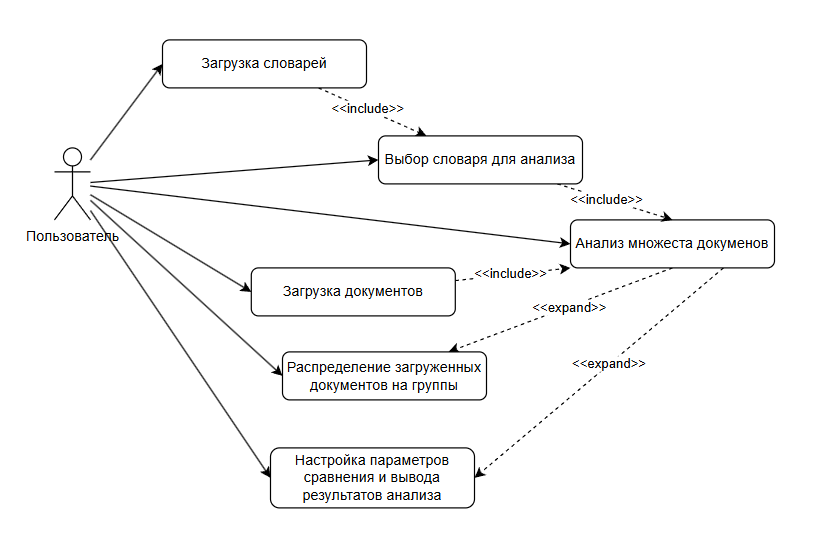


Рисунок 7 – Диаграмма прецедентов пользователя

Загрузка словаря происходит по тому же методу, что и загрузка словарей в случае эксперта-политолога, документы загружаются по тому же принципу, за исключением того, что загрузка предполагает загрузку как одного, так и множества файлов в разных форматах, таких как .docx, .pdf, .txt. После загрузки файлов пользователь может расставить файлам отдельные группы, если файлов много. После выбора файлов/файла для анализа можно поменять словарь, который будет использоваться для анализа, далее пользователь выбирает термины, по которым проводится анализ, выбирает настройки и тип вывода результатов и запускает процесс анализа. По окончанию анализа пользователь получает результаты в выбранной им форме.

### 3.1.2 Функции системы

В таблице 6 показаны требования к функционалу системы анализа политических документов.

Таблица 6 – Требования к функционалу системы анализа политических документов.

