

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение   
высшего образования

**«Дальневосточный федеральный университет»**

(ДВФУ)

**ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ**

**Департамент программной инженерии и искусственного интеллекта**

СМОЛЯНИНОВ ЮРИЙ АЛЕКСАНДРОВИЧ

РАЗРАБОТКА ПЛАТФОРМЫ ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННЕГО ЯЗЫКА.АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

**КУРСОВОЙ ПРОЕКТ**

по дисциплине «Методы и технологии интеллектуализации программных систем»

по образовательной программе подготовки бакалавров по направлению

09.03.04 - Программная инженерия

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |  |  | Студент гр. Б9120-09.03.04прогин | | |
|  |  |  | | Ю. А. Смолянинов |
|  | | | | | | |  |  | (подпись) | |  |
|  | | | | | | | | | | | |
| Защищён с оценкой | | | | | | |  |  | Руководитель | | |
|  | | | | | | |  |  | профессор, к.ф.-м.н., доцент департамента программной инженерии и искусственного интеллекта | | |
|  | | | |  |  | |  |  | ученая степень, должность |  | Л. И. Прудникова |
| (подпись) | | | |  | (И.О. Фамилия) | |  |  | (подпись) |  | (И.О. Фамилия) |
| « |  | » |  | | | 2023 г. |  |  |  | | |
|  | | | | | | | | | | | |

г. Владивосток

2023

**Оглавление**

[Введение 3](#_Toc134985907)

[1. Обоснование актуальности 5](#_Toc134985908)

[1.1 Обзор существующих платформ для обработки текстов на естественном языке 5](#_Toc134985909)

[1.1.1 Общий функционал 5](#_Toc134985910)

[1.1.2 Особенности платформ 5](#_Toc134985911)

[1.1.3 Сравнительная таблица платформ для обработки текстов на естественном языке 8](#_Toc134985912)

[1.2 Актуальность проекта 9](#_Toc134985913)

[2 Анализ предметной области 10](#_Toc134985914)

[2.1 Объекты предметной области 10](#_Toc134985915)

[2.2 Модель естественного языка Тузова В. А. 11](#_Toc134985916)

# **Введение**

С каждым годом количество информации, с которой сталкивается человек растет экспоненциально. В следствие чего возникает потребность предобработки больших объемов текстов для возможности удобной работы с ними.

Поэтому целесообразно использовать платформы для обработки текста, так как программная платформа - это совокупность программных решений и технологий, позволяющая осуществлять разработку и/или производство программных продуктов в определенной предметной области методом компоновки и настройки уже готовых модулей и модулей, специально разработанных в целях расширения функционала и включаемых затем в платформу [6]. Исходя из определения платформы видно, что она сможет сочетать в себе весь функционал для предобработки и анализа больших объемов текстов. Конечно, уже существуют платформы для обработки текстов на естественном языке. Однако, в основном, они являются коммерческими разработками, поэтому предоставляют свой функционал платно. Помимо этого, существующие платформы либо не обладают полным набором функций для обработки текстов на русском языке, либо не учитывают спецификации русского языка.

С другой стороны, существующие платформы для обработки текстов ориентированы на поиск, менеджмент данных разных корпоративных предприятий, поэтому в основном анализ текстов рассчитан на определение тематики и кластеризацию, а не полноценную обработку текстов.

В связи с этим актуальной задачей является разработка платформы для обработки естественного языка, рассчитанной на полноценную работу с текстом на русском языке и имеющей широкий функционал для работы с текстом, например, реферирование текстов, определение тональности и т.п.

В рамках выпускной работы, была поставлена цель - создание многофункциональной платформы для обработки текста на естественном языке.

Целью курсовой работы является: обзор существующих платформ для обработки текстов на естественном языке.

Для достижения поставленной цели необходимо выполнить следующие задачи:

1. Найти и изучить существующие платформы для обработки текстов на естественном языке.

2. Провести сравнительный анализ существующих платформ.

3. Найти и выделить недостатки существующих платформ.

# **1. Обоснование актуальности**

## **1.1 Обзор существующих платформ для обработки текстов на естественном языке**

В данном разделе представлены особенности существующих платформ для обработки текстов на естественном языке, а именно IQPLATFORM, Elasticsearch, M-Brain Intelligence Plaza, PolyAnalyst Text и их общий функционал.

### **1.1.1 Общий функционал**

Каждая система имеет возможность осуществлять настройку и управление функциональностью, возможность импорта и/или экспорта данных в популярных файловых форматах, позволяют получать систематизированные и визуализированные данные из системы для последующего анализа.

