

Инженерная школа информационных технологий и робототехники

Направление 09.03.01 «Информатика и вычислительная техника»

Отделение информационных технологий

Отчёт по теме

**«Чат-бот ТПУ(Персональный помощник студента ТПУ)»**

по дисциплине «Творческий проект»

Выполнил:

студент гр. 8В11 \_\_\_\_\_\_\_\_\_\_\_\_\_ Агафонов Н.В.

\_\_\_.\_\_\_.\_\_\_\_\_\_

Проверил:

доцент отделения ОИТ ИШИТР

оценка (до 30 б.): \_\_\_\_\_\_\_\_\_\_\_\_\_

к защите допускаю \_\_\_\_\_\_\_\_\_\_\_\_\_ Фадеев А.С.

\_\_\_.\_\_\_.\_\_\_\_\_\_

Томск 2023

**Содержание**

[ВВЕДЕНИЕ 3](#_Toc124512130)

[Цель и задачи 4](#_Toc124512131)

[1. NLP – обработка естественного языка и её подходы. 5](#_Toc124512132)

[2. Выбор инструментов разработки. 6](#_Toc124512133)

[2.1. Выбор языка программирования и библиотеки. 6](#_Toc124512134)

[2.2. Библиотека spaCy. 6](#_Toc124512135)

[2.3. Сложности при работе с NLP. 7](#_Toc124512136)

[3. Разработка алгоритма. 9](#_Toc124512137)

[3.1. Описание алгоритма. 9](#_Toc124512138)

[ЗАКЛЮЧЕНИЕ 12](#_Toc124512139)

[СПИСОК ИСТОЧНИКОВ 13](#_Toc124512140)

### ВВЕДЕНИЕ

Общение с клиентами сегодня требует интуитивных решений, которые ориентированы на пользователя и обеспечивают эффективное взаимодействие с ним, в идеале снижая затраты на обслуживание. Поэтому компании внедряют в свои чат-боты технологию NLP, которая позволяет улучшить качество обслуживания в сторону немедленной и эффективной помощи.

Боты для обработки естественного языка (NLP) предназначены для преобразования текста или речи, вводимых пользователем, в структурированные данные. Данные в дальнейшем используются для выбора соответствующего ответа. НЛП включает в себя важные этапы, такие как токенизация, анализ настроений чат-ботов, распознавание сущностей и анализ зависимостей.

При этом чат-боты внедряются не только предпринимателями, но и больницами, службами такси, государственными структурами и сферами образования.

### Цель и задачи

Цель работы:

Проектирование и внедрение в Телеграм чат-бота «Персональный помощник студента» технологии NLP.

 Для достижения общей поставленной цели необходимо выполнить следующие задачи:

* Изучить инструменты и технологии NLP;
* выбрать оптимальный алгоритм NLP;
* составить требования к проекту;
* разработать алгоритм NLP;
* внедрить алгоритм в телеграмм чат-бота «Персональный помощник студента».

### 1. NLP – обработка естественного языка и её подходы.

Обработка естественного языка (Natural Language Processing, NLP) — пересечение [машинного обучения](https://neerc.ifmo.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5) и математической лингвистики, направленное на изучение методов анализа и синтеза естественного языка. Сегодня NLP применяется во многих сферах, в том числе в голосовых помощниках, автоматических переводах текста и фильтрации текста. Основными тремя направлениями являются: [распознавание речи](https://neerc.ifmo.ru/wiki/index.php?title=%D0%A0%D0%B0%D1%81%D0%BF%D0%BE%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D1%80%D0%B5%D1%87%D0%B8)  (Speech Recognition), понимание естественного языка (Natural Language Understanding ) и генерация естественного языка (Natural Language Generation[[3]](https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D0%B5%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0#cite_note-3)). [1]

NLP имеет несколько основных подходов и операций для распознания текста. Самое первое действие любого NLP-приложения — разбор текста на токены, которые могут быть словами, числами или знаками препинания. Токенизация — самая первая операция в конвейере, поскольку для всех остальных операций необходимо, чтобы текст был разобран на токены. Далее следует извлечь различные лингвистические признаки с помощью специальных операций.