| **Название** | Загрузить словарь |
| --- | --- |
| **Действующие лица:** | Эксперт-политолог/Пользователь |
| **Описание:** | Пользователь открывает интерфейс программы, переходит во вкладку “Словари” выбирает пункт “Загрузить словарь”, далее в открывшемся окне нажимает кнопку “Обзор” и выбирает словарь из файлов компьютера или переносит файл со словарём в окно. |
| **Предусловия:** | Файл словаря имеется на компьютере |
| **Постусловия:** | В списке словарей появляется новый словарь, по умолчанию подписанный названием файла. |
| **Основной поток:** | **1.0 Загрузить файл через обзор компьютера**   1. Пользователь открывает интерфейс программы 2. Переходит во вкладку “Словари” 3. Выбирает пункт “Загрузить словарь” 4. Открывается окно загрузки файлов 5. Пользователь нажимает кнопку “Обзор” и выбирает словарь из файлов компьютера 6. Пользователь нажимает кнопку “Загрузить” |
| **Альтернативные потоки:** | **1.1 Загрузить файл переносом в окно (ответвление шага 5)**   1. Пользователь переносит файл в открывшееся окно загрузки 2. Пользователь нажимает кнопку “Загрузить” |
| **Исключения:** | **1.0.E.1 Файл словаря не может быть прочитан или не обладает нужным форматом**   1. Пользователю выводится сообщение об ошибке 2. Пользователя возвращает на вкладку “Словари” 3. Переход к шагу 3 основного потока. |
| **Включает:** | нет |
| **Приоритет:** | средний |
| **Особые требования:** | нет |
| **Допущения:** | нет |
| **Вопросы:** | нет |
| **Название** | Добавить значение в словарь |
| **Действующие лица:** | Эксперт-политолог |
| **Описание:** | Эксперт-политолог открывает интерфейс программы, переходит во вкладку “Словари” выбирает словарь, далее в открывшемся окне нажимает кнопку “Редактировать” и добавляет значение в словарь. |
| **Предусловия:** | В программе имеются словари |
| **Постусловия:** | В списке словарей, а также в папке словарей, появляется новый словарь или изменится старый. |
| **Основной поток:** | **1.0 Редактирование словаря и сохранение как новый словарь**   1. Эксперт-политолог открывает интерфейс программы 2. Переходит во вкладку “Словари” 3. Выбирает в списке нужный словарь 4. Выбирает пункт “Редактировать словарь” 5. Открывается окно редактирования словаря 6. Эксперт-политолог нажимает кнопку добавить новое слово 7. Заполняет необходимые значения 8. Эксперт-политолог нажимает кнопку “Сохранить как” 9. В открывшемся окне выбирает как назвать новый файл и где его создать 10. Эксперт-политолог нажимает кнопку “Сохранить” 11. Файл словаря появляется в указанном месте и с нужным именем, такой же словарь появляется в списке словарей программы. |
| **Альтернативные потоки:** | **1.1 Сохранить изменения в том же словаре (ответвление шага 15)**   1. Эксперт-политолог нажимает кнопку “Сохранить” 2. Изменения сохраняются в том же словаре, который был открыт изначально. |
| **Исключения:** | **1.0.E.1 Поля значений словаря указаны не верно**   1. Изменение полей или добавление нового значения отменяются 2. Выводится уведомление об ошибочном заполнении данных 3. Переход к шагу 5 основного потока. |
| **Включает:** | нет |
| **Приоритет:** | средний |
| **Особые требования:** | нет |
| **Допущения:** | нет |
| **Вопросы:** | нет |
| **Название** | Редактировать значение в словаре |
| **Действующие лица:** | Эксперт-политолог |
| **Описание:** | Эксперт-политолог открывает интерфейс программы, переходит во вкладку “Словари” выбирает словарь, далее в открывшемся окне нажимает кнопку “Редактировать” и начинает редактировать словарь. |
| **Предусловия:** | В программе имеются словари |
| **Постусловия:** | В списке словарей, а также в папке словарей, появляется новый словарь или изменится старый. |
| **Основной поток:** | **1.0 Редактирование словаря и сохранение как новый словарь**   1. Эксперт-политолог открывает интерфейс программы 2. Переходит во вкладку “Словари” 3. Выбирает в списке нужный словарь 4. Выбирает пункт “Редактировать словарь” 5. Открывается окно редактирования словаря 6. Эксперт-политолог дважды кликает по имеющемуся значению в словаре 7. Поля становятся доступными к редактированию 8. Эксперт-политолог меняет значения словаря 9. Эксперт-политолог нажимает кнопку “Сохранить как” 10. В открывшемся окне выбирает как назвать новый файл и где его создать 11. Эксперт-политолог нажимает кнопку “Сохранить” 12. Файл словаря появляется в указанном месте и с нужным именем, такой же словарь появляется в списке словарей программы. |
| **Альтернативные потоки:** | **1.1 Сохранить изменения в том же словаре (ответвление шага 15)**   1. Эксперт-политолог нажимает кнопку “Сохранить” 2. Изменения сохраняются в том же словаре, который был открыт изначально. |
| **Исключения:** | **1.0.E.1 Поля значений словаря указаны не верно**   1. Изменение полей или добавление нового значения отменяются 2. Выводится уведомление об ошибочном заполнении данных 3. Переход к шагу 5 основного потока. |
| **Включает:** | нет |
| **Приоритет:** | средний |
| **Особые требования:** | нет |
| **Допущения:** | нет |
| **Вопросы:** | нет |
| **Название** | Удаление значения в словаре |
| **Действующие лица:** | Эксперт-политолог |
| **Описание:** | Эксперт-политолог открывает интерфейс программы, переходит во вкладку “Словари” выбирает словарь, далее в открывшемся окне нажимает кнопку “Редактировать” и удаляет значение из словаря. |
| **Предусловия:** | В программе имеются словари |
| **Постусловия:** | В списке словарей, а также в папке словарей, появляется новый словарь или изменится старый. |
