Cài ubuntu

Cài ruby + lib

sudo apt-get install ruby1.9.3

https://help.ubuntu.com/community/ApacheMySQLPHP

sudo apt-get install libmysqlclient-dev
sudo gem install mysql
sudo gem install hpricot
sudo gem install nokogiri (trong lúc cài nếu lỗi thiếu thư viện nào thì cài thêm thư viện đó thêm -dev :: ví dụ thiếu thư viện abcxm thì sudo apt-get instal abcxml-dev)
sudo gem install rest-open-uri
sudo gem install selenium-webdriver

Chung: Chỉnh file lib/config.rb username và password đúng với db

Hướng dẫn chạy file trong thư mục agoda

Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho","linkthanhpho")

Link thành phố lấy như sau:
B1: vào trang http://www.agoda.vn/
B2: gõ vào ô "bạn muốn đi đâu" => "vietnam"
B3: click vào nút "Kiểm tra giá"
B4: click vào thành phố được phân công
B5: lấy link thu được cùng với tên thành phố đưa vào hàm

Cách thực thi: cd đến thư mục agoda thực thi lệnh: ruby exec.rb

Sau khi thực hiện xong send Tùng file .sql

Chú ý:
- Trong quá trình chạy có của sổ firefox tự bật lên, không được đóng cửa sổ này.

Hướng dẫn chạy file trong thư mục chudu24

Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho","linkthanhpho")

Link thành phố lấy như sau:
B1: vào trang http://www.chudu24.com/
B2: kéo trang web xuống cuối cùng phần Khách sạn theo thành phố
B3: click vào thành phố được phân công
B4: lấy link thu được cùng với tên thành phố đưa vào hàm

Cách thực thi: cd đến thư mục chudu24V2 thực thi lệnh: ruby exec.rb

Sau khi thực hiện xong send Tùng file .sql

Chú ý:
- Trong quá trình chạy có của sổ firefox tự bật lên, không được đóng cửa sổ này.

Hướng dẫn sử dụng profile của trình duyệt để tăng tốc lấy dữ liệu - không nhất thiết phải làm

Vấn đề:
- Khi load trang web có nhiều flash nên làm tốn tài nguyên hệ thống khi chạy thời gian dài
- Có nhiều link ảnh load từ nhiều nguồn khác nhau nên chậm -> dễ bị time out
- Khi link ảnh bị lỗi thì không thể load xong được -> chương trình bị dừng hoặc lỗi
- Css cũng gây ra các vấn đề như trên.

Giải pháp:
- Không load ảnh + CSS + flash.

Hệ quả: (Dấu + là tốt, dấu - là không tốt)
+ Thời gian load trang web được cải thiện đáng kể -> không có time out
+ Tài nguyên máy sử dụng ít hơn -> tốt hơn cho các máy có cấu hình thấp
- Một số site bắt buộc phải load css và ảnh thì mới lấy được dữ liệu VD là các site lấy ảnh từ google.map

Khắc phục:
- Chỉnh profile để trình duyệt chỉ cho phép load ảnh từ các site được chỉ định.

Áp dụng:
- Sử dụng file profile được share. File tên là q9og5f51.test.tar.gz đã được share trên dropbox
- Giải nén ra thư mục q9og5f51.test
- Copy thư mục này vào ~/.mozilla/firefox (cp -R q9og5f51.test ~/.mozilla/firefox)
- cd ~/.mozilla/firefox
- Chỉnh sửa 2 dòng trong file profile.ini trong thư mục này như sau:

Name=test
Path=q9og5f51.test

Hướng dẫn chạy file trong thư mục ivivu

Chú ý:
- Trong quá trình chạy có của sổ

Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho","linkthanhpho") Cách thực thi: cd đến thư mục ivivu thực thi lệnh: ruby exec.rb **Sau khi thực hiện xong send Tùng file .sql**	Link thành phố lấy như sau: B1: vào trang http://www.ivivu.com/vi/ B2: Trong dropdown thành phố: chọn thành phố được phân công B3: click vào nút tìm kiếm B4: lấy link thu được cùng với tên thành phố đưa vào hàm	
Hướng dẫn chạy file trong thư mục nto Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho",id_thanhpho) Cách thực thi: cd đến thư mục nto thực thi lệnh: ruby exec.rb **Sau khi thực hiện xong send Tùng file .sql**	Link thành phố lấy như sau: B1: vào link http://www.nto.com.vn/vn/kham-pha-va-trai-nghiem/binh-thuan-d28/ B2: Trong các link thành phố ở trên chọn thành phố được phân công, nếu không thấy thì chọn trong dropdown B3: click vào thành phố để lấy link B4: lấy id của thành phố trên url là 2 số sau tên thành phố trong URL VD: binh-thuan-d28 thì là 28 B5: lấy tên và id của thành phố đưa vào hàm	**Không có chú ý gì**
Hướng dẫn chạy file trong thư mục mytour_* Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho","link thành phố") - Trong site có 2 phần nên phải thực hiện 2 lần cho 2 thực mục mytour_diadiem và mytour_hotel Cách thực thi: cd đến thư mục mytour_* thực thi lệnh: ruby exec.rb **Sau khi thực hiện xong send Tùng file .sql**	Link thành phố lấy như sau: B1: vào link http://mytour.vn/vn/location B2: Cuối trang có list các thành phố B3: click vào thành phố được phân công để lấy link B4: lấy link của thành phố B5: lấy tên và link của thành phố đưa vào hàm, với hotel thì chuyển /localtion/ trên uri thành /hotel/ VD: http://mytour.vn/vn/location/c33/hoa-binh.html --> http://mytour.vn/vn/hotel/c33/hoa-binh.html	**Chú ý:** **- Trong quá trình chạy có của sổ firefox tự bật lên, không được đóng cửa sổ này.**
Hướng dẫn chạy file trong thư mục amthuc365 Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho","link thành phố") Cách thực thi: cd đến thư mục mytour_* thực thi lệnh: ruby exec.rb **Sau khi thực hiện xong send Tùng file .sql**	Link thành phố lấy như sau: B1: vào link http://www.amthuc365.vn/dia-chi-nha-hang.html B2: Bên phải có list các thành phố B3: click vào thành phố được phân công để lấy link B4: lấy link của thành phố B5: lấy tên và link của thành phố đưa vào hàm	**Không có chú ý gì**
Hướng dẫn chạy file trong thư mục foody		**Không có chú ý gì**

Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho",id_thanhpho) Cách thực thi: cd đến thư mục mytour_* thực thi lệnh: ruby exec.rb **Sau khi thực hiện xong send Tùng file .sql**	Link thành phố lấy như sau: B1: vào link http://foody.vn bằng chrome hoặc trình duyệt có cài công cụ bắt Network như của F12 trong chrome B2: chọn thành phố được phân công B3: click vào chữ "Địa điểm" bên cạnh thành phố vừa chọn B4: scroll chuột xuống phía dưới để dữ liệu load thêm 1 lần B5: bật F12, chuyển qua tab "Network" tìm truy vấn bắt đầu bằng "dia-diem?ss=directory&vt=row&st=1&p" B6: xem trên truy vấn có đoạn "provinceId=XXX", thì XXX là code của thành phố B7:lấy tên và link của thành phố đưa vào hàm		
Hướng dẫn chạy file trong thư mục dendau - hiện tại do không có dữ liệu mới nên không cần làm		**Không có chú ý gì**	
Hướng dẫn chạy file trong thư mục lukhach24h		**Không có chú ý gì**	
Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho",id_thanhpho) Cách thực thi: cd đến thư mục nto thực thi lệnh: ruby exec.rb **Sau khi thực hiện xong send Tùng file .sql**	Link thành phố lấy như sau: B1: vào link http://lukhach24h.com/listing/results.php?country_id=6 B2: Chọn ở mục "Phạm vi địa lý" Quốc Gia là: "Việt Nam", thành phố là thành phố được giao B3: click vào nút "Lọc kết quả" B4: ở url mới trên trình duyệt có đoạn "state_id=XXX" thì XXX là mã thành phố B5: lấy tên và id của thành phố đưa vào hàm		

Hướng dẫn chạy file trong thư mục agodaV2

Trong file config.rb Chỉnh các biến $db_user và $db_pass cho đúng với máy mình

Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho"," linkthanhpho")

Link thành phố lấy như sau:
B1: vào trang http://www.agoda.vn/
B2: gõ vào ô "bạn muốn đi đâu" => "vietnam"
B3: click vào nút "Kiểm tra giá"
B4: click vào thành phố được phân công
B5: lấy link thu được cùng với tên thành phố đưa vào hàm

Cách thực thi: cd đến thư mục agodaV2 thực thi lệnh: ruby exec.rb

Sau khi thực hiện xong send Tùng file .tar.gz

Chú ý:
- Trong quá trình chạy có của sổ firefox tự bật lên, không được đóng cửa sổ này.

Hướng dẫn chạy file trong thư mục chudu24V2

Trong file config.rb Chỉnh các biến $db_user và $db_pass cho đúng với máy mình

Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho"," linkthanhpho")

Link thành phố lấy như sau:
B1: vào trang http://www.chudu24.com/
B2: kéo trang web xuống cuối cùng phần Khách sạn theo thành phố
B3: click vào thành phố được phân công
B4: lấy link thu được cùng với tên thành phố đưa vào hàm

Cách thực thi: cd đến thư mục chudu24V2 thực thi lệnh: ruby exec.rb

Sau khi thực hiện xong send Tùng file .tar.gz

Chú ý:
- Trong quá trình chạy có của sổ firefox tự bật lên, không được đóng cửa sổ này.

Hướng dẫn sử dụng profile của trình duyệt để tăng tốc lấy dữ liệu

Vấn đề:
- Khi load trang web có nhiều flash nên làm tốn tài nguyên hệ thống khi chạy thời gian dài
- Có nhiều link ảnh load từ nhiều nguồn khác nhau nên chậm -> dễ bị time out
- Khi link ảnh bị lỗi thì không thể load xong được -> chương trình bị dừng hoặc lỗi
- Css cũng gây ra các vấn đề như trên.

