Нижегородский государственный технический университет
им. Р. Е. Алексеева
Институт радиоэлектроники и информационных технологий
Кафедра «Вычислительные системы и технологии»

Выпускная квалификационная работа

ПРОГРАММНАЯ СИСТЕМА С ЕСТЕСТВЕННО-ЯЗЫКОВЫМ ИНТЕРФЕЙСОМ

Студент: Баринов Р.О. 16-В-2
Научный руководитель: к.т.н., доцент Гай В.Е.

Нижний Новгород
2020

Цель и задачи работы

➢ Цель:

- Разработка программной системы с естественно-языковым интерфейсом для помощи абитуриентам и студентам младших курсов НГТУ им. Р. Е. Алексеева.

➢ Задачи:

1. Обзор существующих методов реализации систем с естественно-языковым интерфейсом;
2. Составить базу данных со всей необходимой справочной информацией;
3. Разработать программное обеспечение для классификации запроса пользователя, поиска необходимой справочной информации и генерации ответа пользователю;
4. Проведение эксперимента для подтверждения корректности работы созданной системы.

Методы реализации систем с естественно-языковым интерфейсом

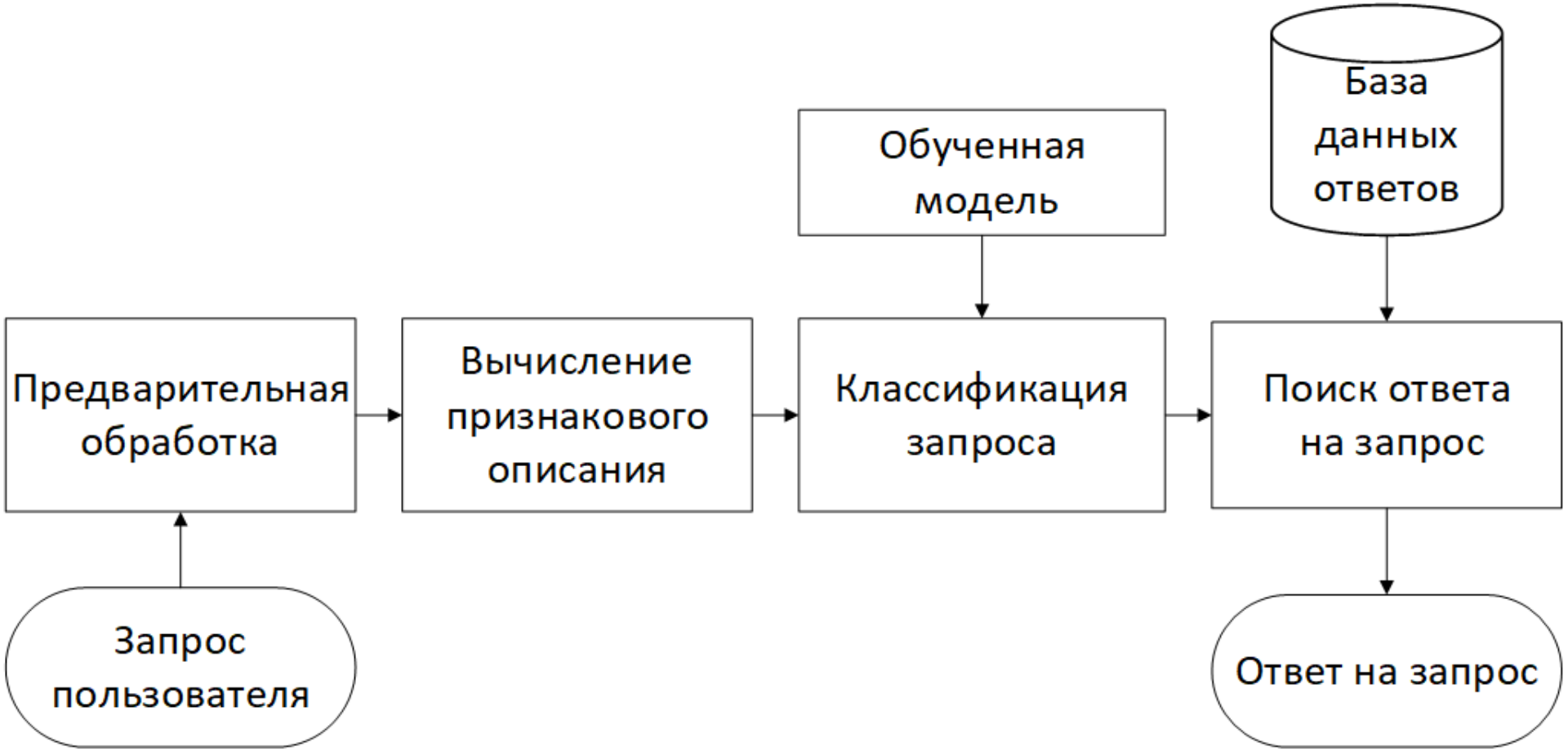
- **Генеративная модель:**
  - Может генерировать произвольный ответ;
  - Склонна к несоответствию ответа и поставленного вопроса;
  - Может генерировать ответ в ошибочной грамматической и синтаксической форме;
  - Требуется большое число данных для обучения модели;