### **1.1.2 Особенности платформ**

В данном разделе представлены отличительные особенности и недостатки каждой из платформ, описанных ранее.

#### **1.1.2.1 Платформа IQPLATFORM**

IQPLATFORM – это цифровая аналитическая платформа, которая позволяет выполнять продвинутую аналитику на базе больших объёмов информации, синтез новых знаний, мониторинг и контроль информационных объектов [1].

Отличительной чертой платформы является поддержка только русского языка. Также IQPLATFORM предназначен для сбора, обработки, хранения и глубинного анализа больших объёмов структурированных и неструктурированных данных из различных типов источников. Программное обеспечение IQPLATFORM первоочерёдно ориентировано на средние и крупные предприятия кредитно-финансовой сферы, промышленности и ТЭК, а также на организации государственного сектора и некоммерческие организации. Поэтому текстовый анализ содержит следующие функции [1]:

* Определение языка документа для десятков наиболее распространенных языков мира с применением классификатора предметных областей;
* Автоматическая классификации входящих документов по списку рубрик с фильтрацией по атрибутам документов или их текстовому содержанию;
* Морфологический анализ слов и синтез стандартных словоформ с помощью как собственных, так и дополнительно подключенных морфологических модулей;
* Нормализация российских адресов для приведения их к стандартному формату с указанием кода БД ФИАС и ОКАТО;
* Нормализация имен для определения фамилии, имени и отчества персоны с коррекцией ошибок, транслитерации и уменьшительно-ласкательных форм;
* Выделение сюжетов - выявление групп входящих документов с аналогичной темой в реальном времени, с возможностью просмотра информации о них из интерфейса системы.

Из выше сказанного можно сделать вывод, что у платформы отсутствует функция реферирования текстов на естественном языке.

Другими недостатками платформы являются отсутствие открытого исходного кода, отсутствие визуализации данных и возможности в реальном времени оперативно получать агрегированную информацию.

#### **1.1.2.2 Платформа** **Elasticsearch**

Платформа Elasticsearch – это программное обеспечение, главной отличительной чертой которого является открытый исходный код, также оно предназначено для поиска, сбора, анализа и хранения текстовых данных с использованием интеллектуальных алгоритмов. Помимо этого, ключевые функции системы Elasticsearch включают индексацию данных, автоматическое восстановление узлов, управление жизненным циклом индекса, ведение журнала аудита и уведомления о проблемах [2].

Аналитика в Elasticsearch основана на полнотекстовом поиске и фасетах. Фасеты в поиске — это некая агрегация по определенному признаку [3]. Платформа ориентирована на поиск и анализ больших данных, откуда следует недостаток - отсутствие обработки текста как такого. Например, у платформы нет функции реферирования текстов, определения тональности текста и т.п.

#### **1.1.2.3 Платформа M-Brain Intelligence Plaza**

M-Brain Intelligence Plaza — то ИТ-платформа для управления потоками информации о рынках и конкурентах для отделов аналитики, продаж, маркетинга, менеджмента. Платформа предназначена для хранения в едином месте, интеллектуального структурирования и распределения деловой информации из различных источников компании. В системе Intelligence Plaza может обрабатываться любой контент, включая внутрикорпоративные данные от разных поставщиков или из внешних источников, контент о бизнес-среде по любым отраслям, в форме кратких деловых сводок, аннотаций и резюме, создаваемых аналитиками M-Brain [4].

Платформа предназначена для менеджмента данных разных корпоративных предприятий, поэтому в основном анализ текстов рассчитан на определение тематики и кластеризацию, а не на полноценную работу с текстом. Поэтому у платформы нет функции реферирования или аннотирования текстов.

#### **1.1.2.4 Платформа Megaputer PolyAnalyst Text**

PolyAnalyst Text - платформа интеллектуального анализа текстовых данных [5]. Система органично сочетает в себе передовые инструменты обработки естественного языка, средства семантического текстового анализа и алгоритмы машинного обучения, что позволяет пользователям получать результаты высокой точности. Основные функции для обработки текста [5]:

* Кластеризация;
* Обезличивание;
* Анализ тональности;
* Извлечение фактов;
* Классификация;
* Резюмирование.

Недостатками является отсутствие открытого исходного кода и, несмотря на широкий функционал обработки текста, данная платформа распространяется только по платному тарифу. Помимо этого, недостатком также является отсутствие полной документации.

### **1.1.3 Сравнительная таблица платформ для обработки текстов на естественном языке**

После рассмотрения существующих платформ для обработки текстов на естественном языке, необходимо построить таблицу, показывающую характеристики каждой из рассмотренных систем.

