Сбор данных

Введение

Все данные мы взяли из открытых источников из интернета. Сам сбор данных был осуществлен с помощью языка программирования Python и дополнительных библиотек (Конкретно для сбора данных это selenium).

Выбор сайта

Изначально мы планировали использовать market.yandex.ru в качестве основы для сбора данных, но в ходе реализации мы понял, на нём очень трудно обойти капчу. В связи с этим мы решили взять данные с ozon.ru. Данные все были успешно собраны, но мы поняли, что их крайне мало и они очень неинформативные. В итоге мы приняли решение взять данные с сайт rc-today.ru, так как довольно много дронов в разных ценовых категориях и на странице каждого дрона описано большое кол-во параметров.

Сравнение данных с различных сайтов:

Сравнение будет сделано для дрона DJI Mini 3 Pro

- Rc-today.ru
  Основная табличка с данными [1]:

Подвес:	3-х осевой
Бренд:	DJI
Категория:	Квадрокоптеры
Наличие камеры:	4K, Есть
Тип комплекта:	RTF
Страна производитель:	Китай
Цвет:	Серый
Тип двигателя:	Бесколлекторный
Складной:	Да
Вес до 250 гр:	Да
3D пилотирование :	Нет
Follow Me (Следуй за мной):	Есть
Headless Mode:	Есть
GPS:	Есть
Waypoint (Облет заданных точек):	Есть
FPV (Онлайн трансляция):	Есть
Point of Interest (Точка интереса):	Есть
Возврат домой:	Есть
Удержание высоты:	Есть
Автоматический взлет и посадка:	Есть

  В добавок довольно большое описание как для данного дрона, так и для всех, которые мы изучали.

- market.yandex.ru
  Основная табличка с данными [2]:

Производитель	DJI
Камера	с камерой
Максимальное разрешение видеосъемки	160p
Назначение	профессиональный, детский, любительский, промышленный
Функции	возвращение в точку взлета, возврат одним нажатием, автоматический взлет и посадка,
Количество винтов	4
Поддерживаемые ОС	Android, iOS
Размер	мини
Линейка	Mavic PRO
Управление	управление жестами, Wi-Fi, Bluetooth, радиоканал
Навигационная система	BeiDou, GPS, Galileo
Датчики	датчик визуального позиционирования, инфракрасный датчик, гироскоп, акселерометр

Характеристики

Тип мультикоптера	квадрокоптер (4 винта)
Максимальное время полета	34 мин.
Максимальная высота полета	4000 м
Максимальная скорость набора высоты	5 м/с
Максимальная скорость снижения	5 м/с
Максимальная скорость полета	16 м/с
Максимальный подъемный вес	0.25 кг
Двигатель	бесколлекторный электродвигатель
Функции	Active Track, Headless Mode, Point of Interest, автоматические флипы, автоматический взлет и посадка, возврат одним нажатием, возвращение в точку взлета
Датчики	акселерометр, гироскоп, датчик визуального позиционирования, инфракрасный датчик

Управление полетом

Управление	Bluetooth, Wi-Fi, радиоканал, управление жестами
Поддерживаемые ОС	Android, iOS
Навигационная система	BeiDou, GPS, Galileo
Тип гироскопа	3-х осевой
Дальность управления по радиоканалу	8000 м
Тип Wi-Fi	g/n/ac
Частота WiFi	2.4 ГГц, 5.8 ГГц
Видеовыход на пульте управления	есть
Особенности	автопилот, пульт управления в комплекте, складная конструкция

Камера

Камера	с камерой
Расположение камеры	встроена в корпус
Разрешение матрицы	48 Мпикс
Угол обзора камеры	82.1 °
Максимальное разрешение видеосъемки	2160p
Разрешение фото по вертикали	6048 пикс.
Разрешение фото по горизонтали	8064 пикс.
Дистанционное управление положением камеры	три оси
Число кадров в секунду при разрешении 4K	60

Питание

Емкость аккумулятора	2453 мА·ч
Напряжение аккумулятора	7.38 В
Время зарядки аккумулятора	64 мин.
Тип питания пульта управления	встроенный аккумулятор

Габариты и комплектация

Размеры (ДхШхВ)	251x362x70 мм
Вес	924 г
Комплектация	квадрокоптер - 1 шт., пульт дистанционного управления - 1 шт., интеллектуальная полетная батарея - 1 шт., Запасные пропеллеры DJI Mini 3 Pro - 1 комплект, защита подвеса - 1 шт., кабель Type-C к Type-C PD - 1 шт., отвертка - 1 шт., винты - 6 шт
Линейка	Mavic PRO

Дополнительно

Срок службы	2 г.
Гарантийный срок	1 г., Гарантия продавца 1 месяц

Данный источник очень хороший в плане информации: большой объем характеристик для каждого отдельно взятого дрона и четкая группировка в соответствующей табличке. Но, к сожалению, на этом сайте подключена система Yandex Smart Captcha, что не позволяет в автоматическом режиме собирать данные.

