Đề tài tìm hiểu
Môn : Xử lý dữ liệu lớn

NHÓM 8

THÀNH VIÊN:

- LÂM MINH TRUNG – 52000817
- NGUYỄN KHẮC VĂN – 52000868
- TRẦN KHÁNH DUY – 52000042
- DƯƠNG TRỌNG CHÍ – 52000742
- NGUYỄN ĐĂNG HƯNG – 52000762
- TRẦN GIA HOÀNG - 52000759

1. Sử dụng thư viện

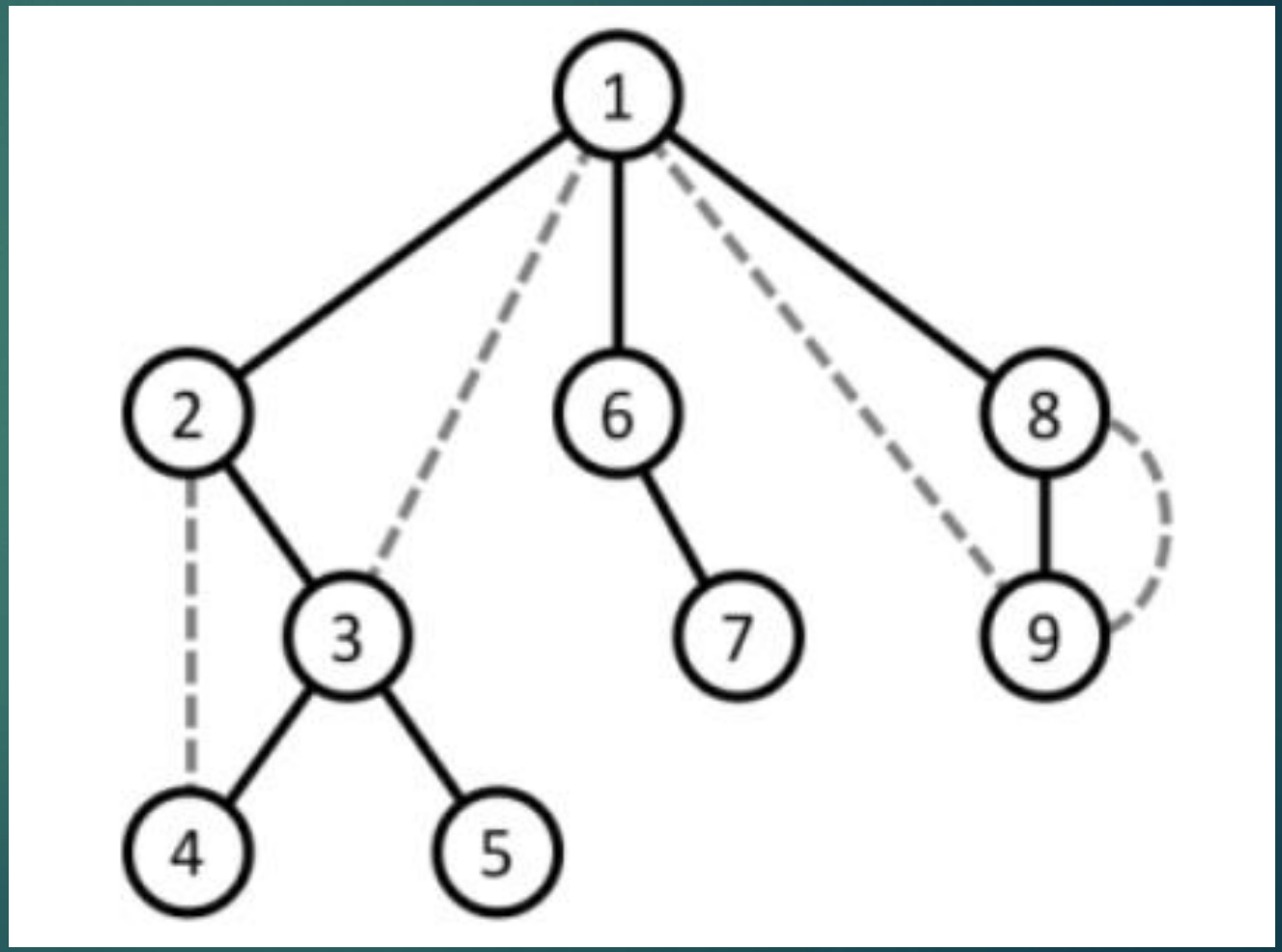
- Thư viện Request

- Thư viện Beautiful Soup

ĐẠI HỌC TÔN ĐỨC THẮNG
TON DUC THANG UNIVERSITY

2. Sử dụng kĩ thuật DFS

Áp dụng kĩ thuật DFS để duyệt các successor



ĐẠI HỌC TÔN ĐỨC THẮNG
TON DUC THANG UNIVERSITY

2. Sử dụng kĩ thuật DFS

Áp dụng kĩ thuật DFS để duyệt các successor

```
def DFS(url, prefix, n):
  d = collections.defaultdict(list) #hashmap with value = []
  stack = [url]
  visited = set()
  try: #handling generator exit
    while stack and n>0:
      vertex = stack.pop() #current url
      d[vertex] = [sub_url for sub_url in get_sub_urls(prefix)]

      #prevent row > n
      if len(d[vertex]) > n:
        d[vertex] = d[vertex][:n]
      n-=len(d[vertex])

      if vertex in visited: #O(1) lookup
        continue
      # if d[vertex]: remove for deadend check
      yield vertex,d[vertex] #ex: (page,[list of it's sub_page])

      visited.add(vertex)
      for neighbor in d[vertex]:
        stack.append(neighbor)

  except GeneratorExit:
    print("clean up first")
```

3. Sử dụng pyspark

Có rất nhiều webpsite hiện nay trên internet
Khó khăn trong việc duyệt các website và chứa trên thanh RAM

```
prefix = 'tdtu.edu.vn'
n = 10000

tree = DFS(main_page,prefix,n) #generator
