CS CAPSTONE TECHNOLOGY REVIEW
NOVEMBER 9, 2018
Auction Hunter
Group 4

Yufei Zeng

Abstract

This document breaks the whole project into several pieces tech reviews. It will mainly talk about the crawler: coding language/framework, code repository / collaboration software, and damage estimation engine.

1 Introduction

If someone totals a vehicle, they are in a serious accident and the vehicle is so badly damaged that it is not worth repairing. At that time, the car will be auctioned by insurance auctions.Then dealers and insurance companies will buy and sell the vehicle. However, as a ordinary consumers, it is hard to know if a car is worth to bid, or how to bid at appropriate prices

The main function of Auction hunter is automatically collecting information on a variety of auction sites. So for the current work, the first task is to look for a reliable web crawler to crawl car's data from eBay Motors, ADESA, OVE, and so on.

In addition, during the process of development, we need to establish a code repository to allows us easily version and share code with other developers.

For damage estimation engine,it is a bonus part for Auction hunter, the involved technologies are too advanced for us. For this part, we would prefer to study rather than prove in the real application.

2 Web Crawler

For collecting car's data, we need to find a method to collect cars photo, bidding information, and winning bid amount which are listed on eBay Motors, ADESA, Auction Auto Mall, Dashub, A better bid, Salvage bid, Smart Auction, OVE.com, Auto Trader, and Insurance Auctions USA Inc. Integrating these information to one place. The Potential solution include that using website crawler technology to collect the relative web images and data from eBayMotors, ADESA, Auction Auto Mall, Dashub, A better bid, Salvage bid, Smart Auction,OVE.com, Auto Trader, and Insurance Auctions USA Inc. Web crawler is a kind of script which will follow a special set of rules to automatically gather information from website. Web crawler also called web scraping, or web spidering. It is used for web classification in World Wide internet. All kinds of search engine use web crawler to provide valuable searching results. In fact, web crawler collects some special HTML image or text contents or some hyperlinks from other website, and view them as an appropriate way. The scraping mainly includes two steps: The first one is that looking for and downloading web pages systematically. The second one is that extracting information from web pages we got in advance.

2.1 Scrapy

Scrapy is a twisted-based Python frame work for web crawling .It can be used to parse or extract HTML,XML,JS and CSV style documents. The main function of it includes: Requests manager, Selectors and Pipelines. For pipelines, once we got the data from website, we are able to pass them through different pipelines. Such as storing the scraped data, cleaning the HTML data, and validating data. For selectors, it is Scrapy's own mechanism, and it is used to extract data. Selector works by selecting the selected parts of HTML files which specified by the expressions of XPath and CSS.[1] For requests manger, it is responsible for downloading pages, and it will work behind the event at the same time. That is the reason why Scrapy costs less time than other web crawler. Scrapy is compatibly with some complicated crawl scenes. For example, for Auction hunter, we need to crawl many pages from different website. Using Scrapy will be a good choice.

2.2 BeautifulSoup

BeautifulSoup is one of most popular Python web scraping library. It is able to construct a Python object which based on HTML code's structure. BeautifulSoup provides several easy way to navigate, search, and modify a parse tree.It will automatically converts the form of incoming documents from Unicode to UTF-8.[2] The advantage of BeautifulSoup includes it performs to be compatible with broken or non-standard HTML and XML. So it means that it provides better compatibility. However, comparing with lxml, BeautifulSoup by itself earns a lower scores on the efficiency. It cannot perform well when parsing some extremely large HTML and XML files.

2.3 lxml

The most important feature of lxml is that it is very fast. In addition, lxml supports standards-compliant XML and broken HTML. it includes a module called soupparser which lets it to be able to fall back on the

functionality of BeautifulSoup. It means when users try to solve some broken HMTL or XML, lxml allows them to switch back to parse data from a broken
HTML by using soupparser.[3]

2.4 Conclusion

For Action Hunter, it focus on helping users to gather data about these cars more quicker and more convenient. It provides a series of information to help customers make purchasing decisions. I think Scrapy fits well.BeautifulSoup cannot deal with web side which has a bunch of items, and lxml has less compatible. Scrapy is one of most famous Python scraping libraries. It will help us to crawl a whole domain, and ignore the content type of page. It saves a lots of time for us to crawl the car's data from other web side.

3 Code Repository / Collaboration Software

The code repository can be seen like a file archive. The purpose of using code repository is to share code, and make a better software. A good code repository allows us easily version and share code with other developers & A good collaboration software will raise our working efficiency.

3.1 GitHub

We are able to store many projects in GitHub repositories, such as some open source projects. For GitHub has good help section and guide document for user who is new to GitHub.
For example, when I try to create a branch for Auction Hunter, and I don't know the flow of GitHub. I can simply search GitHub Flow within GitHub. The guide of flow shows that branching is one of most important concept of Git. When we are working on a project, we will got many features in progress. So we need to create a branch for the project, and any change we make won't affect the master branch. We can feel free to commit changes on the branch.[4] In addition, GitHub allows developer to track the changes done, we can easily find who made the change and when he made the change. And we are able to control the versions of project. If the newest version cannot be compiled, we can rollback to older version. Also, the function SCM of GitHub allows us to store the code in the same file without conflict.
The disadvantage of GitHub is that it is disagreeable towards layman. The use of GitHub is more inclined to programmer. However, all of my group members have been working with GitHub for several years.

3.2 Slack

The advantage of slack includes integration, better search feature, accurate online indication, and abundant messaging. User is able to create custom integration by slack. Slack is one of most convenience collaboration software we are using.

4 Damage Estimation Engine

As the bonus part of Auction Hunter, the purpose of developing AI damage estimation engine is that the cost of manual screening is too expensive and too ineffective, and it might make consumers pay more for the vehicle.It is necessary to finding a way to determine the value of salvage vehicles.Consumers need a tool to help them to automatically bid the vehicle. It is necessary to find a reliable tool to estimate damage images of car. However, due to the lack of relative application, estimating the damage of car is well above current level technology, so we will only focus on study for this part.

4.1 The IBM Watson Visual Recognition service

The IBM Watson Visual Recognition service analyzes a variety of photos by using learning algorithms. Such as it can be used to estimate vehicle damage. The flow of this service is that user sent the images of car after an accident to public cloud, then server side will take the photos for analysis. Finally, Watson Visual Recognition service will classify the photos, such as a broken windshield, or a broken bonnet, lastly, the results will be returned to customers.[5] It is kind of a new technology, and I had never been seen it before. But it is a open source project,and its code can be found on IBM website. I will suggest a future testing with my group.

4.2 TensorFlow

TensorFlow is using deep learning to estimate the damaged images of the car.

4.3 Datasets of damaged cars

Developers are able to train their model with cars' photos after an accident by accessing specific datasets such as Kaggle UK Car Accidents and Data.gov. and UIUC Image Database for Car Detection. Also those datasets would help us to judge which damage estimation engine is more reliable.

Reference

[1] Scrapy https://doc.scrapy.org/en/1.5/topics/selectors.html
[2] BeutifulSoup https://www.crummy.com/software/BeautifulSoup/
[3] lxml https://lxml.de/
[4] GitHub Help https://help.github.com/
[5] D. Scott(Nov 9, 2018), IBM, https://developer.ibm.com/patterns/classify-vehicle-damage-images/