проект инновационного практикума

Оцифровка и извлечение данных из диаграмм

Над проектом работали:
Ярослав Серов
Зырянов Илья
Артем Демидов
Иванин Никита

Менторы проекта:
Марат Хабибуллин(ABBYY)
София Сергиенко(ABBYY)
Виктория Буравкина(ABBYY)

Предыстория проекта

- Идея проекта
- Исследование актуальности
- Постановка цели

Постановка задачи

- Сужение поставленной задачи на столбчатые диаграммы

- Создание консольного приложения, получающего изображение диаграммы и возвращающего Excel-таблицу с данными.

Разбиение на подзадачи

В ходе разбора изначальных задач были выделены главные подзадачи:

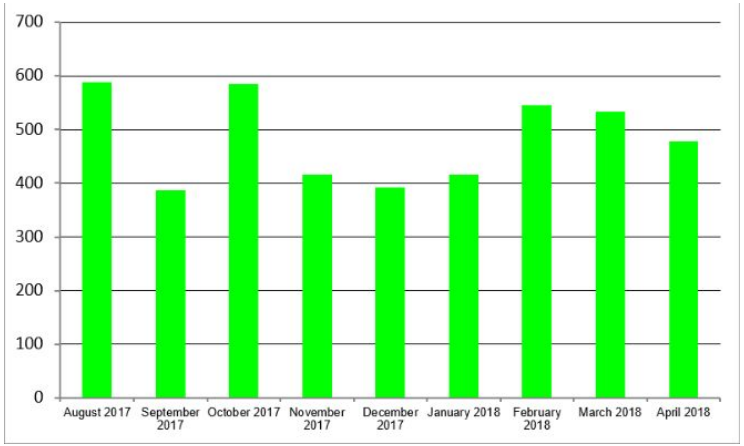
- Реализовать алгоритм распознавания столбцов
- Реализовать распознавание осей и подписей к ним
- Реализовать генератор случайных таблиц и диаграмм для создания синтетической базы диаграмм для тестирования приложения

Использованные технологии

- Методы машинного обучения и нейронные сети для детекции столбцов
- ABBYY Online OCR для распознавания текста
- Библиотека OpenCV для работы с изображениями
- Библиотеки XlsxWriter и win32 для преобразования данных в таблицы и получения изображений

Генератор случайных диаграмм

August 2017	589
September 2017	387
October 2017	585
November 2017	417
December 2017	392
January 2018	416
February 2018	545
March 2018	533
April 2018	478



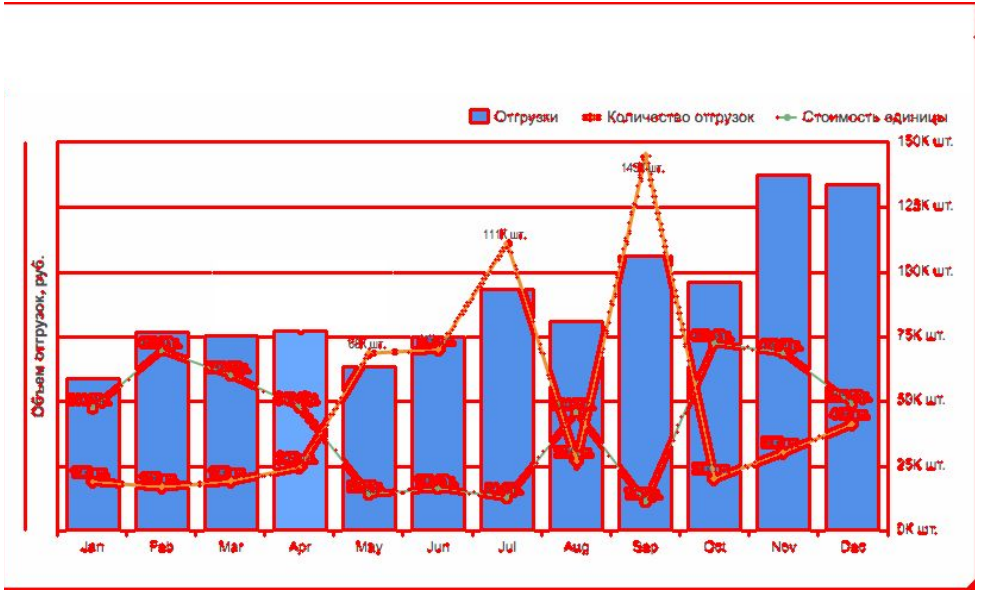
- Исследование характерных параметров диаграмм
- Генерация:
  - Генерация исходных данных
  - Создание синтетической диаграммы

Распознавание столбцов

- Кластеризация цветов на картинке
- Аппроксимация контуров областей прямоугольниками
- Отбор столбцов из найденных контуров на основе геометрических параметров