| **Основной поток:** | **1.0 Редактирование словаря и сохранение как новый словарь**   1. Эксперт-политолог открывает интерфейс программы 2. Переходит во вкладку “Словари” 3. Выбирает в списке нужный словарь 4. Выбирает пункт “Редактировать словарь” 5. Открывается окно редактирования словаря 6. Эксперт-политолог нажимает на значение в словаре правой кнопкой мыши 7. Появляется кнопка удаление записи в словаре 8. Эксперт-политолог нажимает на кнопку удаления 9. Значение в словаре удаляется 10. Эксперт-политолог нажимает кнопку “Сохранить как” 11. В открывшемся окне выбирает как назвать новый файл и где его создать 12. Эксперт-политолог нажимает кнопку “Сохранить” 13. Файл словаря появляется в указанном месте и с нужным именем, такой же словарь появляется в списке словарей программы. |
| **Альтернативные потоки:** | **1.1 Сохранить изменения в том же словаре (ответвление шага 15)**   1. Эксперт-политолог нажимает кнопку “Сохранить” 2. Изменения сохраняются в том же словаре, который был открыт изначально. |
| **Исключения:** | нет |
| **Включает:** | нет |
| **Приоритет:** | средний |
| **Особые требования:** | нет |
| **Допущения:** | нет |
| **Вопросы:** | нет |
| **Название** | Загрузка документов для анализа |
| **Действующие лица:** | Пользователь |
| **Описание:** | Пользователь открывает интерфейс программы, переходит во вкладку “Анализ” выбирает пункт “Загрузить документы”, далее в открывшемся окне нажимает кнопку “Обзор” и выбирает документы из файлов компьютера или переносит документы в окно. |
| **Предусловия:** | нет |
| **Постусловия:** | В списке документов появятся все загруженные документы |
| **Основной поток:** | **1.0 Загрузить файлы через обзор компьютера**   1. Пользователь открывает интерфейс программы 2. Переходит во вкладку “Анализ” 3. Выбирает пункт “Загрузить документы” 4. Открывается окно загрузки файлов 5. Пользователь нажимает кнопку “Обзор” и выбирает документы из файлов компьютера 6. Пользователь нажимает кнопку “Загрузить” |
| **Альтернативные потоки:** | **1.1 Загрузить файл переносом в окно (ответвление шага 5)**   1. Пользователь переносит файлы в открывшееся окно загрузки 2. Пользователь нажимает кнопку “Загрузить” |
| **Исключения:** | **1.0.E.1 Файлы не могут быть прочитан или сохранены в неподдерживаемом формате**   1. Пользователю выводится сообщение об ошибке 2. Переход к шагу 4 основного потока. |
| **Включает:** | нет |
| **Приоритет:** | высокий |
| **Особые требования:** | нет |
| **Допущения:** | нет |
| **Вопросы:** | нет |
| **Название** | Анализ одного документа |
| **Действующие лица:** | Пользователь |
| **Описание:** | Пользователь открывает интерфейс программы, переходит во вкладку “Анализ” выбирает документ из списка, далее выбирает параметры анализа и запускает анализ. |
| **Предусловия:** | В программу были загружены файлы и есть хотя бы один словарь |
| **Постусловия:** | Пользователь получает отчёт по анализу документов |
| **Основной поток:** | **1.0 Запуск анализа документов**   1. Пользователь открывает интерфейс программы 2. Переходит во вкладку “Анализ” 3. Выбирает нужный документ 4. Открывается окно анализа одного документа 5. Пользователь выбирает какой словарь будет использоваться для анализа 6. Пользователь вводит термины, по которым будет производиться анализ 7. Пользователь нажимает кнопку “Анализ” 8. Программа производит анализ текста 9. На основе полученных в ходе анализа данных формируется информативный отчёт в виде таблиц графиков |
| **Альтернативные потоки:** | нет |
| **Исключения:** | **1.0.E.1 Пользователь не выбрал документ для анализа**   1. Кнопка “Анализ” не может быть нажата 2. Переход к шагу 5 основного потока.   **1.0.E.2 Пользователь не ввёл термины для анализа**   1. Кнопка “Анализ” не может быть нажата 2. Переход к шагу 6 основного потока. |
| **Включает:** | нет |
| **Приоритет:** | высокий |
| **Особые требования:** | нет |
| **Допущения:** | нет |
| **Вопросы:** | нет |
| **Название** | Анализ нескольких документа |
| **Действующие лица:** | Пользователь |
| **Описание:** | Пользователь открывает интерфейс программы, переходит во вкладку “Анализ” выбирает группу документов из списка, далее выбирает параметры анализа и запускает анализ. |
| **Предусловия:** | В программу были загружены файлы и есть хотя бы один словарь |
| **Постусловия:** | Пользователь получает отчёт по анализу документов |
| **Основной поток:** | **1.0 Запуск анализа документов**   1. Пользователь открывает интерфейс программы 2. Переходит во вкладку “Анализ” 3. Выбирает группу документов 4. Открывается окно анализа для группы документа 5. Пользователь опционально формирует подгруппы для документов и даёт им названия 6. Пользователь выбирает какой словарь будет использоваться для анализа 7. Пользователь вводит термины, по которым будет производиться анализ 8. Пользователь нажимает кнопку “Анализ” 9. Программа производит анализ текста 10. Пользователь опционально выбирает как группы должны между собой сравнивать результаты анализа 11. На основе полученных в ходе анализа данных и выбранных параметров сравнения формируется информативный отчёт в виде таблиц графиков |
| **Альтернативные потоки:** | нет |
| **Исключения:** | **1.0.E.1 Пользователь не выбрал документы для анализа**   1. Кнопка “Анализ” не может быть нажата 2. Переход к шагу 5 основного потока.   **1.0.E.2 Пользователь не ввёл термины для анализа**   1. Кнопка “Анализ” не может быть нажата 2. Переход к шагу 6 основного потока. |
| **Включает:** | нет |
| **Приоритет:** | высокий |
| **Особые требования:** | нет |
| **Допущения:** | нет |
| **Вопросы:** | нет |