rdd = sc.parallelize(tree)\
        .flatMap(lambda n: [(n[0], x) for x in n[1]])\
        .distinct()
```

Tạo rdd để chứa dữ liệu duyệt các website sucessor

```
+--------------------+--------------------+
|                Page|           Successor|
+--------------------+--------------------+
|https://tdtu.edu.vn|https://college.t...|
|https://tdtu.edu.vn|https://vfis.tdtu...|
|https://tdtu.edu.vn|http://it.tdtu.ed...|
|https://tdtu.edu.vn|http://pharmacy.t...|
|https://tdtu.edu.vn|http://feee.tdtu....|
|https://tdtu.edu.vn|http://aaf.tdtu.e...|
|https://tdtu.edu.vn|http://fss.tdtu.e...|
|https://tdtu.edu.vn|https://fas.tdtu....|
|https://tdtu.edu.vn|http://ssh.tdtu.e...|
|https://tdtu.edu.vn|http://civil.tdtu...|
|https://tdtu.edu.vn|http://laborrelat...|
|https://tdtu.edu.vn|http://law.tdtu.e...|
|https://tdtu.edu.vn|http://enlabsafe....|
|https://tdtu.edu.vn|http://ifa.tdtu.e...|
|https://tdtu.edu.vn|http://ffl.tdtu.e...|
|https://tdtu.edu.vn|http://fba.tdtu.e...|
|https://tdtu.edu.vn|http://finance.td...|
|https://tdtu.edu.vn|http://fms.tdtu.e...|
|https://tdtu.edu.vn|https://internati...|
|https://tdtu.edu.vn|https://incos.tdt...|
+--------------------+--------------------+
only showing top 20 rows
```

4. Dead-end-page

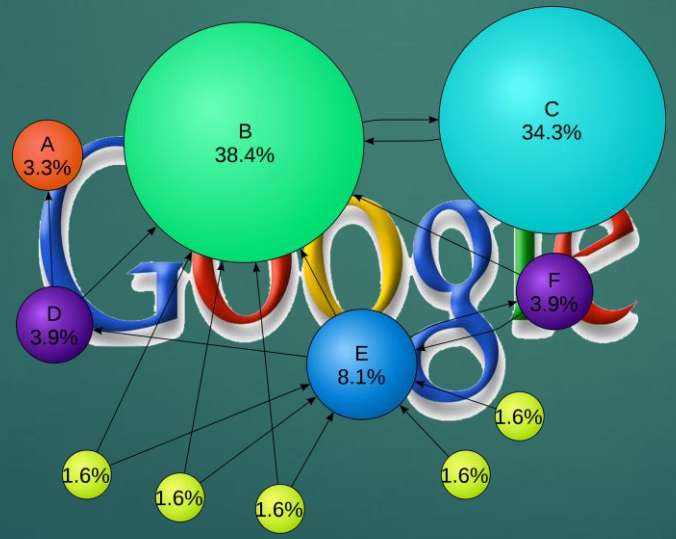
Mặt khác, dead-end page là một trang web không liên kết với bất kỳ trang web nội bộ nào khác hoặc bất kỳ trang web bên ngoài nào. Vì vậy, nó giống như đang tạo ra một cái "ngõ cụt".

