

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
- 
- "economy": 0,
 - "sports": 1,
 - "society": 2,
 - "life": 3,
 - "entertainment": 4,
 - "technology": 5,
 - "science": 6,
 - "russia": 7,
 - "history": 8

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

- RNN
- CNN
- GRU
- BiLSTM
- XgBoost
- SVM

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

XgBoost

- Алгоритм градиентного бустинга, который объединяет несколько слабых моделей (обычно деревьев решений) в одну сильную модель.

Преимущества XGBoost:

- Эффективность
- Гибкость
- Обработка пропущенных значений
- Интерпретируемость

SVM

- Мощный алгоритм машинного обучения, используемый для задач классификации и регрессии. Он работает на основе концепции нахождения гиперплоскости, которая разделяет данные на классы.

Преимущества:

- Эффективность в высоких измерениях
- Устойчивость к переобучению
- Гибкость
- Отсутствие предположений о распределении данных
- Хорошая производительность на сложных задачах

Сравнение точностей обучения

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

	RNN	CNN	GRU	BiLSTM	XgBoost	SVM
test	0.8164	0.8354	0.8355	0.8161	0.8649	0.8801

Роли участников команды

- **Амина:** Подготовка данных. Обучение CNN и RNN
- **Ксюша:** Обучение GRU, BiLSTM, XgBoost, SVM. Разработка telegram-бота
- **Соня:** Организационная деятельность. Поиск и обучение еще одной модели. Разработка telegram-бота
- **Денис:** MIFlow. Подготовка тестовых запросов