## 3.2 Архитектура системы

Система анализа политических документов состоит из 3 основных модулей

* Модуль работы со словарями – модуль, отвечающий за загрузки, редактирования, сохранения словарей эмоциональных окрасов слов;
* Модуль анализа документов – данный модуль отвечает за анализ документа или множества документов, на вход данный модуль получает документы для анализа, словари, параметры анализа;
* Модуль визуализации – данный модуль отвечает за графический интерфейс, визуализацию результатов анализа документов. Данные, полученные после анализа, сравниваются между собой, и представляются в виде графиков и таблиц.

Для реализации системы анализа политических документов были выделены сущности и на их основе были созданы классы и их функции, которые отображены на рисунке 8.

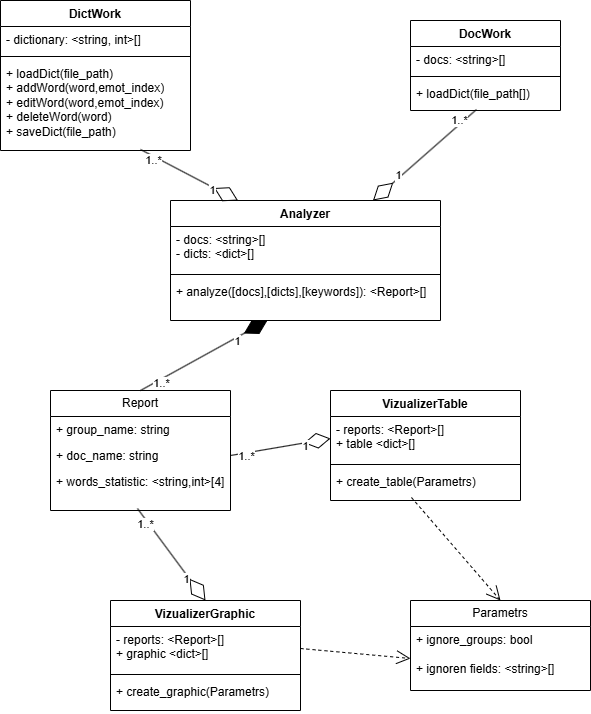


Рисунок 8 – Диаграмма классов

Для реализации модуля работы со словарями, был выделен класс:

