

# Перечень проектов по направлению «Большие данные, искусственный интеллект, финансовые технологии и машинное обучение»

## Тема проекта № 1

### Тренажер для подготовки к ЕГЭ по русскому языку

Всем вам скоро предстоит такой этап в жизни, как написание экзаменов. К сожалению, не всегда достаточно обладать хорошими знаниями по математике, информатике и физике, чтобы поступить в технический вуз. Больше всего проблем как раз возникает с экзаменом по русскому языку, и нужно понять, как к нему можно подготовиться.

**Цель проекта:** создать тренажер для подготовки к егэ по русскому языку, который по вашим орфографическим ошибкам будет искать задания для отработки данного типа ошибок.

#### Основные задачи:

- 1) Ознакомиться с существующими (алгоритмическими похожие на то, что мы планируем разработать) тренажерами, например, «Решуегэ»
- 2) Найти какое-то количество текстов с пропущенными буквами и их правильные варианты, на которых как раз и будут тренироваться пользователи.
- 3) Собрать базу данных из слов со следующими свойствами:
  - данные должны содержать различные части речи,
  - слова должны подходить для отработки различных правил (безударные гласные, словарные слова, правила в приставках и суффиксах).

Для набора базы данных слов можно воспользоваться данными с портала «Решуегэ» или открытым банком заданий.

- 1) Обучить модель для получения хороших эмбедингов (векторные представления) или воспользоваться уже обученной. Нам это понадобится, так как мы хотим по слову определять его часть речи, определять в какой части слова была допущена ошибка, чтобы сгенерировать аналогичные задания.
- 2) Понять, как мы хотим сравнивать эмбединги слов, чтобы распределять их по группам.
- 3) Создать бота или приложение для тренажера

#### Материалы для подготовки:

- 1) <https://habr.com/ru/company/ods/blog/329410/> - Почитать про эмбединги.
- 2) <https://rus-ege.sdangia.ru/> - для сбора текстов.
- 3) <https://fipi.ru/ege/otkrytyy-bank-zadaniy-ege> - открытый бланк заданий.

## Тема проекта № 2

### Wi-Fi Sensing

В последние годы в научном сообществе набирает популярность Wi-Fi sensing - использование сигнала, который передает Wi-Fi для определения действий, происходящих в окружающей среде. Например, определять, что в квартире упал человек, и, если он пожилой, отправлять сигнал родственникам. Подобные чудеса стали возможны благодаря машинному обучению и математикам, которые нашли ему применение в столь нестандартной области. Вам предлагается пройти путем такого исследователя и самостоятельно решить одну из задач Wi-Fi sensing-a.

**Цель проекта:** создание алгоритма для идентификации действий на основе изменения Wi-Fi сигнала.

#### Основные задачи:

- 1) Ознакомиться с принципами передачи сигнала Wi-Fi и статьями, посвященными сенсингу
- 2) Найти датасеты, подходящие для задач Wi-Fi сенсинга
- 3) На основе найденных датасетов выделить задачу
- 4) Разработать алгоритм, решающий данную задачу

#### Материалы для подготовки:

- 1) [https://www.researchgate.net/publication/305722198\\_A\\_survey\\_on\\_WiFi\\_Channel\\_State\\_Information\\_CSI\\_utilization\\_in\\_Human\\_Activity\\_Recognition](https://www.researchgate.net/publication/305722198_A_survey_on_WiFi_Channel_State_Information_CSI_utilization_in_Human_Activity_Recognition) - хорошая обзорная статья, объясняется что такое CSI (с которыми вам и придется работать) и как к ним прикрутить распознавание сценариев
- 2) [https://www.davidrojas.co.uk/wp-content/uploads/2017/10/falldefi\\_ubiquitous\\_fall\\_detection\\_wifi.pdf](https://www.davidrojas.co.uk/wp-content/uploads/2017/10/falldefi_ubiquitous_fall_detection_wifi.pdf) - пример алгоритма для Wi-Fi сенсинга
- 3) [https://www.researchgate.net/publication/336449005\\_Wiar\\_A\\_Public\\_Dataset\\_for\\_Wifi-Based\\_Activity\\_Recognition](https://www.researchgate.net/publication/336449005_Wiar_A_Public_Dataset_for_Wifi-Based_Activity_Recognition) - пример датасета, который можно использовать

## Тема проекта № 3

### Чат-бот для малого и среднего бизнеса

Сейчас сложно представить хоть какой-то бизнес без контакта в мессенджерах. Однако предпринимателям сложно постоянно быть онлайн и отвечать потенциальным клиентам. Чтобы упростить этот процесс, участникам предстоит разработать платформу для автоматизации таких переписок.

**Цель проекта:** Создать платформу (чат-бота в Telegram), которая будет автоматизировать рутинные переписки с клиентами.

#### Основные задачи:

- 1) Найти предобученную модель или обучить текстовую модель или алгоритм, понимающий русский язык. Алгоритм (модель) должен уметь получать хорошие эмбединги (векторные представления) текста.
- 2) Научить модель понимать к какому из нескольких предзаданных текстов относится другой, поданный на вход текст. Это делается так. Берутся  $N$  эмбедингов заготовленных текстов, и эмбединг поданного на вход. Сравниваются попарно по расстоянию (можно поиграться с различными метриками близости). Получаем  $N$  расстояний. Самым похожим (близким по смыслу) считается текст, наиболее близкий по векторному расстоянию.
- 3) Подготовить Telegram-бота. В нем должна быть возможность создавать и редактировать пользовательские сценарии из самого мессенджера (например, это будет полезно для владельцев бизнеса).
- 4) Прикрутить модель к боту. Модель должна обрабатывать первое сообщение от потенциальных клиентов, понимать к какому сценарию оно относится и запускать его (сценарий).
- 5) Развернуть проект на сервере или упаковать в Docker контейнер.

#### Материалы для подготовки:

- 1) <https://huggingface.co/> - огромный репозиторий с предобученными моделями
- 2) <https://onnxruntime.ai/> - технология для конвертации моделей в удобный для прода формат
- 3) <https://docs.aiogram.dev/> - библиотека Python для быстрой разработки telegram ботов
- 4) <https://docs.aiogram.dev/en/latest/examples/index.html> - примеры простых ботов с использованием aiogram
- 5) <https://www.heroku.com/> - платформа для разворачивания сервисов в облаке
- 6) <https://www.docker.com/> - решение для удобной упаковки сервиса в