МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение  
высшего образования

**"Южно-Уральский государственный университет**

**(национальный исследовательский университет) "**

**Высшая школа электроники и компьютерных наук**

**Кафедра системного программирования**

**ОТЧЕТ**  
**о выполнении научно-исследовательской работы**бакалавра направления 02.03.02 «Фундаментальная информатика и информационные технологии»

|  |  |
| --- | --- |
|  | Автор работы:  студент группы КЭ-201  Циммерман Юлия Владимировна  Научный руководитель:  Доктор физико-математических наук, доцент кафедры  «Системное программирование»  М.Л. Цымблер  Оценка: \_\_\_\_\_\_\_\_\_\_\_  Дата: \_\_\_\_\_\_\_\_\_\_\_  Подпись: \_\_\_\_\_\_\_\_\_\_\_ |

Челябинск 2021МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение  
высшего образования

**"Южно-Уральский государственный университет**

**(национальный исследовательский университет)"**

**Высшая школа электроники и компьютерных наук**

**Кафедра системного программирования**

УТВЕРЖДАЮ

Зав. кафедрой СП

\_\_\_\_\_\_\_\_\_\_ Л.Б. Соколинский

01.03.2021 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

студенту группы КЭ-201

Циммерман Юлия Владимировна

обучающемуся по направлению 02.03.02 «Фундаментальная информатика и информационные технологии»

1. **Тема НИР в семестре**Применение технологий искусственного интеллекта в виртуальных собеседниках на примере разработки простого чат-бота
2. **Срок сдачи студентом законченной работы**  31.05.2021 г.
3. **Перечень подлежащих разработке вопросов**

Обзор научных статей по тематике применения искусственных нейронных сетей для разработки чат-ботов

**Научный руководитель**

Доктор физико-математических наук, М.Л. Цымблер  
доцент кафедры   
«Системное программирование»

**Задание принял к исполнению** Ю.В. Циммерман

**Оглавление**

[Введение 4](#_Toc69768395)

[1. Программа чат-бот и история ее возникновения 5](#_Toc69768396)

[1.1 Понятие «чат-бот» 5](#_Toc69768397)

[1.2 История возникновения чат-ботов 5](#_Toc69768398)

[1.3 Классификация чат-ботов 6](#_Toc69768399)

[2. Реализация чат-бота 8](#_Toc69768400)

[2.1 Нейронные сети в обработке естественного языка 8](#_Toc69768401)

[2.2 Обучение чат-ботов 9](#_Toc69768402)

[2.3 Сложности при создании чат-бота 9](#_Toc69768403)

[2.4 Первичный морфологический анализ данных 10](#_Toc69768404)

[2.4.1 Токенизация 10](#_Toc69768405)

[2.4.2 Нормализация 10](#_Toc69768406)

[2.4.3 Лемматизация 10](#_Toc69768407)

[2.4.4 Удаление стоп-слов 11](#_Toc69768408)

[2.5 Векторизация текста 11](#_Toc69768409)

[2.6 Работа с Telegram при разработке чат-бота 12](#_Toc69768410)

[3. Проектирование 13](#_Toc69768411)

[3.1 Архитектура программы 13](#_Toc69768412)

[3.2 Модульная структура программы 14](#_Toc69768413)

[3.2.1 Модуль «Первичный морфологический анализ» 14](#_Toc69768414)

[3.2.2 Модуль «Векторизация текста» 15](#_Toc69768415)

[3.2.3 Модуль «Генератор ответа» 16](#_Toc69768416)

[4. Реализация программы 17](#_Toc69768417)

[4.1. Модуль «Морфологический анализ текста» 17](#_Toc69768418)

[4.2. Модуль «Векторизация текста» 17](#_Toc69768419)

[4.3. Модуль «Генератор ответа» 17](#_Toc69768420)

[Заключение 18](#_Toc69768421)

[Приложение 1. Модуль «Первичный морфологический анализ текста». 21](#_Toc69768422)

[Приложение 2. Модуль «Векторизация текста» 24](#_Toc69768423)

[Приложение 3. Генератор ответа. 25](#_Toc69768424)

[Приложение 4. Реализация модуля «Морфологический анализ текста». 26](#_Toc69768425)

**Обзор научных работ по тематике разработки чат-ботов**

# Введение

Целью данной работы является обзор статей по тематике применения искусственных нейронных сетей для разработки чат-ботов.

Актуальность тематики по разработке чат-ботов обусловлена тем, что виртуальное общение становится основным видом коммуникации в современном мире. В настоящее время чат-боты представлены в широком спектре [19]. Так, в маркетинге их основными задачами являются поддержание интереса потребителя, ответы на вопросы и помощь в покупке какого-либо товара. Помимо этого, многие компании прибегают к использованию виртуальных помощников для проведения финансовых услуг и распространения рекламы. Также чат-боты оказали большое влияние на сферу образования за счет различных информационных рассылок и завлечения в образовательные курсы и вебинары. В таких сферах, как туризм и банки, чат-боты в основном выполняют функцию консультантов, таким образом, помогают потребителю подобрать подходящий тариф или путевку. Согласно исследованию Business Insider, рынок чат-ботов к 2024 году возрастет на 30%. Внедрение чат-ботов поможет сэкономить затраты на поддержку клиентов, а возможность коммуникации с компанией будет доступна в любое время [22].