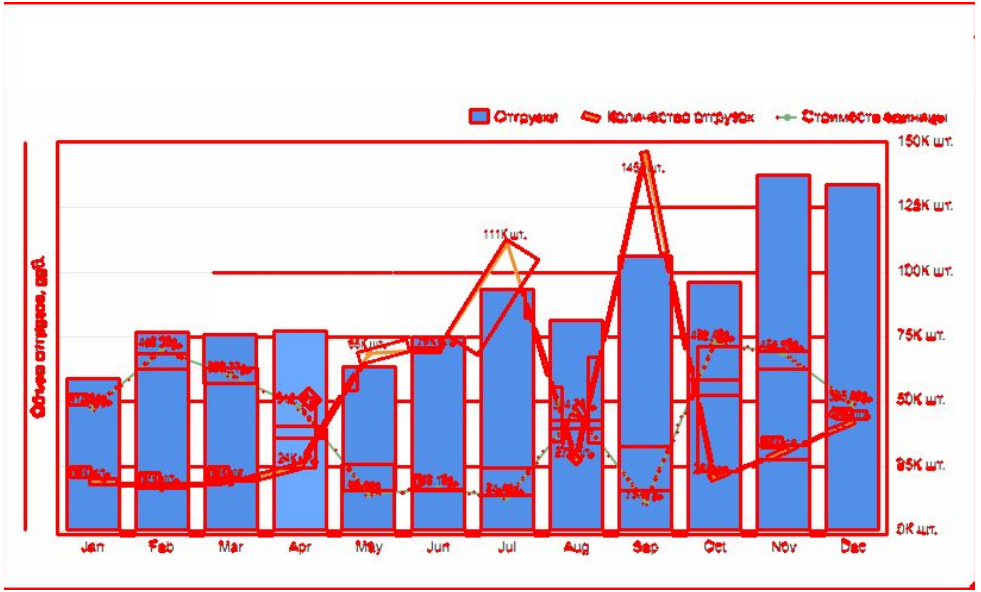
Распознавание столбцов

- Кластеризация цветов на картинке
- Аппроксимация контуров областей прямоугольниками
- Отбор столбцов из найденных контуров на основе геометрических параметров



Распознавание столбцов

- Кластеризация цветов на картинке
- Аппроксимация контуров областей прямоугольниками
- Отбор столбцов из найденных контуров на основе геометрических параметров



Распознавание столбцов

- Кластеризация цветов на картинке
- Аппроксимация контуров областей прямоугольниками
- Отбор столбцов из найденных контуров на основе геометрических параметров

Распознавание осей и подписей к ним

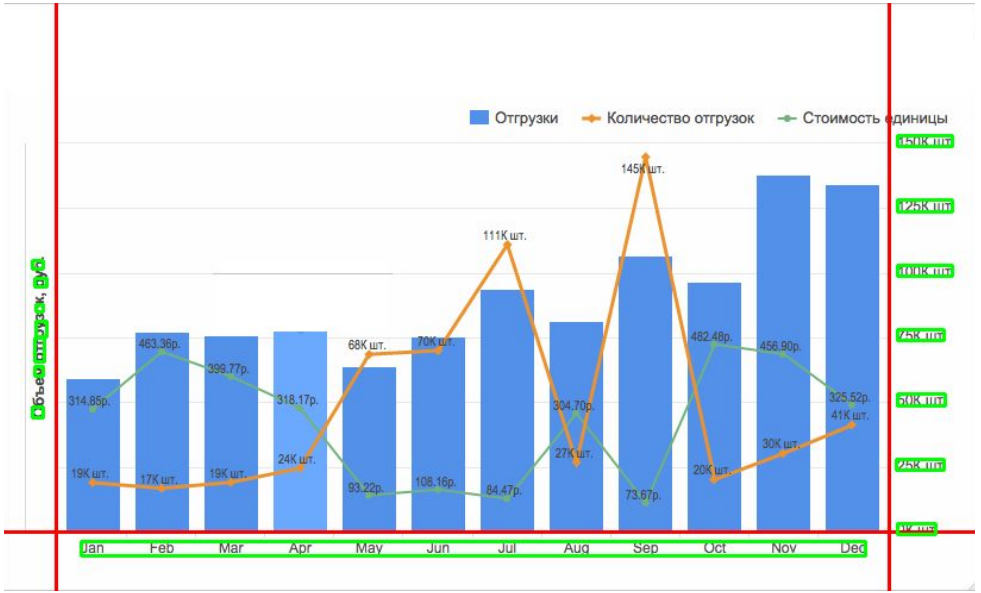
- Поиск осей при помощи преобразования Хафа
- Распознавание текста в областях, определяемых положением осей
- Выбор осей из кандидатов на основании распознанных подписей

Распознавание осей и подписей к ним

- Поиск осей при помощи преобразования Хафа
- Распознавание текста в областях, определяемых положением осей
- Выбор осей из кандидатов на основании распознанных подписей

