ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

NGÀNH KHOA HỌC MÁY TÍNH

KHOA HỌC WEB

BÁO CÁO LAB 03

16/12/2019

Thông tin sinh viên:

Họ tên: Huỳnh Minh Huấn

MSSV: 161285

Nội dung báo cáo:

Em sử dụng môi trường jupyter lab + python để cài đặt thuật toán pagerank theo hướng dẫn.

Các bước cài đặt:

- Import các thư viện cần thiết: ở đây em sử dụng thư viện numpy, cùng với thư viện network để so sánh thuật toán.

```
[1]: import networkx as nx
     import numpy as np
```

- Hàm đọc vào đồ thì từ file dữ liệu:
  - Input: tên file dữ liệu
  - Output: đồ thị G được lưu trữ dưới dạng dictionary, key là đỉnh, value: là danh sách đỉnh mà nó trỏ tới.
- Đọc đồ thị từ file:

```
# đọc đồ thị dữ liệu từ file
G = graph('web-NotreDame.txt')
len(G.keys())
```

- Hàm thuật toán PageRank:

```
def pageRank(G, beta = 0.85, iter = 100, teleport_list = None, eps = 1e-8):
```

Phần thân chương trình được cài đặt theo hướng dẫn, em có sử dụng thêm thư viện numpy để hỗ trợ, cải thiện tốc độ tính toán ở một số bước.

- Kết quả đạt được:

```
%time rank_list = pageRank(G, eps=1e-08)
print(np.sum(rank_list))
```

```
Wall time: 13 s
1.0000000000000002
```

- Đọc đồ thị diG (directed Graph) theo cấu trúc của thư viện yêu cầu

Using NetworkX

```
diG = nx.DiGraph()
for key in G:
    for el in G[key]:
        diG.add_edge(key, el)
```

- Chạy thuật toán PageRank

PageRank algorithm

```
%time pagerank = nx.pagerank(diG, tol=1e-08)
```

```
Wall time: 1min 22s
```

- Vì pagerank output là dictionary nên cần chuyển về dạng numpy array để tiện so sánh

convert pagerank to numpy array

```
pageRank_list = []
for page in pagerank:
    pageRank_list.append(pagerank[page])
```

```
pageRank_list = np.array(pageRank_list)
```

```
pageRank_list
```

```
array([5.52490324e-03, 4.85749916e-04, 2.78690533e-04, ...,
       2.84880796e-06, 2.90259020e-06, 2.04667652e-06])
```

- Chạy thuật toán Hits

Hits algorithm

```
%time hits = nx.hits(diG, tol=1e-04)
```

```
Wall time: 10min 26s
```

(Vì thời gian chạy thuật toán khá lâu nên trong source em sẽ comment lại)

- Output của thuật toán Hits là một tuple (**hubs, authorities**) ứng với hits[0] và hits[1]. **Hubs** là ước tính giá trị nút dựa trên các liên kết ra, **authorities** ước tính giá trị nút dựa trên các liên kết đến.
- Ta sẽ sử dụng authorities để so sánh

So sánh kết quả giữa thuật toán PageRank tự cài đặt và thuật toán PageRank do thư viện NetworkX hỗ trợ:

Các độ đo đều cho ra kết quả với độ chênh lệch gần như bằng 0.

Sử dụng các độ đo so sánh 2 thuật toán

root mean squared error

```
[14]: np.sqrt(np.mean(np.power((np.array(pageRank_list) - rank_list), 2)))
```

```
[14]: 4.1214435370337915e-18
```

absolute mean error

```
[15]: np.mean(np.abs(np.array(pageRank_list) - rank_list))
```

```
[15]: 4.0434020954854955e-19
```

So sánh kết quả giữa thuật toán PageRank tự cài đặt và thuật toán Hits do thư viện NetworkX hỗ trợ:

root mean squared error với rank_list

```
#np.sqrt(np.mean(np.power((np.array(hits_list) - rank_list), 2)))
```

```
4.796825002639911e-05
```

absolute mean error với rank_list

```
#np.mean(np.abs(np.array(hits_list) - rank_list))
```

```
5.715781779943539e-06
```

So sánh kết quả giữa thuật toán PageRank thư viện NetworkX và thuật toán Hits do thư viện NetworkX hỗ trợ:

root mean squared error với pageRank_list

```
#np.sqrt(np.mean(np.power((np.array(hits_list) - pageRank_list), 2)))
```

```
4.7968250026397676e-05
```

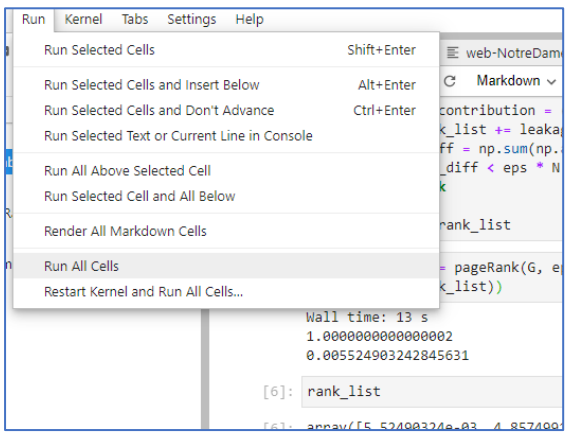
absolute mean error với pageRank_list

```
#np.mean(np.abs(np.array(hits_list) - pageRank_list))
```

```
5.715781779943196e-06
```

Hướng dẫn thực thi:

- Mở command line tại thư mục chứa file + file dữ liệu.
- Mở file jupyter notebook bằng lệnh ‘jupyter lab’ trong command line.
- Chọn Run -> Run All Cells



Tham khảo:

[1] File hướng dẫn lab 03.

[2] Hits: https://networkx.github.io/documentation/networkx-1.9.1/reference/generated/networkx.algorithms.link_analysis.hits_alg.hits.html

[3] pagerank: https://networkx.github.io/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html

[4] tài liệu lý thuyết trên lớp.