1. **Программа чат-бот и история ее возникновения**
   1. **Понятие «чат-бот»**

*Чат-бот* – это программа, специализированная на общении с пользователем при помощи текста или голоса, имитирующая человеческую речь [21]. Основными задачами, выполняемыми чат-ботами, являются понимание естественного языка, ведение диалога и синтез ответа пользователю [16].

* 1. **История возникновения чат-ботов**

История чат-ботов насчитывает около 70 лет [1]. Идея возникновения чат-ботов была придумана Аланом Тьюрингом в 1950 [10]. Именно он первым задался вопросом, можно ли создать программу, поддерживающую с пользователем такой диалог, что его было бы невозможно отличить от реального общения с человеком.

Однако первый чат-бот был разработан лишь в 1966 году Джозефом Вейценбаумом [12]. Программа под названием *«Элиза»* выполняла работу психотерапевта, реализуя технику активного слушания. Конечно, ее функции были ограничены, но именно она послужила дальнейшему развитию данного направления.

Так, следующим чат-ботом была программа под названием *«Parry»*, созданная в 1972 году американским психиатром Кеннетом Колби, имитирующая поведение больного шизофренией. Главным его отличием считается то, что данный чат-бот имеет «личность», соответственно по-особенному реагирует на сообщения пользователя [2]. Несмотря на это, Parry также являлся чат-ботом с низким уровнем возможностей.

Настоящим прорывом в данном направлении послужило создание *«A.L.I.C.E»* в 1995 году. Это чат-бот, способный поддерживать диалог на любом естественном языке и на любую тему. Конечно, он так же, как и другие, был далек от совершенства, и не мог генерировать человекоподобные ответы с выражением эмоций [11].

В 2001 году был сделан еще один шаг в развитии чат-ботов. Была разработана программа с развитием *SmarterChild*, доступная на таких платформах, как America Online и Microsoft, способная помогать людям решать бытовые задачи [7]. Всю информацию бот брал из различных источников, таких, как новости, погода, фильмы, спортивные результаты и др.

Первым персональным помощников в истории чат-ботов стала так называемая «*Siri»*, принадлежащая компании Apple. Siri дает рекомендации пользователю в виде аудио-, видео- и графических ответов, на основе различных Интернет-ресурсов [8]. Еще один персональный помощник по имени *«Cortana»*, разработанный в 2014 году, был способен анализировать и реагировать на голосовые команды пользователя. Так, он выполнял различные функции, как поддержка напоминаний, рассылка писем, общение с пользователем, игры, определение времени и др.

Новый этап в развитии чат-ботов случился в 2016 году, когда на платформах социальных сетей появилась возможность создавать ботов для улучшения сервиса собственных брендов [1]. Так, огромное количество ботов могло обмениваться сообщениями с клиентами абсолютно разных областей, отвечать на вопросы и решать их проблемы.

* 1. **Классификация чат-ботов**

В настоящее время существуют два вида классификации чат-ботов. Так, выделяются *бизнес-классификация*, разделяющая ботов исходя из выполняемых ими задач, и *техническая классификация*, основанная на принципе их работы [21].

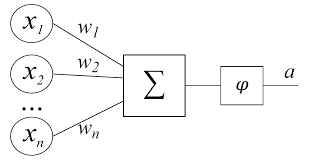
*Бизнес-классификация* разделяет ботов на три категории. К ним относятся разговорные чат-боты, главной задачей которых является поддержание беседы с пользователем, персональные ассистенты, помогающие пользователю выполнять различные задачи и Q&A боты, дающие ответы на конкретные вопросы.

*Техническая классификация* также выделяет три типа чат-ботов. К ним относятся программы, основанные на бизнес-правилах. Диалог с ними ведется по определенному пути, где пользователь принимает решения, но не может получить ответ на вопрос на отвлеченную тему. Противоположностью данного типа являются чат-боты, основанные на искусственном интеллекте. Они не имеют определенного пути и постоянно учатся, используя предыдущие диалоги. Третьим типом являются *гибридные* боты, которые включают в себя элементы первого и второго типов. Таким образом, такие боты ведут диалог по определенному пути, но используют искусственный интеллект при разговоре с пользователем.

1. **Реализация чат-бота**
   1. **Нейронные сети в обработке естественного языка**

В настоящее время нейронные сети являются лучшим способом распознавания речи и обработки естественного языка [20]. Чат-бот на основе нейронных сетей обучается в процессе работы, именно поэтому он является лучшим интеллектуальным виртуальным помощником на рынке чат-ботов.

*Искусственные нейронные сети* являются объединением искусственных нейронов, которые расположены на слоях внутри нейронной сети. Понятие искусственного нейрона возникло в 1943 году, в его основе лежал искусственный нейрон. Данная идея была представлена в работе Уоррена МакКаллока и Уолтера Питтса, согласно которой в *нейрон* (рис. 1) поступают сигналы, равные бинарным величинам (xi), которые внутри нейрона умножаются на вес (wi), соответствующий входу данного значения [5]. Затем функция активации (φ) преобразует суммарный импульс в выходное значение нейрона.



**Рис. 1.** Модель искусственного нейрона.

Обучение нейронных сетей реализуется за счет определения весов соединений между нейронами. Необходимо сделать это таким образом, чтобы сеть приближала функцию к заданной точности.