**DictWork –** представляет класс для работы со словарями

* **Атрибуты:**
  + dictionary: <string, int>[] - словарь эмоциональных окрасов слов

**Методы:**

* + loadDict(file\_path): Открывает указанный файл и пробует считать его как словарь эмоциональных окрасов слов, в случае успеха записывает данные в dictionary, иначе возвращает ошибку о неправильном формате файла;
  + addTo(word, emot\_index): Добавляет в словарь dictionary, новое значение с ключом word и значением emot\_index, которые характеризует эмоциональных окрасов слов;
  + editWord(word, emot\_index): В словаре dictionary изменяет значение с ключом word на emot\_index, который характеризует эмоциональных окрасов слов;
  + deleteWord(word): В словаре dictionary удаляет значение с ключом word;
  + saveDict(file\_path): Сохраняет dictionary в указанный файл

Данный модуль преобразует взаимодействия эксперта-политолога с пользовательским интерфейсом в изменение заданного словаря, с последующим его сохранением

Для реализации модуля анализа документов, были выделены следующие класс:

**DocWork –** представляет класс для работы документами

**Атрибуты:**

* + Docs: <string>[] - массив строк

**Методы:**

* + load\_docs(file\_path[]): По очереди открывает указанные файл и пробует считать их, в случае успеха записывает данные в docs.

**Report –** класс представление результатов анализа

**Атрибуты:**

* + group\_name: string - название группы документов, которая характеризует принадлежности результата к одной из групп определённых пользователем;
  + doc\_name: string - название документа, характеризует название документа, который был проанализирован;
  + words\_statistic: <string,int>[4] – пары значений характеризующие, количество вхождений и их эмоциональный оценки.

**Analyzer –** представляет класс для анализа документов

**Атрибуты:**

* + docs: <string>[] - массив документов
  + dicts: <dict>[] - массив словарей

**Методы:**

* + analyze([docs],[dicts],[keywords]) <Report>[]: По очереди анализирует переданные в метод документы files, чередуя переданные словари dicts, анализ производится по заданным ключвым словам(терминам) keywords

Данный модуль анализирует документ или множество документов по заданным терминам, используя словарь, или если требуется сравнить словари, по нескольким словарям. Алгоритм анализа был представлен в главе 2.

Для реализации модуля визуализации, были выделены следующие классы:

**Parameters –** представляет класс для хранения параметров преобразования

**Атрибуты:**

* + ignore\_groups: bool – флаг игнорирования групп при анализе
  + ignore\_fields: <string>[] – параметр показывающий какие поля игнорировать при анализе

**VizualizerTable –** представляет класс для преобразования данных анализа, в виды таблиц.

**Атрибуты:**

* + reports: <Report>[] массив результатов анализа;
  + table: <dict>[] массив словарей хранящий данные, которые будут преобразованы в отображение интерфейса

**Методы:**

* + create\_table(Parameters): Представляет данные анализа в виде таблиц, используя при этом указанные параметры для удаления и добавления некоторых полей и столбцов.

**VizualizerGraphic –** представляет класс для преобразования данных анализа, в виды графиков.

**Атрибуты:**

* + reports: Report[] массив результатов анализа
  + graphic: dict{<dict>} словарь словарей хранящий данные, которые будут преобразованы в отображение интерфейса

**Методы:**

* + create\_graphic(Parameters): Представляет данные анализа в виде графика, для этого данные анализа сравниваются между собой, используя при этом указанные параметры для определения параметров сравнения.