- **Выборочная (поисковая) модель:**
  - Ограниченный набор заготовленных ответов;
  - Всегда генерирует ответ в правильной (заданной) грамматической и синтаксической форме;
  - Не требуется большое число данных для обучения модели;

Примеры запросов:

- «Врач 6 корпус»
- «Где найти врача в 6 корпусе»

Структурная схема системы с естественно-языковым интерфейсом

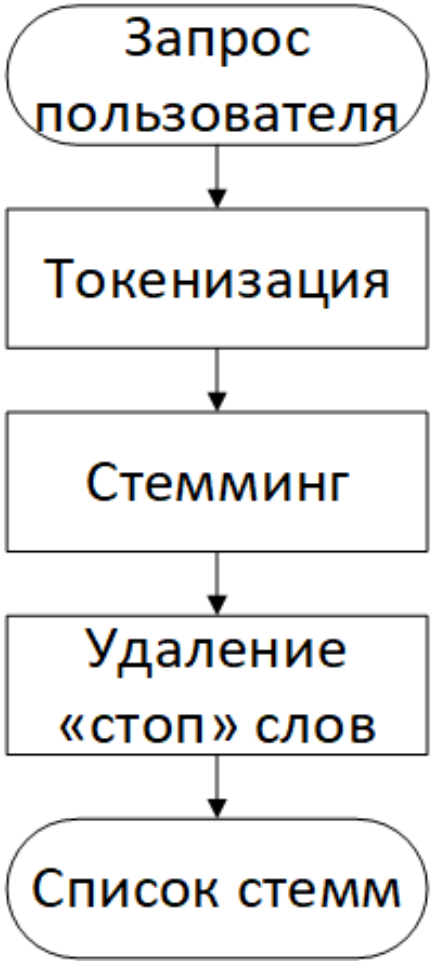


Формирование базы данных ответов

- Было определено 15 различных тем (классов):
  - Банкоматы, библиотеки, второй отдел, СОЛ «Ждановец», история НГТУ, медицинская служба, общежития, парковки, питание, профком, расписание, РСМ НГТУ, структура НГТУ, студенческий клуб, транспорт.
- Каждая тема включает в себя различное число ответов на вопросы.
  - Пример из класса «Медицинская служба»: Медицинский кабинет в 6 корпусе: Врач: Гущина Галина Ивановна, телефон: 257-86-64, время работы: с 8:30 до 15:00, кроме субботы и воскресенья.
- В качестве формата хранения данных выбран формат «JSON».