Существуют три способа обучения нейронных сетей [20]. Так, при первом методе, имеющем название *обучение с учителем*, на вход нейронной сети поступает набор сигналов с известным результатом, следовательно, веса будут меняться в соответствии с правильностью выходного сигнала. Отличием *обучения без учителя* является то, что нейронная сеть не знает ожидаемый верный результат. Третий подход – *обучение с подкреплением*, в данном случае нейронная сеть так же, как и в предыдущем методе, не получает результат заранее, а выводит его после контакта с внешней средой.

* 1. **Обучение чат-ботов**

Для того чтобы обучить чат-бота необходим большой набор текстовых данных, которые улучшат качество и логичность его ответов.

Существует большое количество источников, находящихся в свободном доступе, которые можно использовать в качестве данных для обучения чат-бота. К данным источникам относятся субтитры к фильмам и сериалам, тексты художественной литературы, сервис микроблогинга Twitter, публичные чаты мессенджеров, комментарии в социальных сетях, различные тематические веб-форумы и др. [16]. Затем полученные данные должны пройти несколько этапов обработки текста: первичный морфологический анализ и векторизация текста, в результате чего будет получена база данных – основа анализа запроса пользователя и получения ответа на него.

* 1. **Сложности при создании чат-бота**

Работа с текстом вызывает множество сложностей при разработке чат-бота. Одной из основных проблем является *мультиязычность*, часто встречающаяся в разговорном языке [3]. Так, в тексте может встречаться сленг и заимствованные из других языков слова или цитаты. Для решения данной проблемы необходима модель, способная обобщать разные языки в общем векторном пространстве [16].

Также важными деталями при разработке чат-бота являются неконсистентность ответов, то есть получение разных ответов на одинаковые по смыслу, но по-разному сформулированные вопросы, и ответы в виде общих фраз. Именно поэтому при создании чат-ботов чаще всего прибегают к выводу ответа из заранее заготовленного набора [16].

Однако важнейшей проблемой является *нечеловечность* чат-бота, которая не позволяет диалогу достичь *реалистичности*. Основным способом ее решения является использование различных источников, располагающих способами коммуникации между людьми. Так, на основе этих данных чат-бот получает различные варианты ответа на пользовательский запрос, которые будут соответствовать реальному общению между людьми [9].

* 1. **Первичный морфологический анализ данных**

Для наиболее качественного получения признаков из исходного текста для его дальнейшей классификации необходимо проводить первичный морфологический анализ полученных на входе данных [14]. С помощью *классификации* текста чат-бот подбирает наиболее подходящие запросу ответы.

*Предварительная обработка текста* включает в себя несколько этапов, каждый из которых необходим для дальнейшего обучения программы.

* + 1. **Токенизация**

*Токенизация* представляет собой разбиение текста на более мелкие части, такие как абзацы, предложения и слова, которые имеют название *токены*. Существует множество библиотек и программ-токенизаторов в особенности для таких популярных языков программирования, как Python и JavaScript. Для токенизации текста в Python чаще всего используются такие библиотеки, как «NLTK» и «re» [15].

* + 1. **Нормализация**

Основной задачей *нормализации* является приведение запроса к стандартному виду, то есть преобразование текста к единому регистру, удалению знаков препинания и словесному написанию чисел [14]. Все эти действия можно осуществить благодаря доступным функциям работы со строками различных языков программирования.

* + 1. **Лемматизация**

*Лемматизация* отвечает за преобразование слов к своей первоначальной форме. Так, глаголы приводятся в инфинитивную форму, а существительные и прилагательные к именительному падежу.

Функций, реализующих лемматизацию, как правило, нет в базовом наборе библиотек у какого-либо языка программирования. Однако существует множество доступных для установки программ и библиотек, в их числе такие популярные библиотеки, как «NLTK», «SpaCy» и PyMorphy2, однако лишь PyMorphy2 поддерживает возможность обработки текста на русском языке [14].

* + 1. **Удаление стоп-слов**

Данный этап можно осуществить путем удаления всех шумовых слов, то есть за счет исключения всех слов, которые не удалось идентифицировать. Такие слова не должны влиять на смысловое значение полученного запроса, поэтому данный этап реализует функцию очистки текста от лишнего.

Удаление шумовых слов можно произвести с помощью встроенного списка стоп-слов в библиотеке «NLTK» на Python. Также можно создать собственный список слов, которые не несут смысловой нагрузки в текст, чтобы в дальнейшем программа могла их удалить, не изменив суть запроса.

* 1. **Векторизация текста**

После проведения всех этапов предобработки запроса пользователя полученный текст необходимо преобразовать в векторный формат. В процессе векторизации текстовые данные кодируются в числовой вид, которые на основе заданных алгоритмов будут формировать логичный ответ пользователю [17].

Существует большое количество *моделей векторизации текста*. Так, в модели «TF-IDF» [4] вычисляется произведение частотности слов в тексте и обратной частотности слова в базе данных. *Частотность* слов определяет вероятность появления слова в тексте, а *обратная частотность* его уникальность. Размерность векторов зависит от количества слов в запросе. Так, наибольший вес приобретают слова, часто встречающиеся в конкретном запросе, а не в изначальной базе данных.

Методы векторизации текста получили наибольшую популярность после возникновения нового подхода в 2013 году, который получил название «*Word2Vec*». В данном алгоритме векторное представление показывает контекстную близость слов, опираясь на модель непрерывного мешка слов [6].