Таблица 1 — Сравнительная таблица платформ обработки естественного языка

| **Платформы**    **Характеристики** | PolyAnalyst Text | IQPLATFORM | Elasticsearch | M-Brain Intelligence Plaza |
| --- | --- | --- | --- | --- |
| Свободное ПО | - | - | + | - |
| Возможность работы с текстом | + | + | + | + |
| Поддержка русского языка | + | + | - | + |
| Наличие подробной документации | - | + | + | - |
| Бесплатный доступ | - | - | + | - |
| Функция реферирования | - | - | - | + |
| Импорт/экспорт данных | + | + | + | + |

## **1.2 Актуальность проекта**

Каждая из вышерассмотренных платформ для обработки текстов на естественном языке не сочетает в себе бесплатный доступ, наличие подробной документации, открытый исходный код, поддержку русского языка. Вследствие чего, актуальной задачей является проектирование, и разработка многофункциональной платформы, которая будет включать в себя вышеописанные пункты.

**2 Анализ предметной области**

Рассматриваемая профессиональная деятельность — это разработка платформы для обработки текстов на естественном языке.

В рамках данной работы выделяются следующие подзадачи:

1. Определение основных требований к платформе:

* Выделение основных функциональных возможностей платформы;
* Какие естественные языки должна поддерживать платформа;
* Определение структур данных и методов;
* Проектирование интерфейса платформы.

1. Выбор средств для реализации функциональных возможностей платформы:

* Выбор языка программирования;
* Выбор библиотек, фреймворков и т.д.

1. Реализация платформы.

**2.1 Объекты предметной области**

С профессиональной деятельностью в рассматриваемой предметной области связаны следующие объекты:

* Текст на естественном языке — текст, написанный на русском языке и обладающий смысловой целостностью, состоящий из одного или нескольких предложений.
* Словарь — множество записей, каждая из которых записана в новой строке.
* Открытая\закрытая система
* Платформа — это совокупность программных решений и технологий, позволяющая осуществлять разработку и/или производство программных продуктов в определенной предметной области методом компоновки и настройки уже готовых модулей и модулей, специально разработанных в целях расширения функционала и включаемых затем в платформу [7].
* Модуль — функционально законченный фрагмент программы, оформленный в виде отдельного файла с исходным кодом [8].
* Выбор предметной области. Область применения
* Язык платформы - формальный язык, предназначенный для записи компьютерных программ [9].

## **2.2 Модель естественного языка Тузова В. А.**

В модели естественного языка Тузова выделяется следующие компоненты модели [10]:

**1. Формальный («семантический») язык**, позволяющий в строгом однозначном виде записать любое предложение естественного языка. Формальный язык представляет собой алгебраическую систему , где – базисные функции на языке, – структура данных языка [11].

Базисные функции являются основным средством описания (формализации) смысла слов естественного языка. Класс базисных функций представляет собой некоторую иерархию. На вершине этой иерархии находятся самые абстрактные, а потому наиболее универсальные функции. Они являются универсальными в том смысле, что область применения и их порождающая способность практически не ограничены [11].

Почти половина глаголов русского языка являются каузирующими, поэтому примером базисной функции является функция :

*.*

Аналогично глаголам в модели Тузова В. А. дается определение через базисные функции следующим частям речи: предлогом (простым, сложным), объектом (существительным), операциям над объектами (прилагательным), действиям (глаголам), действиям над функциями (наречия), союзам и т.д. например:

Так как в русском языке четко выделены части речи – глагол, существительное и т.д. то каждая часть речи играет определенную роль в организации синтаксической структуры предложения [11]:

* Существительные, являясь аргументами функций, образуют структуру данных языка.
* Прилагательные – простейшие функции на существительных; эквивалентны операторам присваивания.
* Краткие прилагательные – простейшие функции на существительных; отображаются на одноместные предикаты.
* Глаголы – хорошо развитые функции в основном на существительных; отображаются на многоместные предикаты.
* Наречия – функции на глаголах; эквивалентны операторам присваивания.
* Простые предлоги – функции на существительных; преобразуя грамматический тип существительного, организуют связь слов в предложении.
* Сложные предлоги и союзы – функции, аргументами которых являются суперпозиции функций; способны связывать целые предложения.

В итоге получается, что предложение является алгоритмом и каждое слово должно быть описано некоторой совокупностью формул. Множество таких описаний есть семантический словарь.