Предварительная обработка пользовательского запроса

- **Токенизация** (сегментация) – это процесс разделения однородной структуры (текста или отдельных предложений), написанной на естественном языке, на более мелкие структуры – предложения или слова.
- **Стемминг** – это эвристический процесс, который убирает некоторые словообразующие морфемы от корня слова.
- **«Стоп» слова** - это частицы, предлоги, союзы и т.д., то есть слова, которые не несут смысловой нагрузки.



Признаковое описание.

Bag of Words (Мешок слов)

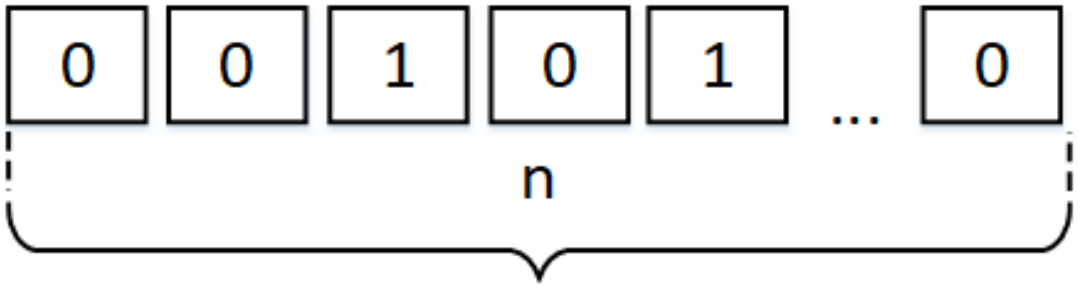
Словарь	здорово	машинное	обучение	очень	это
Машинное	0	1	0	0	0
обучение	0	0	1	0	0
это	0	0	0	0	1
здорово	1	0	0	0	0
Итоговый вектор для предложения	1	1	1	0	1

Принятие решения

- На вход классификатору (нейронной сети) подается векторизованный пользовательский запрос.

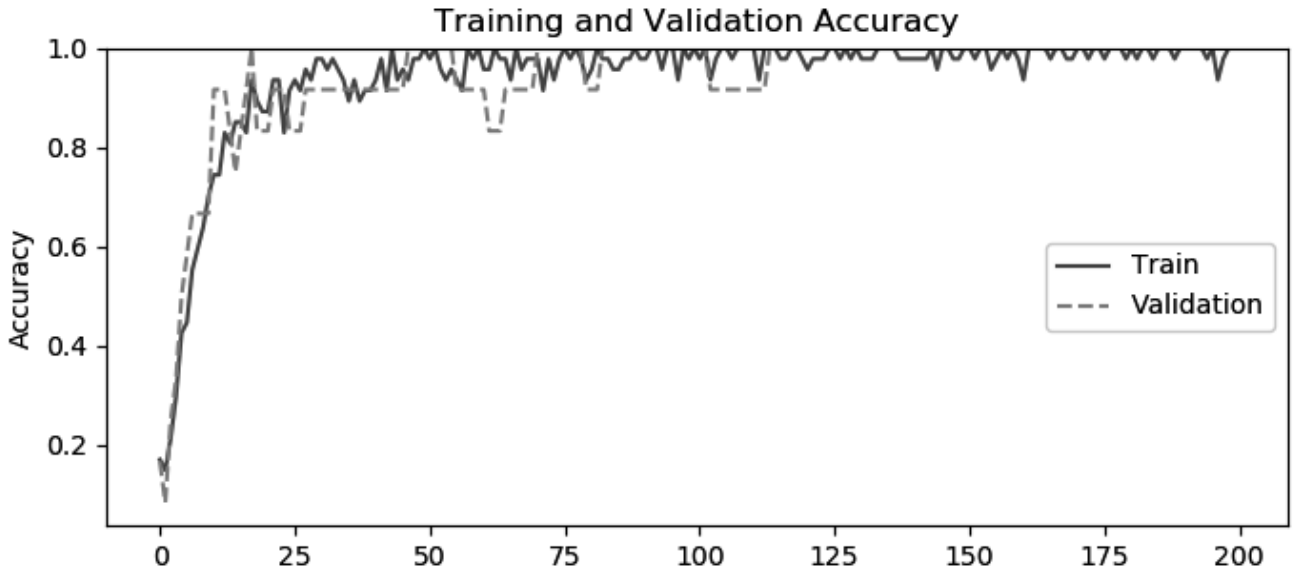
Пример запроса: «где находится врач в 6 корпусе?»

