

# Категоризация новостей с помощью чат-бота в Telegram

## Состав команды:

Афанасьев Денис,

Боттаева Амина,

Гусева Софья,

Склезнёва Ксения

# Постановка задачи

- Разработать Telegram-бот, в который пользователь загружает текст новости, а на выходе получает категорию (тему) загруженной новости
- **Дальнейшие перспективы:**  
Telegram-бот выдает новости за указанный период по одной конкретной теме
- **Используемые данные:** датасет, составленный из новостей с сайта Lenta.ru за 2020 год

# План работы над проектом

- ✓ Сбор данных
- Поиск подходящих моделей
- Обучение моделей
- Тестирование и усовершенствование моделей
- 5. Создание телеграм-бота
- 6. Тестирование телеграм-бота

# Обзор решений задачи

## 1. Предобработка текста:

- Приведение текста к нижнему регистру
- Удаление специальных символов и цифр
- Токенизация
- Удаление стоп-слов
- Приведение к нормальной форме
- Объединение токенов обратно в строку

## 2. Обучаемые модели:

- CNN
- RNN

# Модель CNN (сверточные нейронные сети)

- Принцип работы: состоит из нескольких типов слоев, каждый из которых выполняет свою функцию:
  1. Сверточные слои (Convolutional Layers)
  2. Слои подвыборки (Pooling Layers)
  3. Полносвязные слои (Fully Connected Layers)
  4. Активационные функции
- Преимущества:
  - Автоматическое извлечение признаков
  - Параметрическая эффективность
  - Устойчивость к смещениям

# Модель RNN (рекуррентные нейронные сети)

- Принцип работы: RNN имеют внутреннее состояние (или память), которое позволяет им сохранять информацию о предыдущих входах. Это достигается за счет рекуррентных связей, которые позволяют передавать информацию от одного шага времени к следующему.
- Улучшения RNN:
  - LSTM
  - GRU

# Сравнение RNN и CNN

- Сравнение точностей обучения моделей на тренировочной, валидационной и тестовой выборках

	train	validation	test
RNN	0.9421	0.8128	0.8164
CNN	0.9659	0.8267	0.8354

# Роли участников команды

- Амина: подготовка данных, обучение CNN и RNN
- Ксюша: разработка telegram-бота
- Соня: поиск и обучение еще одной модели, организационная деятельность
- Денис: MIFlow, подготовка тестовых запросов