Перечисленные выше методы называются *статическими*, так как в них не учитываются многозначность слов и их контекстно-зависимая природа [17]. Для решения данной проблемы были созданы *динамические* методы векторизации текста. Так, наиболее успешной оказалась модель под названием «*BERT*», повлекшая дальнейшее развитие данного направления. Данная модель может восстанавливать текст на основе контекста, что имеет название *эмбеддинг*.

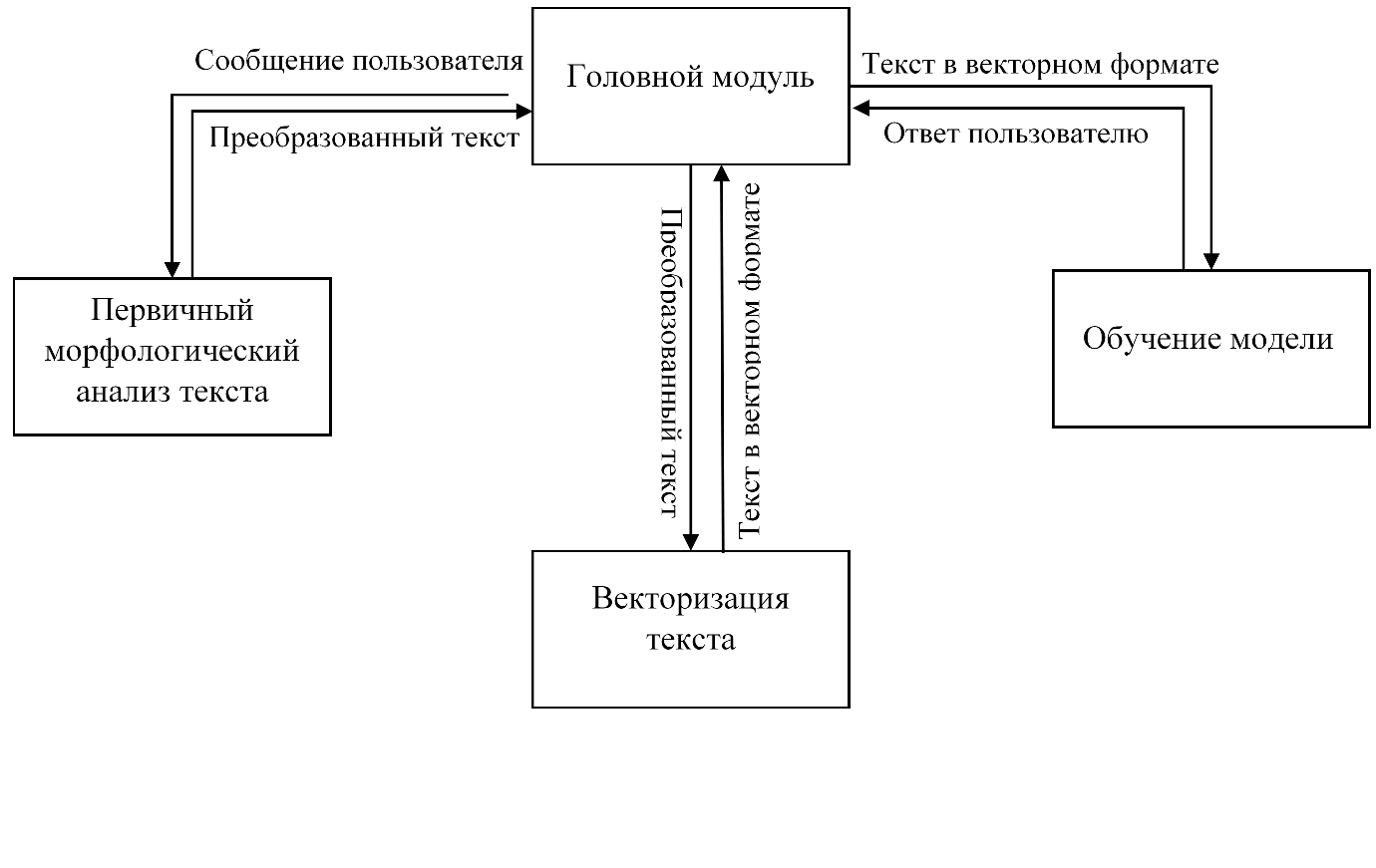
* 1. **Работа с Telegram при разработке чат-бота**

В настоящее время наиболее удобным мессенджером, поддерживающим подключение чат-бота, является Telegram. Начиная с 2015 года, Telegram добавил функцию создания и подключения собственного чат-бота, что повлекло за собой появление различных Интернет-помощников по реализации широкого спектра функций [18].

Для подключения собственного чат-бота к Telegram существует специальная библиотека «*pyTelegramBotAP*» для Python. Данный модуль включает в себя различные редакторы, такие как добавление команд, базовых сообщений и др. [13].