- www.ozon.ru
  Основная табличка [3]:

Характеристики

Тип	Квадрокоптер	Макс. емкость карты памяти	1 ТБ
Бренд	DJI	Цвет	Светло-серый
Радиус действия, м	8000	Гарантийный срок	30 дней
Питание радиоуправляемой модели	От сети 220В	Максимальный возраст ребенка	До 18 лет
Управление радиоуправляемой моделью	Пульт Д/У	Минимальный возраст ребенка	От 3 лет
Время работы, мин	35	Камера	Встроенная
Питание пульта	Встроенный аккумулятор	Пол ребенка	Унисекс
Количество батареек	1	Дальность полета, м	6000

Специальные возможности: Видеозапись на карту памяти, Видеосъемка, Трансляция на андроид-устройство, Фотосъемка

Особенности радиоуправляемых игрушек: Встроенная камера, Звуковые эффекты, Оптическая стабилизация, Световые эффекты

По сравнению с предыдущими двумя тут очень скудный набор данных, от которых очень трудно получить достоверную математическую модель.

Программная реализация:

Исходный код [4]:

Основной скрипт (Rc_Today.py):

```python
import pickle
import time
import traceback

from Selenium import StartSelenium, ConnectToURL
from Selenium.findElementByXpathToSelenium import search


class RC_Today(object):
    def __init__(self):
        self.characteristic = {}
        self.all_links = []
        # self.all_links = pickle.load(open("all_links_RS_STODAT_v3.pkl", "rb"))
        # self.characteristic = pickle.load(open("all_characteristic_RS_STODAT_v2.pkl", "rb"))

    def start_selenium(self, headless=False, cookies=True):
        self.driver = StartSelenium.create_driver()
        try:
            self.driver.execute_cdp_cmd("Page.addScriptToEvaluateOnNewDocument", {
                'source': '''
                delete window.cdc_adoQpoasnfa76pfcZLmcfl_Array;
                delete window.cdc_adoQpoasnfa76pfcZLmcfl_Promise;
                delete window.cdc_adoQpoasnfa76pfcZLmcfl_Symbol;
            '''
            })
            print("links len: ", len(self.all_links))
            print("charact len: ", len(self.characteristic.keys()))

            self.go_to_main_page()
            self.go_every_link()
        except:
            print(traceback.format_exc())
        finally:
            time.sleep(3)
            pickle.dump(self.all_links, open("E://all_links_RS_STODAT_v3.pkl", "wb"))
            pickle.dump(self.characteristic, open("E://all_characteristic_RS_STODAT_v2.pkl", "wb"))
            self.driver.close()
            self.driver.quit()
```

```python
    def go_to_main_page(self):
        ConnectToURL.connect("https://rc-today.ru/kvadrokoptery/?v[price1]=150000&v[price2]=1666700&p=ALL", self.driver)
        self.find_all_links()

    def find_all_links(self):
        all_titles_on_page = search(self.driver, "//a[@class = 'product_item__name__link']", _list=True)
        print(len(all_titles_on_page))

        for headder in all_titles_on_page:
            self.all_links.append(headder.get_attribute("href"))

    def go_every_link(self):
        for index, link in enumerate(self.all_links):
            if link not in self.characteristic.keys():
                ConnectToURL.connect(link, self.driver)
                try:
                    description = search(self.driver, "//div[@itemprop = 'description']").text
                    price = search(self.driver, "//div[@class = 'product_page__buttons']//span")
                    price = price.text.replace("руб.", "").replace(" ", "").strip()
                    keys = search(self.driver, "//tr/td[1]", _list=True)
                    values = search(self.driver, "//tr/td[2]", _list=True)

                    info_dict = {}
                    for key_index in range(len(keys)):
                        info_dict[keys[key_index].text] = values[key_index].text

                    self.characteristic[link] = [price, info_dict, description]

                    print(index, [price, info_dict, description])
                except:
                    print(traceback.format_exc())
            else:
                print(index, "link already parsed")
```

Вспомогательный скрипт для запуска WebDriver (StartSelenium.py):

```python
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.chrome.options import Options


def create_driver():
    s = Service(ChromeDriverManager().install())
    options = Options()
    options.add_argument('--remote-debugging-port=9222')
    options.add_argument("--disable-dev-shm-usage")
    options.add_argument("--v=99")
    options.add_argument("--no-sandbox")
    options.add_argument(f"--user-data-dir=Cookies")
    options.add_argument(
        "user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36")
    driver = webdriver.Chrome(service=s, chrome_options=options)
    return driver
```

