**Nhận biết sale và not sale**

Đầu tiên em tải data về ( có khoảng 13 triệu data nhưng em chỉ tải 8 triệu data) rồi import vào local của em để có thể xử lý.

Trong số data lấy về thì có 257 domain nhưng có rất nhiều domain không liên quan gì đến bất động sản nên em lọc bớt những domain ko liên quan thì thu được 8 domain liên quan đến bds:

"alonhadat.com.vn"246355

"batdongsan.com.vn" 90967

"bannhachungcu.vn"22

"batdongsangroup.com"562

"batdongsanquangbinh.vn"4408

"nhadat101.com"72224

"nhadat24h.net"53123

"sanbatdongsanhanoi.com"9421

Em lấy tất cả những data của những domain này

Em chia data thành nhiều file để quá trình lọc nhanh hơn.=> thu được data liên quan đến bds.( **Em thu được khoảng 460.000 data**)

Dựa vào Url em truy cập vào một số trang liên thì thấy có nhiều url không tìm thấy nữa nên em lọc tiếp và chỉ lấy những url có bài viết.( quá trình này cũng khá lâu )

**Nhận xét:**

bannhachungcu.vn => 22 url nhưng đều không dùng đượcnữa

batdongsangroup.com => 562 url đều không liên quan đến bds

batdongsanquangbinh.vn => url bị lỗi 404

nhadat101.com => 53123 bài viết chỉ mỗi tiêu đề

nhadat24h.net =>53123 các link này cũng bị hỏng

**sanbatdongsanhanoi.com : 9421, alonhadat.com.vn: 246355, batdongsan.com.vn: 9421** lànhững data này là những data có liên quan đến bds và sử dụng để lấy được sđt

Emlấy khoảng 24.000 data của **alonhadat.com.vn** để lấy số điện thoại vì máy em chạy mất khá nhiều thời gian nên em chỉ lấy nhiêu đó để xử lý.

Nhận xét : để phân biệt sale với not sale:

nếu như người đó không phải sale thì: người đó chỉ đăng 1 bài trên **alonhadat.com.vn** rồi đóng tiền duy trì bài viết mỗi ngày cho trang đến khi nào bán được chứ không cần phải đăng nhiều bài viết cùng một nội dung lên một trang làm gì cả.=> nếu trong data số điện thoại : **số điện thoại nào xuất hiện 1-2 lần thì đó là not sale**

nếu như người đó là sale thì người đó sẽ đăng nhiều bài khác nhau về nhiều ngôi nhà căn hộ khác nhau => tần suất số điện thoại cao hơn nhiều so với not sale => **số điện thoại nào xuất hiện nhiều lần thì đó là sele**

**Hướng làm:**

**B1: từ file dữ liêu sđt tạo ra 1 túi số điện thoại**

**B2: tính tần suất xuất hiện sđt đó trong file sđt**

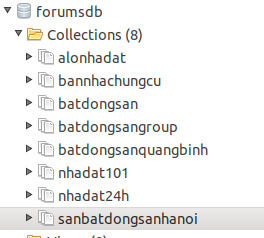
**Code:**

B1: File "domainList.py" : Tìm những "Domain" khác nhau trong bảng Articles và lưu về file "domains.txt"

B2: File "removeDomain.py" : Tìm những domain liên quan đến bất động sản sau đó xóa những data có domain không liên quan đến bds.

=> sản phẩm: ta có một tập data có domain liên quan đến bds trên local

Em chia mỗi domain thành các bảng khác nhau để dẽ quan sát và xử lý( vì số lượng domain khá ít -8 cái)



B3: File " getLinks.py": lấy hết link Url của các bài viết, chỉ lấy nhưng link chạy ra bài viết những link nào lỗi, hoặc không tồn tại nữa thì bỏ. Tất cả nhưng link bài viết sử dụng được lưu về file " ....txt" (mỗi domain là 1 file) => phân tích xong thì chỉ con 3 domain dùng được link.

B4: File " phoneNumber.py" để lấy số điện thoại - em chỉ làm thử trên 24000 data của domain " alonhadat"

B5: File index.py để xử lý phần phân loại sale và not sale.