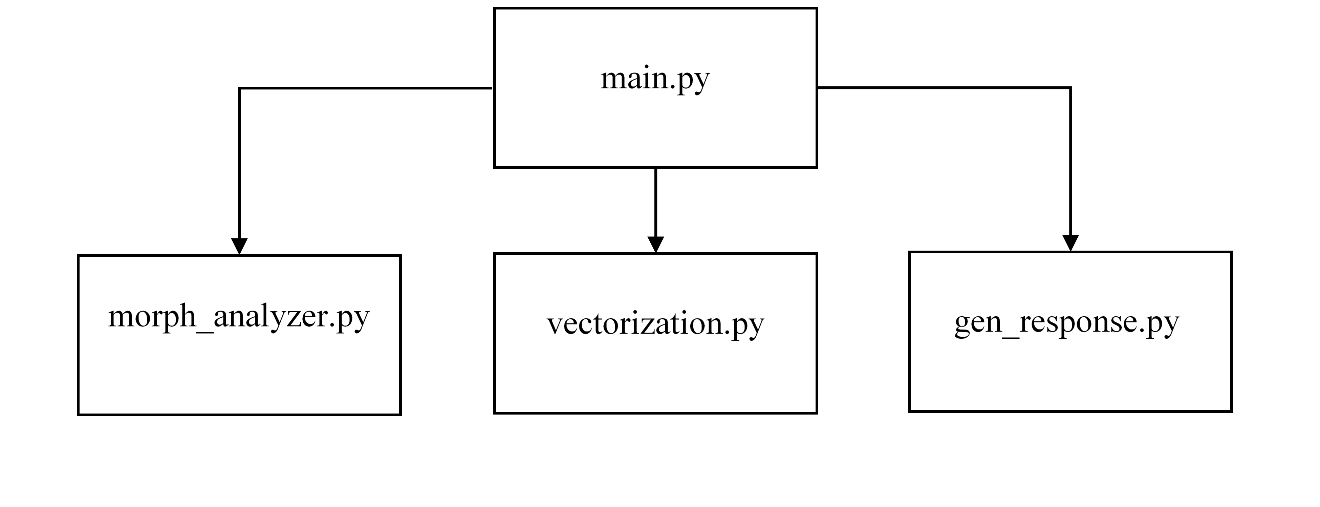
1. **Проектирование**
   1. **Архитектура программы**

Архитектура программы (*Рис. 2*) представляет собой трехэтапный процесс, состоящий из первичного морфологического анализа текста, векторизации текста и обучения модели. Так, перед обучением нейронной сети текст проходит этапы обработки и векторизации.

**Рис. 2.** Архитектура программы.

* 1. **Модульная структура программы**

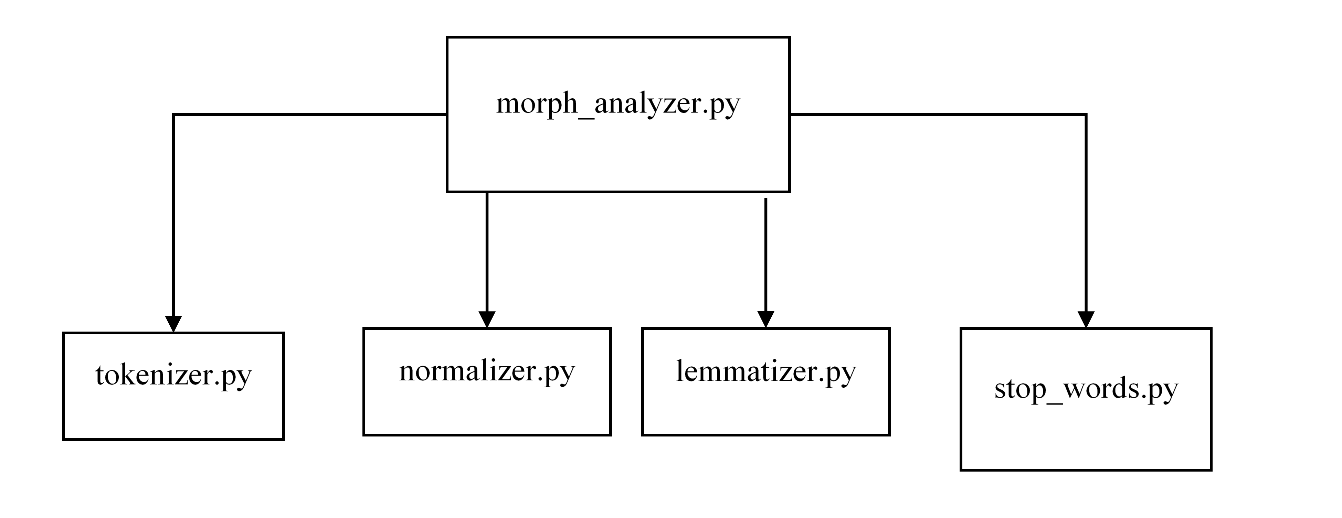
На рисунке 3 представлена модульная структура программы в общем виде. Головной модуль управляет запуском программного продукта.



**Рис. 3.** Модульная структура программы.

### **3.2.1 Модуль «Первичный морфологический анализ»**

Первичный морфологический анализ реализуется в модуле *morph\_analyzer.py*, который преобразовывает сообщение пользователя в формат, подходящий для дальнейшей с ним работы. Морфологический анализ включает в себя такие этапы, как токенизация, нормализация, лемматизация и удаление стоп-слов (*Рис. 4*).



**Рис. 4.** Модульная структура модуля "Первичный морфологический анализ текста".

На этапе токенизации в подмодуле *tokenizer.py* запрос пользователя записывается в двумерный массив, состоящий из сообщения, разбитого на предложения, а предложения, в свою очередь, на слова. Для данной процедуры используются функции python-библиотеки *nltk*.