Вспомогательный скрипт для поиска элементов на странице(findElementByXpathToSelenium.py):

```python
from selenium.webdriver.common.by import By

def search(driver, xpath, _list=False, debug=True):
    try:
        if _list == False:
            return driver.find_element(By.XPATH, f"{xpath}")
        else:
            return driver.find_elements(By.XPATH, f"{xpath}")
    except:
        if debug:
            print(xpath)
            print('find error')
        return None

```

В основном классе есть два поля:

1. characteristic – словарь, будет хранить все данные о каждом дроне.
2. all_links – будет хранить все ссылки на дроны.

Метод start_selenium запускает парсинг (сбор данных с сайтов). Создается WebDriver, который открывает обычный браузер Chrome, но под управление скриптов. Далее выполняются два метода: go_to_main_page и go_every_link. Первый позволяет собрать все ссылки на страницы дронов, второй собирает с каждой страницы нужную информацию. Поиск элементов осуществляется с помощью скрипта findElementByXpathToSelenium.py. В случае критической ошибки все собранные данные будут сохранены, и программа завершит работу.

В методе go_to_main_page мы просто открываем одну ссылку и получаем все ссылки.

В методе go_every_link мы проходимся по каждой ссылке, подключаемся к каждой. Затем собираем данные из основной таблички и копируем все описание дрона и получаем его цену. Все эти данные заносим в словарь characteristic. Структура словаря: ключ- ссылка на дрон, значение это список, где первым элементом является цена дрона, вторым - словарь, где представлены данные из основной таблички, а третий элемент — это строка – описание дрона.

Работа с «сырыми» данными:

В скрипте data.py основной функцией является main (). У нас происходит распаковка данных, затем создается список из столбиков для будущей таблицы, после чего циклом проходимся по данным. Каждая итерация отвечает за один дрон. Из описания каждого дрона мы пытаемся получить его длину, ширину, высоту, вес, дальность полета, наличие барометра и акселерометра. Последним шагом для каждой итерации будет занесение данных в DataFrame (библиотека pandas). После того как все данные приведены в нормальный вид и структурированы мы конвертируем DataFrame в Excel таблицу.

Источники:

1. https://rc-today.ru/product/radioupravlyaemii-kvadrokopter-dji-mini-3-pro-pult-rc-rtf-6941565929419/
2. https://market.yandex.ru/product--kvadrokopter-dji-mini-3-pro-dji-rc/1755156947/spec?cpc=bXzoLRBNbhjGRJGOzHEeCjqDONEN6AuZgEo000cnN_26Y8eMEPd3hSW1Ji2LO2nwmjdjbkmcFQjUkpyWQwL9sHWEXgGHH6OqO0Bqhs7_Yaw-SEAuq95B1nZO1C8q3_DW1PjxHbfDjcFS1RvlN4UGidil4XZceWxgDAZ5h8BoszUvQTC63VTtJxASBiPnIBHrQfNOkN70Kw1895elBGIciNR2twKDFRKlnRIYh6snlE6PAbZZx4HqFBnVSLOh3udb&track=char&sku=101765080732&cpa=1&nid=18042097
3. https://www.ozon.ru/product/kvadrokopter-dji-mini-3-pro-rc-740782080/?asb=eHPdxBClVCxJxQACdPRQCYjb5RDBGgvucmreMwpWMJU%253D&asb2=MADlUbkUzPCJ3VbHblFcR04MG7QsigYzN6gD2uRJeqOtwNXrVKa8xTLnWBjupH5M&avtc=1&avte=1&avts=1683984637&keywords=dji+mini+3&sh=84-w8-Uo0g
4. https://github.com/TRAIANUSssS/DroneParser

1. Изначально наш план состоял в том, чтобы взять данные с яндекс маркета, но, к сожалению, в ходе работы я столкнулся с капчей, которую нереально пропустить. Яндекс маркет мы выбрали так как там очень много параметров для каждого дрона и все они структурированы и удобно занесены в табличку.
2. Затем мы решили брать данные с озона. Я написал сборщик данных, но мы опять столкнулись с проблемой. Дронов, у которых достаточное кол-во информации для анализа очень мало.
3. Было принято решение подобрать сайт, который специализируется исключительно на радиоуправляемой технике и выбор пал на rc-today. На этом сайте довольно большое кол-во дронов и для каждого расписаны постоянные параметры. А также у каждого из дронов присутствует довольно большое описание, из которого мы вытащили еще некоторые параметры.
   В добавок к вот этим параметрам мы еще взяли габариты, вес, дальность полета, присутствие барометра и акселерометра