Также, может случится ситуация, когда слова, наиболее часто встречающиеся во множестве слов одного из классов словаря, синтаксически различаются. Для задачи анализа тональности текста разница в синтаксисе не влияет на принятие решений, поскольку важна лишь оценочная семантика слова. Такими словами, могут быть, например, слова *плохо, плохое, плохая* и т.д. Эти слова семантически одинаковые, а синтаксически разные. Такое различие в синтаксисе приводит к тому, что подсчет веса для этих слов будет производится раздельно, что будет оказывать влияние на точность модели. Чтобы исправить данную ситуацию, применяются методы обработки естественных языков для приведения словоформ в каноническую форму. Для рассматриваемой задачи такими методами могут быть:

1) Стемминг – это процесс нахождения основы слова для заданного исходного слова. Основа слова не обязательно совпадает с морфологическим корнем слова. Например: *кровать* *крова*;

2) Лемматизация – процесс приведения словоформы к лемме – её нормальной (словарной) форме. Например: *стояли* *стоять*, *раскидистые* *раскидистый*;

3) Удаление стоп-слов – удаление из текста документа слов, которые самостоятельно не несут никакой смысловой нагрузки. Например: местоимения *я, ты, мы, вы*. [2]

### 2. Выбор инструментов разработки.

2.1. Выбор языка программирования и библиотеки.

NLP можно реализовать на чем угодно PHP, Python, Node is, Go, семейство C. Cложность написания на всех вышеперечисленных языках будет примерно одинаковой, поэтому для разработки был выбран Python, так как этот язык достаточно прост в изучении, функционален и есть множество библиотек для работы с NLP.

Также нужно выбрать библиотеку, с помощью которой будет реализован алгоритм распознавания текста. Из общего множества была выбран spaCy, т.к. эта библиотека имеет обширную документацию, включая удобное для начинающих руководство, бесплатный интерактивный онлайн-курс и ряд [видеоуроков](https://www.youtube.com/c/ExplosionAI).

2.2. Библиотека spaCy.

spaCy — это бесплатная библиотека с открытым исходным кодом для расширенной обработки естественного языка (NLP) в Python. Он разработан специально для производственного использования и помогает создавать приложения, которые обрабатывают и «понимают» большие объемы текста. Его можно использовать для создания систем извлечения информации или понимания естественного языка.[3]

Обзор функций:

* Поддержка 72+ языков
* 80 обученных пайплайнов для 24 языков
* Многозадачное обучение с предварительно обученными преобразователями , такими как BERT
* Предварительно обученные векторы слов
* Непревзойденная скорость
* Готовая к производству система обучения
* Лингвистически мотивированная токенизация
* Компоненты для распознавания именованных сущностей , тегирования частей речи, анализа зависимостей, сегментации предложений, классификации текста , лемматизации, морфологического анализа, связывания сущностей и т . д.
* Легко расширяемый с помощью настраиваемых компонентов и атрибутов
* Поддержка пользовательских моделей в PyTorch , TensorFlow и других фреймворках
* Встроенные визуализаторы синтаксиса и NER
* Простая упаковка модели , развертывание и управление рабочим процессом
* Надежная, тщательно оцененная точность[3]

2.3. Сложности при работе с NLP.

Нейросети сегодня могут говорить, писать, слушать и понимать, благодаря обработке естественного языка (NLP). Она извлекает значения из сообщений и структурирует их. Но нам все еще сложно найти общий язык с машинами, т.к возникает ряд проблем при распознавании текста.

Язык синонимичен и многозначен. Так, сложно математически выразить нюансы фразы с омонимами: «график (художник) не вписался в график (план)». Подобных примеров много.

Язык — это не только система правил, но и исключений. Это также эмоции, жесты, культурный и бытовой контекст, метафоры, игра слов или сарказм. Компьютер легко переводит с английского на суахили, но проблема в том, что ни того ни другого он не знает по-настоящему. Поэтому главная задача исследователей — научить алгоритмы извлекать смыслы из слов и работать с образами. Человек, например, запоминает смысл высказывания, а не его форму.

За извлечение смысла отвечает семантический анализ текста.  Правильное применение слов в контексте обеспечивает прагматический анализ. Есть и другие уровни обработки естественного языка: морфологический, синтаксический, фонологический. Но с ними алгоритмы уже справляются.