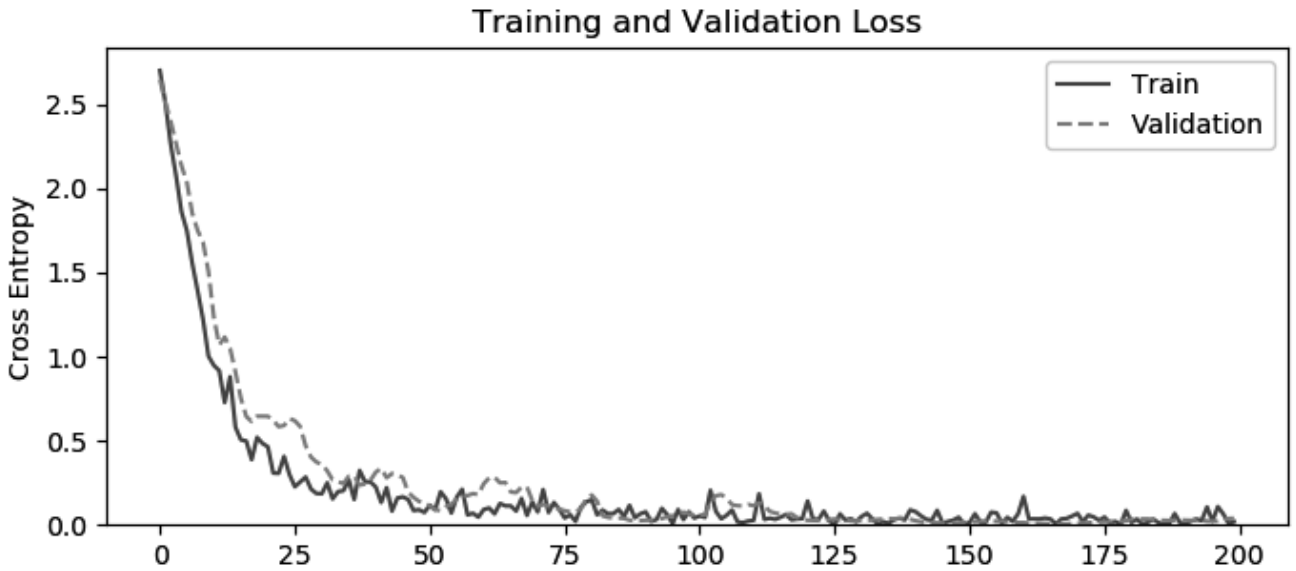
Пример векторизованного запроса, где *n* – размер словаря:

