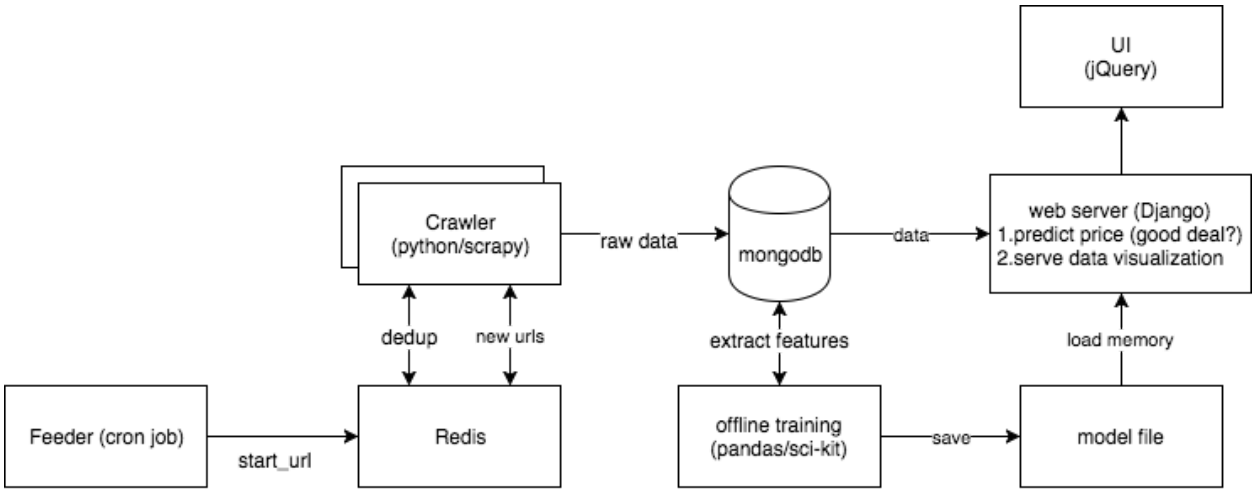
Scenario:

基于Python的网络爬虫，抓取Craigslist上的二手车信息
存储到后端数据库
进行可视化处理
应用机器学习进行价格预测
(Optional) 设计离线检测系统，如果发现Craigslist上的新广告有价格优势，可以实时提示用户有优惠

Application:



- Feeder: cron + python
    - Publish start urls into “sfbay_redis:start_urls” queue every 30 mins
    - Deploy in VM ad docker container
- Crawler: implement two crawlers, use scrapy + redis + mongodb
    - sfbay_redis: (only sf bay area)
        - Run as daemon to crawl used car ad, download and save used car information into mongodb
        - It will fetch start_urls from redis and save pending requests in redis
        - Handle dedup using bloom filter or set in redis
        - Deploy as docker container
        - Use proxy

    - new_cars:
        - Get new car information including year, make, model from www.msn.com
        - Run once as a script and save it into file

- Offline trainer: pandas + scikit-learn
    - Dump mongodb collection into file
    - Clean up duplicate information
    - Based on new car data to extract features from used car data, including price, year, make, model, trim, odometer and city
    - Use linear regression to train the model to predict car for testing data
    - Save the model into file

- Web server + UI: Django + D3
    - Display used car data, grouped by city, make, model, year, trim
    - Display used car list with real price and predicted price
    - Add search or filter for used car list

APIs:

- /cars?filter=make:toyota
- /cars?q=toyota

Kilobytes:

Collection in mongodb:

1. Used_car: id, price, year, make, model, trim, odometer, city