**2. Семантический словарь**, содержащий формальные представления слов естественного языка. В словаре слова делятся на две группы: слова, которые не выражаются через другие, т.е. базовые или же лексемы (примерно 20000 слов) и производные, т.е. их смысл выражается через суперпозицию базисных функций и базовых слов.

Для базового слова можно только указать его класс (см. табл. 2) и, в ряде случаев, краткое формальное описание. Например, слово абразив принадлежит классу «поделочные материалы» и описано как «вещество, используемое для обработки металлов» [10]:

Таблица 2 – Пример базовых слов

|  |  |
| --- | --- |
| **Слово** | **Класс** |
| Иванов | Существительное / Физический объект / Живой / Человек / Индивид / ФИО / Фамилия |
| радость | Существительное / Психика / Душа / Чувства / Довольство / Радость |
| абразив | Существительное / Физический объект / Неодушевленный / Материалы / Поделочные |

Таблица 3 – Пример производных слов

|  |  |
| --- | --- |
| **Название** | **Описание** |
| жечь |  |
| аморфный |  |
| мулатка |  |

В семантическом словаре каждому из слов языка сопоставлены одна или несколько лексем. Каждая из лексем снабжена собственным обозначением и толкованием. Толкование лексемы определяет ее допустимое множество сущностей (ее значений) в зависимости от состава, вида и значения ее аргументов. Каждое такое описание включает морфологические характеристики, семантический класс, список слов и конструкций, присоединяемых этим словом и описание семантики слова с помощью семантических функций [12].

Например,

Таким образом описание слова в семантическом словаре содержит следующие элементы [12]:

* Морфологические признаки;
* Семантический класс;
* Семантико-грамматический тип,
* Описание моделей управления: слова и конструкции каких типов и классов может присоединять данное слово;
* Описание семантики слова в виде формулы.

Для ряда слов некоторые элементы описания отсутствуют, например, описание семантики для базовых слов.

Описывая слова подобным образом, т.е. давая определение базовым словам через базисные функции и производным через лексемы и базисные функции получается, что для толкования слов дается через формальный («семантический») язык, при этом толкование производного слова дается в виде формулы, устанавливающей зависимость этого слова от других слов.

**3. Анализатор**, преобразующий текст на естественном языке к формальному виду.

**Список литературы**

1. iqmen.ru [Сайт]. URL: <https://www.iqmen.ru/iqplatform>
2. elastic.co [Сайт] URL: httop://www.elastic.co/elasticsearch/
3. ElasticSearch 1.0 — новые возможности аналитики [Электронный ресурс] // habr.com, 2014 — Режим доступа: https://habr.com/ru/company/vk/blog/213849/
4. m-brain.com [Сайт]. URL: https://www.m-brain.com/intelligence-plaza/
5. megaputer.ru [Сайт]. URL: https://www.megaputer.ru/produkti/polyanalyst-text/
6. Исследование программных платформ для создания информационных систем Вооруженных Сил Российской Федерации / О. В. Масленников, В. А. Шевченко, А. Г. Карманов, Р. С. Веревкин. - (Информатизация вооруженных сил). - Текст: непосредственный // Военная мысль. - 2021. - № 2. - С. 91-97: 1 табл. - Библиогр. в примеч. - Примеч.: с. 97 (4 назв.). - ISSN 0236-2058.
7. Гуляева К.А., Артемьева И. Л. Особенности применения онтологического подхода в разработке интеллектуальных систем для некоторых задач химии / К.А. Гуляева, И.Л. Артемьева // Онтология проектирования. – 2020. – Т.10, №3(37). - С.307-326. – Режим доступа:<https://www.ontology-of-designing.ru/article/2020_3(37)/Ontology_Of_Designing_3_2020_5_K.A.Gulyaeva,I.L.Artemieva.pdf>
8. В. Фаронов. Создание приложений с помощью C#, Москва, ЭКСМО, 2008, с. 12
9. ISO/IEC/IEEE 24765:2010 Systems and software engineering — Vocabulary
10. Мозговой М. В. Машинный семантический анализ русского языка и его применения / Мозговой М. В. // Диссертация на соискание ученой степени кандидата физико-математических наук. – 2006. – С.110.
11. Тузов В. А. Компьютерная семантика русского языка. - СПб.: Изд-во СПбГУ, 2003. – XXX с.
12. Кутарба А. Ю. Построение семантического словаря для обработки англоязычных текстов / Кутарба А. Ю. // АВТОРЕФЕРАТ диссертации на соискание ученой степени кандидата физико-математических наук. – 2006. – С.20.