# 4. Описание и тестирование программного средства

## 4.1 Инструменты разработки

4.1.1 Языки программирования

Для реализации всех подсистем был использован язык программирования Python 3.10 — это язык программирования, который широко используется в интернет-приложениях, разработке программного обеспечения, науке о данных и машинном обучении (ML). Разработчики используют Python, потому что он эффективен, прост в изучении и работает на разных платформах.

4.1.2 Редактор исходного кода и среда разработки

Для разработки был выбран Visual Studio Code. Visual Studio Code — это упрощенный, но мощный редактор исходного кода, он поддерживает множество языков программирования, включая JavaScript, Python, C++, Java, Go и Rust, благодаря встроенным функциям и расширениям. Интеллектуальное завершение кода (IntelliSense) ускоряет написание программ, предлагая авто дополнение, подсказки и анализ синтаксиса.

4.1.3 Библиотеки

В модуле анализа документов используется библиотека NLTK (Natural Language Toolkit) — это одна из самых старых и наиболее полных библиотек для NLP(Natural Language Processing) на Python. Она предоставляет множество инструментов для различных задач NLP, таких как токенизация, стемминг, лемматизация, парсинг и многое другое. NLTK широко используется в академических кругах и является отличным инструментом для обучения основам NLP.

Для считывания документов типа pdf в модуле анализа документов используется библиотека pypdf — это библиотека для работы с PDF-файлами в Python, которая предоставляет широкие возможности по манипуляции и анализу данных в формате PDF.

## 4.2 Модуль работы со словарями

Модуль работы со словарями отвечает за добавление, удаление и изменение значений из словарей, а также сохранение изменений словаря в файлах.

Для реализации данного модуля использованы следующие инструменты:

* Библиотека pickle для сохранения словаря в файл

Данный модуль имеет следующий функционал:

1. Загрузка словаря из файла нужной кодировки и расширения

* Файл словаря выбирается через обзор файлов операционной системы

1. Работа со значениями в загруженного в программу словаря

* Добавление нового значения
* Удаления существующего значения
* Изменение существующего значения

1. Сохранение изменённого словаря

* Сохранение в изначальный файл, из которого был загружен словарь
* Сохранение в новый файл с указанием названия и места сохранения через обзор файлов операционной системы

## 4.3 Модуль анализа документов

Модуль анализа документов отвечает за анализ документов, а именно считывает текст из файлов, проводит предобработку текста и анализирует на основе выбранного словаря.

Для реализации данного модуля использованы следующие инструменты:

* Библиотека NLTK, а именно инструменты sent\_tokenize для разбимения текста на предложения, word\_tokenize для разбиения текста на слова, stopwords для проверки второстепенных слов, SnowballStemmer для стемминга(приведение к корневой форме).
* Функция PdfReader из библиотеки pypdf для чтения содержимого PDF-файлов

Данный модуль имеет следующий функционал:

1. Загрузка указанного словаря из файла
2. Загрузка указанных файлов документов
3. Считывание содержимого текста из документа

* Проверка расширенрия указанного файла
* Применение соответствующего расширению файла метода считывания файла

1. Предобработка текстов документов

* Разбиение текста на предложения
* Разбиение текста в предложениях на отдельные слова и символы
* Удаление второстепенных слов из текста
* Стемминг всех слов в тексте
* Стемминг слов в словаре