Dead-end dễ dàng được khắc phục bằng cách thêm liên kết đến nội dung trên trang của bạn. Chưa kể, đảm bảo rằng điều hướng trên thanh sidebar hoặc phần dưới website được điền trên mọi trang.

5. Thuật toán PageRank (PR)

PageRank là một thuật toán được Google Tìm kiếm sử dụng để xếp hạng các trang web trong kết quả của công cụ tìm kiếm.

Thuật toán này lặp lại cập nhật xếp hạng cho từng tài liệu bằng cách cộng các đóng góp từ các tài liệu liên kết với nó.



PageRank

1. Bắt đầu từ mỗi page có hạng bằng 1

2.rank(p) / |neighbor(p)|

3. Đặt lại hạng của mỗi trang 0.15 + 0.85 * contributions

ĐẠI HỌC TÔN ĐỨC THẮNG
TON DUC THANG UNIVERSITY

5. Thuật toán PageRank (PR)

```
from operator import add
def computeContributes(urls, rank):
    for url in urls:
        yield (url, rank / len(urls))

lines = df.select(["Page","Successor"])\
            .rdd
links = lines.groupByKey()
ranks = links.mapValues(lambda neigbor: 1.0)
for _ in range(10):
    contribute = links.join(ranks).flatMap(
        lambda url_rank: computeContributes(url_rank[1][0], url_rank[1][1]))
    ranks = contribute.reduceByKey(add).mapValues(lambda rank: round(rank * 0.85 + 0.15,3))

ranks = ranks.toDF(["Page","PageRank"])
lines = lines.groupByKey()\
             .mapValues(list)\
             .map(lambda n: (n[0], n[1],len(n[1]), 0 if len(n[1])>0 else 1))\

lines = lines.toDF(["Page","Successor","Out-degree","Dead-ends"])
lines.join(ranks,on='Page').show()
```

ĐẠI HỌC TÔN ĐỨC THẮNG
TON DUC THANG UNIVERSITY

5. Thuật toán PageRank (PR)

Page	Successor	Out-degree	Dead-ends	PageRank
https://raic.tdtu...	[https://college....	45	0	0.234
https://science.t...	[https://college....	45	0	0.234
https://admission...	[https://college....	45	0	0.234
https://ecc.tdtu....	[https://college....	45	0	0.234
https://science.t...	[https://college....	45	0	0.234
https://vfis.tdtu...	[https://college....	45	0	0.234
https://emas.tdtu...	[https://college....	45	0	0.234
https://undergrad...	[https://college....	45	0	0.234
https://science.t...	[https://college....	45	0	0.234
https://nhatrang....	[https://college....	45	0	0.234
https://clc.tdtu....	[https://college....	45	0	0.234
https://student.t...	[https://college....	45	0	0.234
https://discovery...	[https://college....	45	0	0.234
http://tracuuvanb...	[https://college....	45	0	0.234
https://lib.tdtu....	[https://college....	45	0	0.234
https://vietnames...	[https://college....	45	0	0.234
https://ceca.tdtu...	[https://college....	45	0	0.234
https://science.t...	[https://college....	45	0	0.234
https://fostect.t...	[https://college....	45	0	0.234

ĐẠI HỌC TÔN ĐỨC THẮNG
TON DUC THANG UNIVERSITY

5. Tài liệu tham khảo

python - What does the "yield" keyword do? - Stack Overflow

github.com - ashishvshenoy/pagerank-spark

ĐẠI HỌC TÔN ĐỨC THẮNG
TÔN ĐỨC THANG UNIVERSITY

“Nothing great in the world has ever been accomplished without passion”

GEORGE HEGEL

Cảm ơn thầy và các bạn đã lắng nghe tổ 8 thuyết trình