Распознавание осей и подписей к ним

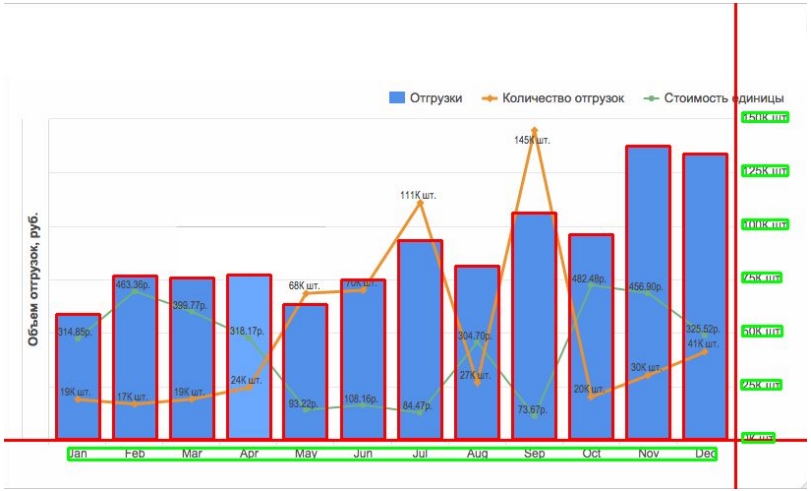
- Поиск осей при помощи преобразования Хафа
- Распознавание текста в областях, определяемых положением осей
- Выбор осей из кандидатов на основании распознанных подписей



Распознавание осей и подписей к ним

- Поиск осей при помощи преобразования Хафа
- Распознавание текста в областях, определяемых положением осей
- Выбор осей из кандидатов на основании распознанных подписей

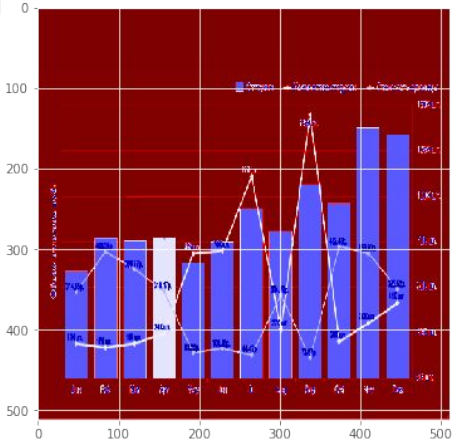
Извлечение данных из диаграмм

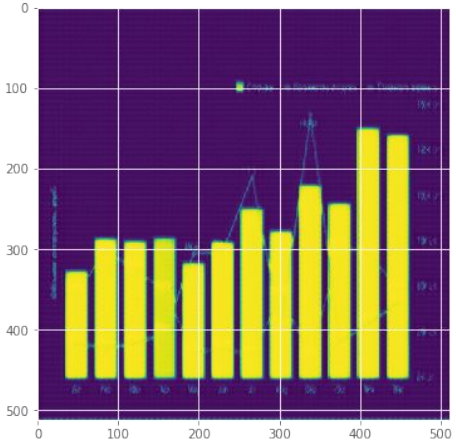


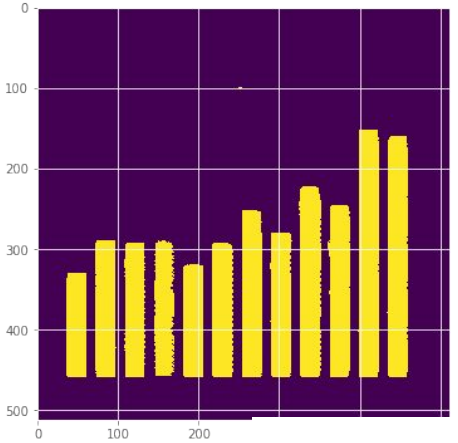
Jan	58k
Feb	75k
Mar	75k
Apr	76k
May	62k
Jun	74k
Jul	92k
Aug	80k
Sep	105k
Oct	95k
Nov	136k
Dec	132k

Используя результаты, полученные при детекции столбцов и распознавании осей и подписей может быть получена таблица с данными

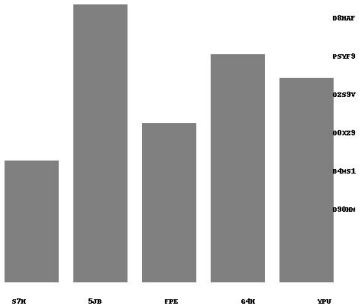
Дополнительный способ поиска столбцов на диаграмме с помощью нейронной сети







- Архитектура - U-NET
- Генерация картинок, похожих на диаграммы, как обучающей выборки, из-за отсутствия размеченных данных
- Неплохое качество на маленьком количестве эпох и маленькой обучающей выборке



Возможности дальнейшего развития

- Улучшение функциональности
  - Повышение точности распознавания диаграмм
  - Расширение приложения для распознавания других типов диаграмм

- Применение полученных результатов
  - Интеграция в существующие продукты ABBYY
  - Коммерческая реализация

Демонстрация работы приложения

Спасибо за внимание!