Подмодуль *normalizer.py* отвечает за нормализацию текста: удаляет все знаки препинания в массиве. Для реализации данного процесса используются функции базового набора python.

В подмодуле *lemmatizer.py*, реализующем лемматизацию, элементы массива приводятся к своей начальной форме, что осуществляется в результате использования функции python-библиотеки *pymorphy2*.

Последний этап морфологического анализа – удаление стоп-слов. Он осуществляется в подмодуле морфологического анализа stop\_words.py. Для реализации используется пакет *nltk.corpus* от python-библиотеки *nltk*, откуда импортируется список стоп-слов русского языка.

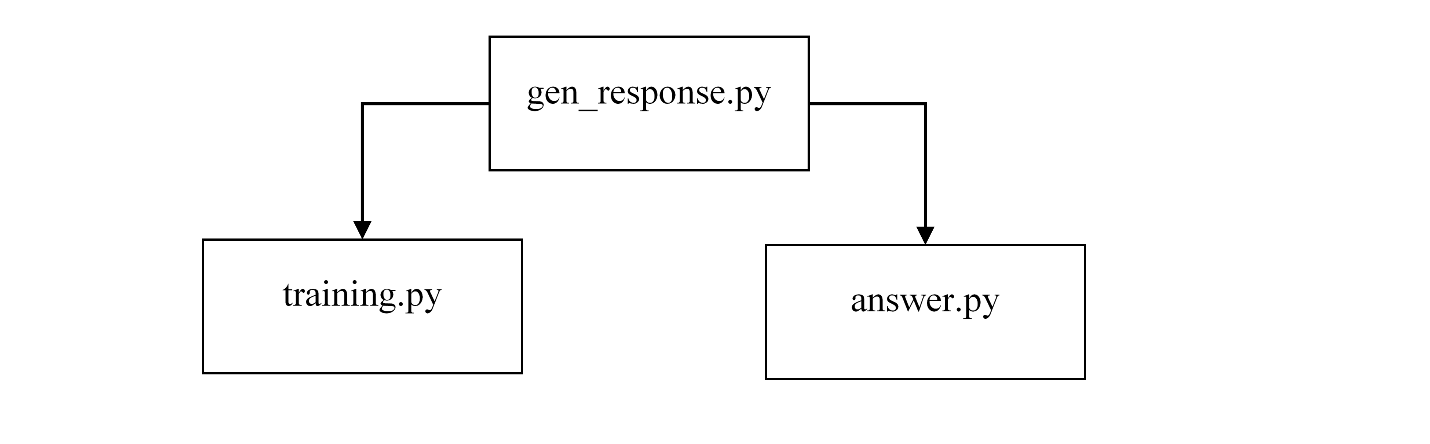
Подробное описание модуля morph\_analyzer.py и его подмодулей представлено в приложении 1.

### **3.2.2 Модуль «Векторизация текста»**

Модуль *vectorization.py* приводит сообщение пользователя, преобразованное морфологическим анализом, и тексты, требующиеся для обучения нейронной сети в векторный формат. Для данного действия используется python-библиотека *sklearn*. Этот модуль реализует метод TF-IDF, который делится на два этапа. Так, TF определяет частоту появления конкретного слова в одном текста, а IDF вычисляет, как редко встречается это слово во всех доступных текстах. Таким образом, коэффициент TF-IDF — это вес, который будет использоваться для обработки информации и оценки важности слова для текста в некотором наборе текстов. Подробное описание модуля представлено в приложении 2.

### **3.2.3 Модуль «Генератор ответа»**

Модуль *gen\_response.py* (*Рис. 5*) реализуется с помощью нейронной сети. Главной её задачей является генерация логичного ответа пользователю на его запрос. Для начала необходимо обучить нейронную сеть на основе готовых текстов, для чего потребуется отдельный подмодуль training.py. Второй подмодуль answer.py будет отвечать за вывод ответа. Детальные интерфейсы данного модуля представлены в приложении 3.



**Рис. 5.** Модуль «Генератор ответа».

1. **Реализация программы**

## **4.1. Модуль «Морфологический анализ текста»**

Листинг данного модуля представлен в приложении 4.

## **4.2. Модуль «Векторизация текста»**

Заготовка

## **4.3. Модуль «Генератор ответа»**

Заготовка

# Заключение

Чат-бот – это программа, формирующая корректные, логически обоснованные ответы, схожие с реальным разговором людей, на основе полученного от пользователя запроса.

В настоящее время чат-боты используются в широком спектре и затрагивают большое количество сфер. В основном они реализуют функции персонального помощника в коммерческих компаниях и функцию поддержания диалога в различных сервисах. Чат-боты способны решать различные задачи, такие как общение, развлечение, медицинские консультации, проведение финансовых услуг, заказа товаров и др.

Для получения логических ответов на запрос пользователя важно провести его морфологический анализ, а затем представить данные в векторном формате. Чтобы добиться создания интеллектуального чат-бота, необходимо решить следующие проблемы: мультиязычность, неконсистентность и нечеловечность.