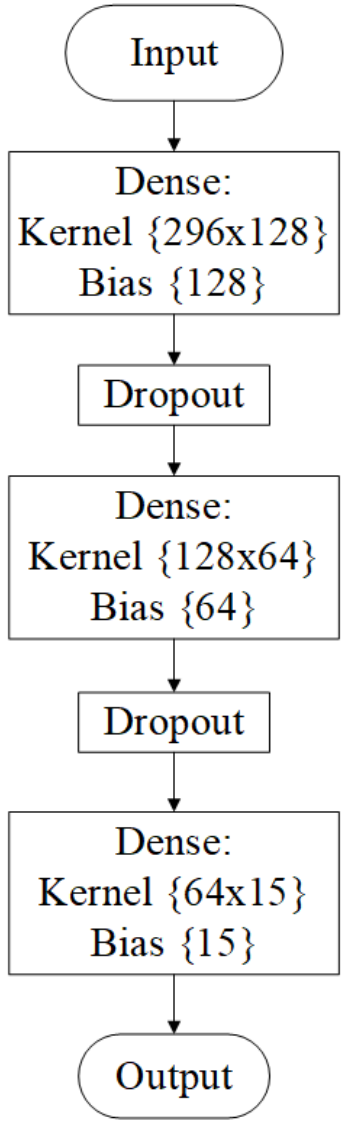


- Классификатор вычисляет вероятность принадлежности признакового описания пользовательского запроса к каждому из 15 классов.
- Решение принимается на основе вероятности принадлежности признакового описания к одному из классов.

Структура нейронной сети и результаты обучения модели







Поиск конечного ответа

- Алгоритм поиска ответа на запрос состоит из четырёх этапов:

1. Векторизация пользовательского запроса с учётом весов слов;
2. Вычисление сходства между векторами (косинусного коэффициента);
3. Выбор ответа из БД по косинусному коэффициенту;
4. Интерпретация вектора в ответ на естественном языке.

Векторизация запроса с учётом весов слов

1. $$tf(t,d) = \frac{n_t}{N}$$

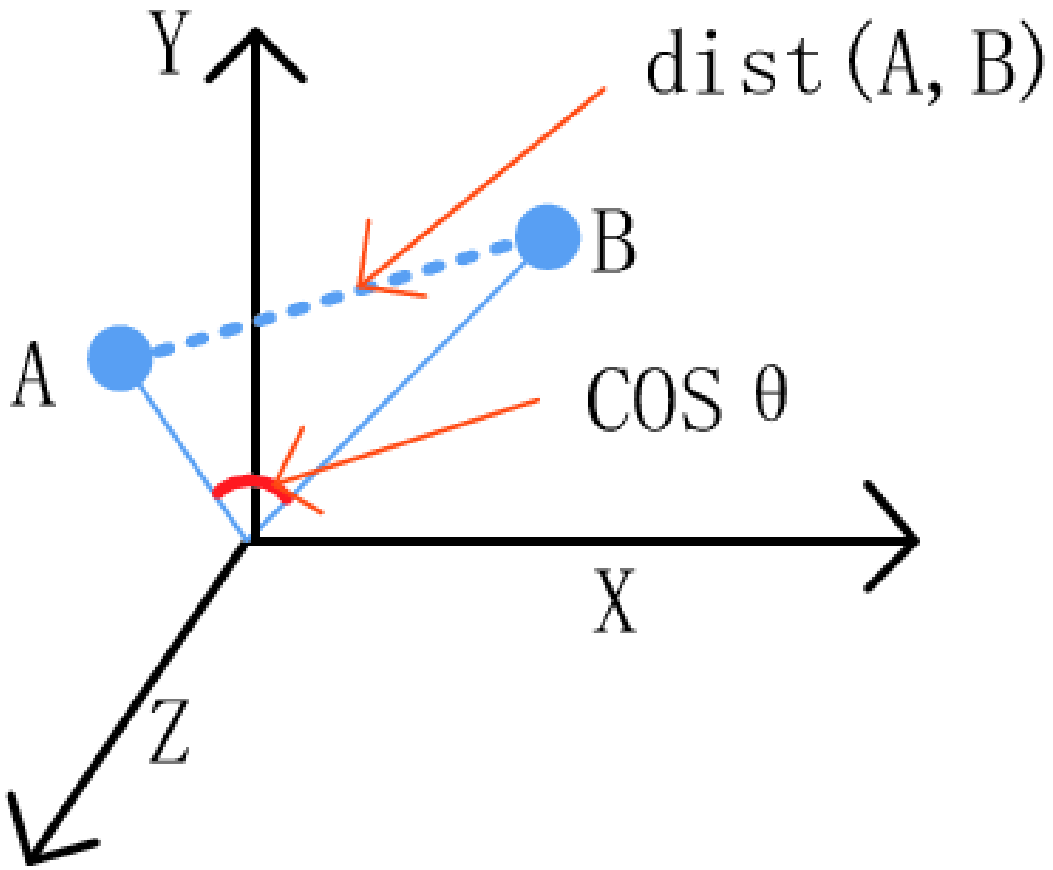
где
t – слово, для которого считается коэффициент;
d – текущий документ;
n_t – количество слов *t* в документе *d*;
N – общее количество слов в документе *d*.