Giải pháp:
- Không load ảnh + CSS + flash.

Hệ quả: (Dấu + là tốt, dấu - là không tốt)
+ Thời gian load trang web được cải thiện đáng kể -> không có time out
+ Tài nguyên máy sử dụng ít hơn -> tốt hơn cho các máy có cấu hình thấp
- Một số site bắt buộc phải load css và ảnh thì mới lấy được dữ liệu VD là các site lấy ảnh từ google.map

Khắc phục:
- Chỉnh profile để trình duyệt chỉ cho phép load ảnh từ các site được chỉ định.

Áp dụng:
- Sử dụng file profile được share. File tên là q9og5f51.test.tar.gz đã được share trên dropbox
- Giải nén ra thư mục q9og5f51.test
- Copy thư mục này vào ~/.mozilla/firefox (cp -R q9og5f51.test ~/.mozilla/firefox)
- cd ~/.mozilla/firefox
- Chỉnh sửa 2 dòng trong file profile.ini trong thư mục này như sau:

Name=test
Path=q9og5f51.test

Hướng dẫn chạy file trong thư mục ivivuV2

Chú ý:
- Trong quá trình chạy có của sổ firefox tự bật lên, không được đóng cửa sổ này.

Trong file config.rb Chỉnh các biến $db_user và $db_pass cho đúng với máy mình		
Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho"," linkthanhpho")	Link thành phố lấy như sau: B1: vào trang http://www.ivivu.com/vi/ B2: Trong dropdown thành phố: chọn thành phố được phân công B3: click vào nút tìm kiếm B4: lấy link thu được cùng với tên thành phố đưa vào hàm	
Cách thực thi: cd đến thư mục ivivuV2 thực thi lệnh: ruby exec.rb		
Sau khi thực hiện xong send Tùng file .tar.gz		
Hướng dẫn chạy file trong thư mục lukhach24hV2		**Để chạy được cần gem inststall hashie**
Trong file config.rb Chỉnh các biến $db_user và $db_pass cho đúng với máy mình		
Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho", id_thanhpho)	Link thành phố lấy như sau: B1: vào trang lukhach24h.com/listing/results.php?category_id=50&country_id=6&state_id=223 B2: Trong dropdown quốc gia chọn việt nam, trong dropdown thành phố chọn thành phố được phân công B3: click vào nút tìm kiếm B4: lấy id của thành phố trên url trong mục state_id=XXX B5: lấy tên và id của thành phố đưa vào hàm	
Cách thực thi: cd đến thư mục lukhach24hV2 thực thi lệnh: ruby exec.rb		
Sau khi thực hiện xong send Tùng 5 file XML		
Quy trình với 1 thành phố		
Các site bao gồm:	**Các bước:**	
agoda	- Lấy link	
chudu24	- Lấy raw	
ivivu	- Lấy info	
lukhach24h	- Sinh xml	
nto	- Merge dữ liệu (nhóm khác đang làm)	
amthuc365		
Hướng dẫn chạy file trong thư mục ntoV1		**Không có chú ý gì**
Trong file config.rb Chỉnh các biến $db_user và $db_pass cho đúng với máy mình		
Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho", id_thanhpho)	Link thành phố lấy như sau: B1: vào link http://www.nto.com.vn/vn/kham-pha-va-trai-nghiem/binh-thuan-d28/ B2: Trong các link thành phố ở trên chọn thành phố được phân công, nếu không thấy thì chọn trong dropdown B3: click vào thành phố để lấy link B4: lấy id của thành phố trên url là 2 số sau tên thành phố trong URL VD: binh-thuan-d28 thì là 28 B5: lấy tên và id của thành phố đưa vào hàm	

Cách thực thi: cd đến thư mục ntoV1 thực thi lệnh: ruby exec.rb			
Sau khi thực hiện xong send Tùng file tar.gz			
Hướng dẫn chạy file trong thư mục mytour_*V2		**Chú ý:** **- Trong quá trình chạy có của sổ firefox tự bật lên, không được đóng cửa sổ này.**	
Trong file config.rb Chỉnh các biến $db_user và $db_pass cho đúng với máy mình			
Trong file exec.rb thay đổi tham số cho hàm execFile như ví dụ: execFile("tenthanhpho"," link thành phố") - Trong site có 2 phần nên phải thực hiện 2 lần cho 2 thực mục mytour_diadiemV2 và mytour_hotelV2	Link thành phố lấy như sau: B1: vào link http://mytour.vn/vn/location B2: Cuối trang có list các thành phố B3: click vào thành phố được phân công để lấy link B4: lấy link của thành phố B5: lấy tên và link của thành phố đưa vào hàm, với hotel thì chuyển /localtion/ trên uri thành /hotel/ VD: http://mytour.vn/vn/location/c33/hoa-binh.html --> http://mytour.vn/vn/hotel/c33/hoa-binh.html		
Cách thực thi: cd đến thư mục mytour_*V2 thực thi lệnh: ruby exec.rb			
Sau khi thực hiện xong send Tùng file tar.gz			