**Список литературы**

1. Eleni Adamopoulou, Lefteris Moussiades. CHATBOT: Architecture, Design, & Development // Machine Learning with Applications. 2020.
2. Colby K.M., Weber S., Hilf F.D. Artificial paranoia. Artificial Intelligence. 1971. Vol. 2, No 1. P. 1–25.
3. Holger S., Douze M. Learning Joint Multilingual Sentence Representations with Neural Machine Translation // ACL workshop on Representation Learning for NLP. 2017. P. 157-167.
4. Manning C.D., Raghavan P., Schütze H. An Introduction to Information Retrieval // Cambridge University Press, 2009, p. 581.
5. McCulloch W.S., Pitts W. A Logical Calculus of the Ideas Immanent in Nervous Activity // The Bulletin of Mathematical Biophysics. 1943. Vol. 5, No. 4. P. 115–133.
6. Mikolov T. Distributed Representations of Words and Phrases and their Compositionality // Proceedings of Workshop at ICLR. 2013.
7. Molnar G., Zoltán S. The role of chatbots in formal education. International Symposium on Intelligent Systems and Informatics. 2018.
8. Nora Ni Loideain, R. Adams. From Alexa to Siri and the GDPR: The gendering of Virtual Personal Assistants and the role of Data Protection Impact Assessments // Computer Law & Security Review. 2020. Vol. 36.
9. Rui Yan. “Chitty-Chitty-Chat Bot”: Deep Learning for Conversational AI // Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018.
10. Turing A.M. Computing machinery and intelligence. Mind. 1950. Vol. LIX, No. 236. P. 433–460.
11. Wallace R.S., R. Epstein, G. Roberts, G. Beber. The anatomy of A.L.I.C.E. Parsing the turing test: philosophical and methodological issues in the quest for the thinking computer. 2009. P. 181–210.
12. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. Commun. 1966. Vol. 9, No. 1. P. 36–45.
13. Бийбосунов Б.И., Бийбосунова С.К., Жолочубеков Н.Ж. Описание концепции Telegram ботов и их разработка // Colloquium-journal. 2020. Vol. 7, No. 59. P. 3-6.
14. Бородин А.И., Вейнберг Р. Р., Литвишко О. В. Методы обработки текста при создании чат-ботов // Хуманитарни Балкански изследования. 2019. Т. 3, № 5. С. 108-111.
15. Гречанин В.А. К вопросу о токенизации текста // Международный научно-исследовательский журнал. 2016. Т. 6-4, № 48. С. 25-27.
16. Жеребцова Ю.А., Чижик А.В. Создание чат-бота: обзор архитектур и векторных представлений текста // International Journal of Open Information Technologies. 2020. Т. 8, № 7. С. 50-66.
17. Жеребцова Ю.А., Чижик А.В. Сравнение моделей векторного представления текстов в задаче создания чат-бота // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2020. Т. 18, № 3. С. 16-34.
18. Козлов А.А., Батищев А.В. Телеграм-бот как простой и удобный способ получения информации // Территория науки. 2017. № 5, С. 55-64.
19. Параскевов А.В., Каданцева А.А., Мороз С.И. Перспективы и особенности разработки чат-ботов // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2017. № 130. С. 395-404.
20. Созыкин А.В. Обзор методов обучения глубоких нейронных сетей // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2017. Т. 6, № 3. С. 28–59.
21. Ураев Д.А. Классификация и методы создания чат-бот приложений // International scientific review. 2019. № LXIV. С. 30-33.
22. Будущее чат-ботов: 10 исследований и прогнозы экспертов [Электронный ресурс]. URL: <https://vc.ru/services/175276-budushchee-chat-botov-10-issledovaniy-i-prognozy-ekspertov> (дата обращения: 26.12.2020).

# 

# Приложение 1. Модуль «Первичный морфологический анализ текста».

**Модуль morph\_analyzer.py:** def morph(raw), token\_text, norm\_text, lem\_text, fin\_text.

Спецификация: morph\_analyzer.py импортирует подмодули tokenizer.py, normalizer.py, lemmatizer.py и stop\_words.py и реализует их вызов. Внутри реализуется функция def morph(raw), входным параметром которой является строка raw, введенная пользователем и переданная из головного модуля. Параметр token\_text – результат работы подмодуля tokenizer.py. Переменная norm\_text – выходной параметр подмодуля normalizer.py. Параметр lem\_text – результат работы подмодуля lemmatizer.py. Переменная fin\_text – выходные данные подмодуля tokenizer.py и всего модуля.

**Подмодуль tokenizer.py**: def tokenize(raw), sent\_tokens, a, i, j, word.

Спецификация: подмодуль tokenizer.py реализует разбиение исходного сообщения пользователя на предложения, которые, в свою очередь, разбиваются на слова. Для реализации этих задач используется библиотека nltk, осуществляющая функции разделения текста на предложения и слова, sent\_tokenize и word\_tokenize соответственно. Параметр raw, запрос пользователя (строка), приводится к нижнему регистру. Параметр sent\_tokenize – массив предложений параметра raw. Переменная a – изначально пустой массив, пополняющийся в ходе выполнения циклов. Переменные i, j – целочисленные счетчики циклов. Первый цикл выполняется количество раз, численно равное количеству элементов переменной sent\_tokenize. В нем определяется параметр word – одно из слов массива sent\_tokenize, добавляющееся в массив a в следующем цикле. Выходным параметром данного подмодуля является параметр a – двумерный массив предложений, разбитых на слова.