Компьютеры понимают не все языки одинаково хорошо. Лучше всего нейросети работают с популярных языками с фиксированной структурой (например, английским).

Также важно учитывать количество описанных слов, наличие текстов в разных жанрах для обработки (книги, журналы, пресса), наличие тезаурусов и словарей с аннотированными текстами. Всего 20 языков (из более 7000) подготовлены для работы с нейросетью. Хорошо обработаны западноевропейские языки, китайский и японский. Отсутствуют ресурсы для многих африканских и азиатских, а также вымирающих языков и местных диалектов. Практически невозможно обработать языки без письменности.

Модели обнаруживают простейшие закономерности, объясняющие данные.  Например,  есть классификатор изображений, где представлены собаки и кошки. Изображения собак имеют оттенки серого, а кошки — цветные

Модель, скорее всего, зафиксирует ложную корреляцию между наличием/отсутствием цвета и тегами. Если попадётся изображение собаки в цвете, программа, вероятно, идентифицирует ее как кошку.

Огромный объем данных, необходимых для машинного обучения (в частности, глубокого обучения), берут из интернета.  Из-за этого наборы наследуют нежелательные свойства текстов (дублирование, статистические ошибки, ложь), которые сложно обнаружить и удалить. Это тоже приводит к неверным корреляциям.[4]

### 3. Разработка алгоритма.

3.1. Описание алгоритма.

Алгоритм должен быть внедрен в чат-бота, чтобы связать её с сервисом «help.tpu.ru». Тем самым автоматизировать работу специалистов, отсекая вопросы, на которые есть ответ в базе данных бота. Также, если бот не сможет автоматически ответить на вопрос, должна быть возможность переслать вопрос специалисту и получить ответ на вопрос в этом же чате.

Исходя из вышесказанного, можно представить примерную работу алгоритма.

Пользователь задает вопрос, к примеру: "Что делать, если не назначен преподаватель по дисциплине?"

Из этого вопроса отбрасываются знаки пунктуации, стоп-слова, потом происходит токенизация по словам, далее лемматизация, потом находятся подлежащее, сказуемое и ближайшие дополнения к ним. Все слова записываются в список. Этот список сравнивается с каждым значением столбца, в котором хранятся ключевые слова вопроса, в таблице и возвращает процент совпадения, если процент совпадения наибольший, то пользователю возвращается ответ на вопрос и появляется возможность отправить специалисту, если ответ не устроил. Если процент совпадения будет небольшим, то бот сразу дает возможность отправить специалисту, при этом записывает в таблицу ключевые слова, а после ответа специалиста будет ответ записывать к ключевым словам в столбец с ответами.

* 1. Работа с данными.

Для работы с данными потребуется СУБД. На начальном этапе была выбрала SQLite, потому что главное – проверка работы алгоритма, в дальнейшем планируется перейти на другую СУБД.

В SQLite была создана таблица, состоящая из трех столбцов: id, question, answer(рис.1).

