GOIBIBO HOTEL DATA EXTRACTION USING WEB CRAWLER

Indian Institute of Information Technology,Allahabad (Shubham Tewari-MIT2020004)

05/02/2021

Report

Index- SOURCE CODE-> DATA ANALYSIS AND ALGORITHM

Listing 1: Source Code–

```
"""Created on Thu Feb  4 15:51:42 2021
@author: shubham
"""
import pandas as pd
import numpy as np
import re
f = open("go1", "r")
link=[]
for j in f:
  j=j+"#"
  raw_target_link=j.split("/", 1)
  new_target_link=raw_target_link[1].split("*",1)
  link.append(new_target_link[0])

import urllib.request

hotels = []
prices = []
ratings = []
for i in link:
  hyper="https://www.goibibo.com/"+i
  try:
   with urllib.request.urlopen(hyper) as response:
    html = response.read()
  except:
    continue
  from bs4 import BeautifulSoup
  soup = BeautifulSoup(html, 'html.parser')


  city_name_extract=i.split("in-",1)
```

```
32    city_name=city_name_extract[1].split("-",1)
33    hotels.append(" ")
34    prices.append(" ")
35    ratings.append(" ")
36    hotels.append("**************************************************")
37    prices.append(city_name[0].upper())
38    ratings.append("********")
39    hotels.append(" ")
40    prices.append(" ")
41    ratings.append(" ")
42
43    for t in soup.findAll('div', attrs={'class':'↩
          HotelCardstyles__HotelCardInfoWrapperDiv-sc-1s80tyk-6 dfmysf'}):
44      name=t.find('div', attrs={'class':'↩
            HotelCardstyles__HotelNameWrapperDiv-sc-1s80tyk-11 hiiHjq'})
45      price=t.find('div', attrs={'class':'↩
            HotelCardstyles__CurrentPriceTextWrapper-sc-1s80tyk-26 idchau'})
46      rating=t.find('div', attrs={'itemprop':'aggregateRating'})
47      try:
48       x=rating.text
49       u=x.split("/", 1)
50       ans=u[0]
51      except:
52       ans="N/A"
53
54      hotels.append(name.a.text)
55      prices.append(price.text)
56      ratings.append(ans)
57
58  data = {'Hotel Name' : hotels, 'Price': prices, 'Rating': ratings}
59  df = pd.DataFrame(data)
60  df.to_csv('goibibo.csv', index=False,encoding='utf-8')
```

(DATA ANALYSIS AND ALGORITHM)

Preprocessing Steps of links:

Step1- From the given link " www.goibibo.com/robots.txt ". copy all the llink paths whom you target and then make a seperate simple text file of it and name it **"go1"**

In a " www.goibibo.com/robots.txt " we will have path to link like
Disallow: /hotels/fabhotel-nachiappa-ra-puram-hotel-in-chennai- 3985175342096055412/*
Disallow: /hotels/fabhotel-capital-residency-brigade-rd-hotel-in-bengaluru-6383830869265393868/*

As we can clearly see this is not a proper link so we will extract the "/hotels/fabhotel-nachiappa-ra-puram-hotel-in-chennai- 3985175342096055412/" this part using splitting and then concate-

nate with "https://www.goibibo.com/" , so the target link will be "https://www.goibibo.com/hotels/fabhotel-nachiappa-ra-puram-hotel-in-chennai- 3985175342096055412/ "
so we will process every link of "go1" file containing the target areas.

Step2-Now we have the target link in the perfect URL manner i.e "https://www.goibibo.com/hotels/fabhotel-nachiappa-ra-puram-hotel-in-chennai- 3985175342096055412/ "
we will extract the city itself from the unprocessed link " /hotels/fabhotel-nachiappa-ra-puram-hotel-in-chennai- 3985175342096055412/ " we will simply extract that information by spitting string from " in- " and take the list second index as our city as in our case " chennai" convert it into upper case and store it in prices list .so that it can appear in front.

"**TIII NOW WE ONLY PREPROCESS THE LINK AND EXTRACT INFORMATION SUCH AS FOR WHICH CITY WE ARE EXPLORING HOTELS"**
"BUT NOW IN UPCOMING STEPS WE WILL DISCUSS HOW WE WILL EXTRACT ALL HOTEL INFORMATION (NAME,PRICE,RATING) FOR A GIVEN TARGET LINK AND SAME IS DONE FOR ALL LINK USING A LOOP:"

Step3- Analyze the web page its structure, and using **inspect element mode** and then find out the class for the features you want to extract from that page.
"In goibibo page there a preprocessing is also been done for extracting rating because the rating column was with much information and we only need a numerical value , so again a splitting is done in order to get the actual rating number .

Step 4-Convert them to text and and then store them in a seperate list.

Step5 Three feature were extracted for ever hotel and then create and save a fie using pandas library.

Remember your "go1" file should also be in the same folder where your source code file is saved , the program is designed in such a way only, I used "spyder IDE" for above functioning

"A "goibibo.csv" file will be generated into the folder of your source code file , Above implementation is done in " Spyder IDE "