2. $$idf(t,D) = log\frac{D}{d_t}$$

где
D – количество документов в наборе;
d_t – число документов из набора *D*, в которых присутствует слово *t*.

3. $$tfidf(t,d,D) = tf(t,d) * idf(t,D)$$

Вычисление сходства между векторами



Запрос: «контакты кафедры ВСТ»
A – вектор пользовательского запроса
A = (0, …, 0.42, …, 0.36, …, 0.59, 0, 0)

B – вектор из БД ответов
B = (0, …, 0.36, …, 0.59, …, 0.62, …, 0.42, …, 0)

Косинусный коэффициент: 0.41

Ответ: «кафедра: вычислительных систем и технологий (вст): контакты: vt@nntu.ru»

$$cos(\theta) = \frac{A * B}{|A| * |B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Интеграция системы с естественно-языковым интерфейсом с ПО «Telegram Messenger»

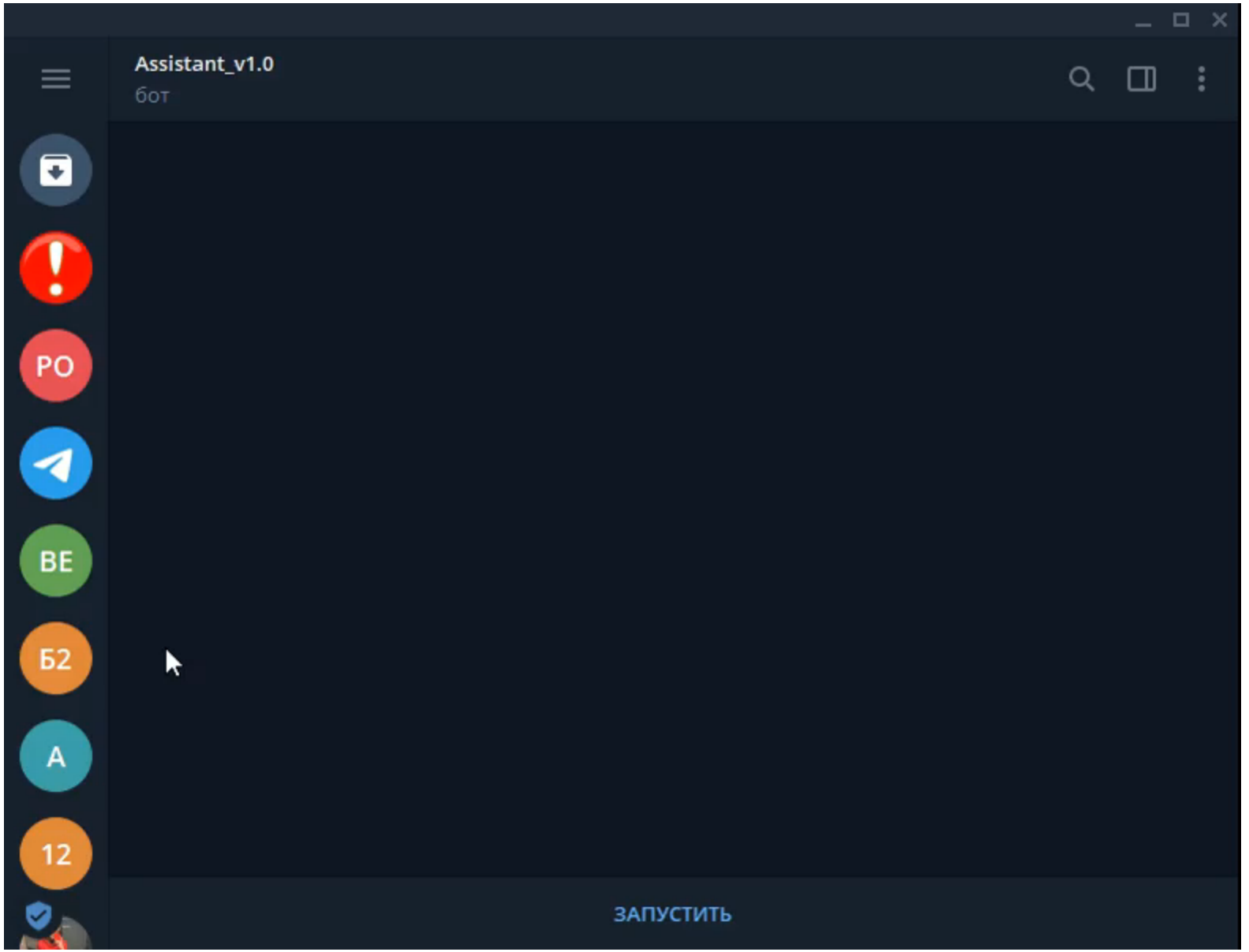
- При создании Telegram-бота выделяется уникальный идентификатор.
- Стандартная функция обработки декарируется разработанной системой.
- Для запуска программной системы необходим интерпретатор Python версии 3.6.x с установленными библиотеками NumPy, Pandas, NLTK, Keras, Sklearn, Telebot, а также ПО «Telegram Messenger».