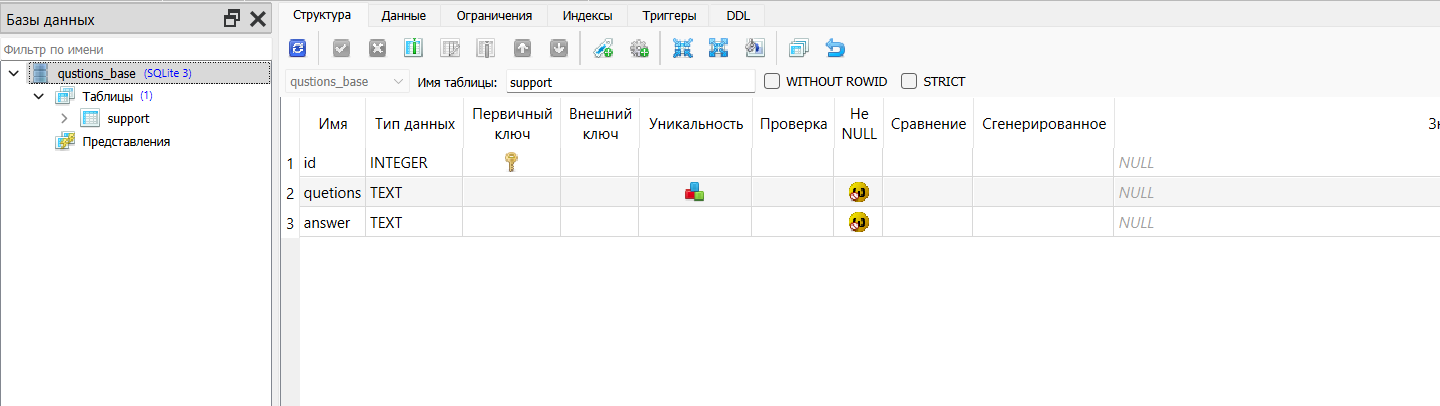


Рис.1. База данных.

* 1. Пример работы №1.

Пользователь задает вопрос, к примеру: "Что делать, если не назначен преподаватель по дисциплине?"[5]. В результате всех преобразований имеется список вида ['не назначать', 'преподаватель', 'дисциплина'].

В таблице созданы значения похожие на этот список и сам список в столбце «questions»(рис. 2.1).

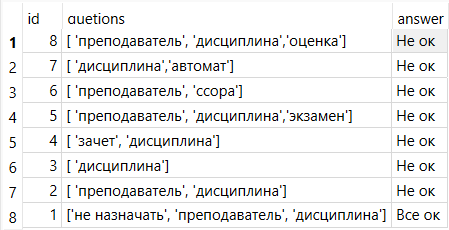


Рис.2.1. Заполненная таблица.

Алгоритм прогоняет преобразованный список по таблице, сверяя его со значениями из столбца «questions», записывая максимальный процент совпадения и значения «id» в словарь.

Для проверки алгоритма в «answer» вписаны значения «Все ок» и «Не ок», чтобы понимать корректность работы алгоритма. И в результате работы программы(рис.2.2) можно заметить, что алгоритм правильно распознал вопрос и вернул корректный ответ

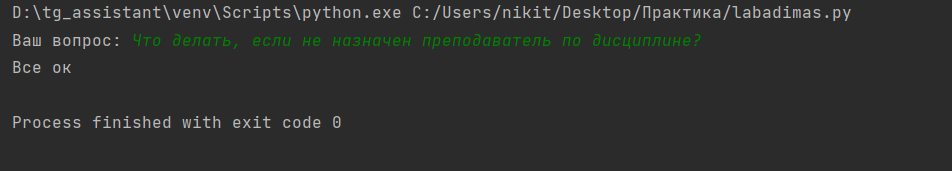


Рис.2.2. Результат работы.

* 1. Пример работы №2.

В предыдущем примере рассмотрен вариант работы алгоритма, когда соответствующий вопрос уже есть в базе, но возможен вариант событий, когда заранее подготовленного ответа может и не быть.

Таблица остается неизменной(рис.2.1.). Пользователь задает вопрос «Когда проходит учебная практика?»[5]. Сравнивая с значениями из таблицы, алгоритм не находит подходящего ответа, поэтому предлагает отправить вопрос специалисту(рис.3)

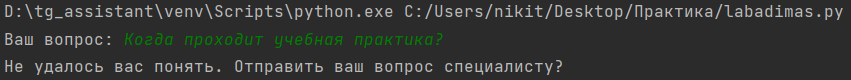


Рис.3. Результат работы.

Пока что не реализована идея переправки вопроса специалисту, потому что для удобства использования вопросы должны передаваться через систему «NAUMEN», на которой работает сервис help.tpu.ru. Для интеграции с этой системой нужно изучать API, которое ещё не было изучено. Но дальнейший алгоритм работы в этом случае таков: после отправки вопроса специалисту бот должен записать обработанный алгоритмом вопрос в таблицу, а после ответа от специалиста зафиксировать ответ на вопрос.

### ЗАКЛЮЧЕНИЕ

За текущий семестр была изучена технология обработки естественного языка(NLP), изучены инструменты для работы с NLP, разработан алгоритм обработки вопросов от пользователей с последующим ответом на них.