1. Анализ пред обработанного текста

* Поиск вхождения ключевого слова
* Анализ контекста вхождения ключевого слова

1. Сохранение результата анализа

## 4.4 Модуль графического интерфейса

## 4.5 Тестирование

# Список использованных источников

1. Ахременко, А. С. Политический анализ и прогнозирование : учеб. пособие для студентов вузов, обучающихся по направлению подгот. ВПО 030200 - "Политология" / А. С. Ахременко ; А. С. Ахременко. – Москва : Гардарики, 2006. – 333 с. – (Disciplinae : D). – ISBN 5-8297-0292-4. – EDN QOFNTH.
2. Митина Ольга Валентиновна, Евдокименко Александр Сергеевич Методы анализа текста: методологические основания и программная реализация // Психология. Психофизиология. 2010. №40 (216). Дата обращения: 24.12.2023 URL: <https://cyberleninka.ru/article/n/metody-analiza-teksta-metodologicheskie-osnovaniya-i-programmnaya-realizatsiya>
3. Попова Ольга Валентиновна «Измерительный инструмент» в сравнительной политологии: к вопросу о нерешенных проблемах // ПОЛИТЭКС. 2009. №1. Дата обращения: 24.12.2023 URL: https://cyberleninka.ru/article/n/izmeritelnyy-instrument-v-sravnitelnoy-politologii-k-voprosu-o-nereshennyh-problemah
4. Шентякова Анна Владимировна, Гришин Николай Владимирович Мобилизация политического протеста молодежи и российские видеоблогеры: результаты когнитивного картирования // Galactica Media: Journal of Media Studies . 2021. №2. Дата обращения: 24.12.2023 URL: <https://cyberleninka.ru/article/n/mobilizatsiya-politicheskogo-protesta-molodezhi-i-rossiyskie-videoblogery-rezultaty-kognitivnogo-kartirovaniya>
5. Реализация и области применения системы ВААЛ-2000 [Электронный ресурс]. - Режим доступа: URL:<http://www.vaal.ru/proekt/vaal2000.php> Дата обращения: 24.12.2023
6. Concordance: software for concordancing and text analysis [Электронный ресурс]. - Режим доступа: <https://www.concordancesoftware.co.uk> Дата обращения: 24.12.2023
7. Simple Word Sorter [Электронный ресурс]. – Режим доступа: <http://content-analysis.ru/index.php/soft/simple-word-sorter-bazovyj-kolichestvennyj-kontent-a/> Дата обращения: 24.12.2023
8. Программа «Лекта»[Электронный ресурс]. – Режим доступа:<https://kontent-analiz.ru/about> Дата обращения: 24.12.2023
9. Маннинг, К. Введение в информационный поиск / К. Маннинг, Р. Прабхакар, Ш. Хайнрих: Пер. с англ. – М.: ООО «И.Д. Вильямс», 2011. - 528 с.
10. Сайт алгоритма Snowball Stemmer библиотеки NLTK,[Электронный ресурс]. – Режим доступа: https://www.nltk.org/api/nltk.stem.SnowballStemmer.html
11. Porter, M. “An algorithm for suffix stripping” Program 14.3 (1980): 130-137.
12. Сайт инструмента sent\_tokenize библиотеки NLTK,[Электронный ресурс]. – Режим доступа: <https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html>
13. Cайт инструмента word\_tokenize библиотеки NLTK, [Электронный ресурс]. – Режим доступа: <https://www.nltk.org/api/nltk.tokenize.word_tokenize.html>
14. Рахмонзода Нуроншохи Довудшох СООТНОШЕНИЕ ПОНЯТИЙ ЭКСПРЕССИВНОСТИ И ЭМОЦИОНАЛЬНОСТИ В ЛИНГВИСТИКЕ // ELS. 2022. №апрель. URL: https://cyberleninka.ru/article/n/sootnoshenie-ponyatiy-ekspressivnosti-i-emotsionalnosti-v-lingvistike (дата обращения: 03.07.2024).
15. Moussa, M., Mohamed, E. & Haggag, M. (2018) A survey on opinion summarization techniques for social media. Future Computing and Informatics Journal. 3 (1). pp. 82-109.
16. Cайт инструмента sentiment библиотеки NLTK, [Электронный ресурс]. – Режим доступа: https://www.nltk.org/howto/sentiment.html
17. Белоцерковец А.П. СОВРЕМЕННЫЕ МЕТОДЫ СЕНТИМЕНТ-АНАЛИЗА: ПЕРСПЕКТИВЫ, ТЕНДЕНЦИИ, ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ // Вестник науки. 2024. №3 (72). URL: https://cyberleninka.ru/article/n/sovremennye-metody-sentiment-analiza-perspektivy-tendentsii-prakticheskoe-primenenie (дата обращения: 03.07.2024).
18. Вигерс Карл, Битти Джой, Разработка требований к программному обеспечению, Издание третье, дополненное (2014) 9-10, 710