**Подмодуль normalizer.py**: def normalize(a), punctuation, i, word.

Спецификация: подмодуль normalizer.py отвечает за удаление знаков препинания, что реализует функция def normalize(a) с входным параметром a – массивом слов, переданным подмодулем tokenizer.py. Параметр punctuation – список, состоящий из возможных знаков препинания. Затем в цикле с целочисленным счетчиком i происходит проверка на вхождение элементов массива a, хранящихся в переменной word, в список punctuation и его удаление в случае вхождения. Выходным параметром является массив a без знаков препинания.

**Подмодуль lemmatizer.py**: def lemmatize(a), lem\_text, morph, i, word, p.

Спецификация: подмодуль lemmatizer.py осуществляет лемматизацию текста, то есть приведение слов к своей начальной форме. Реализация данного действия происходит за счет импортированной библиотеки pymorphy2 и её функций. Входным параметром является текущий массив a, переданный подмодулем normalizer.py. Переменная lem\_text – пустой массив, в котором будет записан результат работы подмодуля. Параметр morph хранит в себе функцию MorphAnalyzer. Переменные i – целочисленный счетчик, word – элемент массива a. Затем каждый отдельно обработанный элемент, хранящийся в параметре p добавляется в массив lem\_text, который является выходным параметром.

**Подмодуль stop\_words.py**: def stop(lem\_text), filtered\_text, stop\_words, i, token.

Спецификация: подмодуль stop\_words.py реализует удаление стоп-слов с помощью пакета nltk.corpus и её функции stopwords. Входным параметром является результат работы подмодуля lemmatizer.py массив lem\_text. Массив stop\_words, используя функцию stopwords импортированного пакета, хранит в себе все базовые стоп-слова русского языка. Параметр filtered\_text – изначально пустой массив, в который в дальнейшем будут добавлены обработанные элементы. Переменные i – целочисленный счетчик цикла, token – элемент массива lem\_text. В цикле происходит проверка на вхождения элемента token в список stop\_words и добавление его в массив filtered\_text в случае невхождения.

# Приложение 2. Модуль «Векторизация текста»

**Модуль vectorizer.py**: text, vector\_text.

Спецификация: модуль vectorizer.py осуществляет преобразование обработанного модулем morph\_analyzer.py текста в массив чисел. Для реализации данной задачи используются пакеты и функции библиотеки sklearn. Так, входной параметр text – преобразованный массив слов вышеупомянутым модулем, передается из модуля main.py. Выходным параметром является параметр vector\_text – массив, хранящий в себе информацию о важности слова и числе его вхождений в текст.

# Приложение 3. Генератор ответа.

**Модуль gen\_response.py:** text, answer.

Спецификация: модуль gen\_response.py на входе получает преобразованное сообщение пользователя text. Выходным параметром является строка answer, которая будет являться ответом на запрос пользователя.

**Подмодуль training.py**: chatbot.txt.

Спецификация: подмодуль training.py обеспечивает обучение нейронной сети на основе готовых текстов.

**Подмодуль answer.py**: text, answer.

Спецификация: подмодуль answer.py реализует получение ответа, заключенное в строковой переменной answer, на запрос пользователя text, являющегося массивом слов.

# Приложение 4. Реализация модуля «Морфологический анализ текста».

1. **Листинг модуля morph\_analyzer.py**

import tokenizer  
import normalizer  
import lemmatizer  
import stop\_words  
  
def morph(raw):  
  
 # Токенизация  
 token\_text = tokenizer.tokenize(raw)  
   
 # Нормализация  
 norm\_text = normalizer.normalize(token\_text)  
  
 # Лемматизация  
 lem\_text = lemmatizer.lemmatize(norm\_text)  
  
 # Удаление стоп-слов  
 fin\_text = stop\_words.stop(lem\_text)  
  
 return fin\_text

1. **Листинг подмодуля tokenizer.py**

import nltk  
  
# Разбиение на предложения и слова  
def tokenize(raw):  
 raw = raw.lower()  
 sent\_tokens = nltk.sent\_tokenize(raw)  
 a = []  
 for i in range(len(sent\_tokens)):  
 word = nltk.word\_tokenize(sent\_tokens[i])  
 a.append([])  
 for j in range(len(word)):  
 a[i].append(word[j])

return a

1. **Листинг подмодуля normalizer.py**

# Нормализация  
punctuation = ['(', ')', '?', ':', ';', ',', '.', '!', '/', '"', "'"]  
print(type(punctuation))  
def normalize(a):  
 for i in range(len(a)):  
 for word in a[i]:  
 if word in punctuation:  
 a[i].remove(word)  
 s = 0; print("Нормализация: ")  
 for i in a:  
 s += 1;  
 print(f"Предложение {s}:")  
 print(i)  
 return a

1. **Листинг подмодуля lemmatizer.py**

# Лемматизация  
import pymorphy2  
  
def lemmatize(a):  
 lem\_text = []  
 morph = pymorphy2.MorphAnalyzer()  
 for i in range(len(a)):  
 lem\_text.append([])  
 for word in a[i]:  
 p = morph.parse(word)[0]  
 lem\_text[i].append(p.normal\_form)

return lem\_text

1. **Листинг подмодуля stop\_words.py**

# Удаление стоп-слов  
from nltk.corpus import stopwords  
  
def stop(lem\_text):  
 filtered\_text = []  
 stop\_words = stopwords.words("russian")  
 for i in range(len(lem\_text)):  
 filtered\_text.append([])  
 for token in lem\_text[i]:  
 if token not in stop\_words:  
 filtered\_text[i].append(token)  
  
 return filtered\_text