Этот алгоритм является полезным функционалом в чат-боте, т.к сможет помочь автоматизировать работу специалистов сервиса help.tpu.ru и получить незамедлительный ответ на свой вопрос для студентов.

Алгоритм был проверен на 50 вопросах, 41 из них были корректно распознаны. Проблемы возникают на вопросах, где ключевые слова одинаковые.

На данный момент алгоритм не внедрен в чат-бота, не интегрирован c системой help.tpu.ru, не доведено до идеала распознавания вопросов. В планах исправить эти проблемы и расширить функционал путем добавления обучения, чтобы бот мог ответить на вопрос, собрав из нескольких ответов один.

### СПИСОК ИСТОЧНИКОВ

1) Обработка естественного языка [Электронный ресурс]. - URL: https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0\_%D0%B5%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE\_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0 (Дата обращения: 12.01.2023).

2) Юлий Васильев Обработка естественного языка. Python и spaCy на практике.: СПб.: Питер СПБ, 2021. - 256 с.

3) Facts & Figures [Электронный ресурс]. - URL: https://spacy.io/usage/facts-figures#comparison-features (Дата обращения: 12.01.2023).

4) 4 проблемы NLP [Электронный ресурс]. - URL: https://robotdreams.cc/blog/87-chto-meshaet-razvitiyu-nlp (Дата обращения: 12.01.2023).

5) Институт электронного обучения - Часто задаваемые вопросы (FAQ) [Электронный ресурс]. - URL: https://portal.tpu.ru/ido-tpu/chasto\_zadavaemie\_voprosi (Дата обращения: 12.01.2023).

6) NER для русского языка в Spacy 3: удобно и легко [Электронный ресурс]. - URL: https://newtechaudit.ru/ner-dlya-russkogo-yazyka-v-spacy-3-udobno-i-legko/ (Дата обращения: 12.01.2023).

7) SpaCy - NLPub [Электронный ресурс]. - URL: https://nlpub.ru/SpaCy (Дата обращения: 12.01.2023).

8) Делаем автореферат со Spacy и экстрактивной суммаризацией [Электронный ресурс]. - URL: https://newtechaudit.ru/delaem-avtoreferat-so-spacy-i-ekstraktivnoj-summarizacziej/ (Дата обращения: 12.01.2023).

9) Суммаризация текста: подходы, алгоритмы, рекомендации и перспективы [Электронный ресурс]. - URL: https://habr.com/ru/post/514540/ (Дата обращения: 12.01.2023).

10) fuzzywuzzy 0.18.0 [Электронный ресурс]. - URL: https://pypi.org/project/fuzzywuzzy/ (Дата обращения: 12.01.2023).

11) Применение библиотеки FuzzyWuzzy для нечёткого сравнения в Python. Расстояние Левенштейна (редакционное расстояние) [Электронный ресурс]. - URL: https://habr.com/ru/post/491448/ (Дата обращения: 12.01.2023).

12) Задача: извлечь ключевые выражения из текста на русском языке. NLP на Python [Электронный ресурс]. - URL: https://habr.com/ru/post/468141/ (Дата обращения: 12.01.2023).

13) Важность обработки естественного языка для поддержки клиентов с помощью чат-ботов [Электронный ресурс]. - URL: https://habr.com/ru/company/otus/blog/567924/ (Дата обращения: 12.01.2023).

14) Machine Learning и не только: как устроены чат-боты [Электронный ресурс]. - URL: https://www.bigdataschool.ru/blog/how-chat-bot-is-made.html (Дата обращения: 12.01.2023).

15) Бот на нейросетках: как работает и учится виртуальный ассистент [Электронный ресурс]. - URL: https://habr.com/ru/post/478572/ (Дата обращения: 12.01.2023).

16) Краткий обзор NLP библиотеки SpaСy [Электронный ресурс]. - URL: https://habr.com/ru/post/504680/ (Дата обращения: 12.01.2023).

17) Advanced NLP with spaCy · A free online course [Электронный ресурс]. - URL: https://course.spacy.io/en/ (Дата обращения: 12.01.2023).

18) Russian language models for spaCy [Электронный ресурс]. - URL: https://github.com/buriy/spacy-ru/ (Дата обращения: 12.01.2023).