Тестирование программной системы

- В тестировании системы принимали участие студенты 2-3 курсов НГТУ им. Р. Е. Алексеева.
- Было обработано 900 пользовательских запросов.
- Результаты тестирования:
  - 772 (85,8 %) запроса были верно классифицированы из них:
    - На 721 – дан верный ответ

    (93,4 % - относительно верно классифицированных запросов и 80,1% - относительно всех);
    - На 51 – ошибочный ответ

    (6,6 % - относительно верно классифицированных запросов и 5,7% - относительно всех);
  - 128 (14,2 %)запросов были классифицированы ошибочно.


Assistant_v1.0
бот
РО
ВЕ
Б2
А
12
ЗАПУСТИТЬ

Итоги

- Разработана и протестирована программная система с естественно-языковым интерфейсом для помощи абитуриентам и студентам младших курсов НГТУ им. Р. Е. Алексеева.

- В ходе разработки системы были решены поставленные задачи, включающие в себя разработку ПО для классификации пользовательских запросов и поиска конечного ответа.

Перспективы развития системы

- Создание подсистемы обработки голосовых запросов;
- Создание подсистемы дообучения модели нейронной сети;
- Внедрение многопоточной обработки пользовательских запросов.

Публикация

Р.О.Баринов, В.Е.Гай Программная система с естественно-языковым интерфейсом// Материалы XXVI международной научно - технической конференции «Информационные системы и технологии - 2020», ИСТ - 2020, Россия, Н. Новгород, 2020г.

Спасибо за внимание!

Нижегородский государственный технический университет
им. Р. Е. Алексеева
Институт радиоэлектроники и информационных технологий
Кафедра «Вычислительные системы и технологии»

ПРОГРАММНАЯ СИСТЕМА С ЕСТЕСТВЕННО-ЯЗЫКОВЫМ ИНТЕРФЕЙСОМ

Студент: Баринов Р.О. 16-В-2
Научный руководитель: к.т.н., доцент Гай В.Е.

Нижний Новгород
2020