# 데이터과학

L15.1: PageRank Practice

**Kookmin University** 

#### 모듈 불러오기

• 사용할 모듈 import 하기

```
import matplotlib.pyplot as plt
import numpy as np
import requests
```

## 데이터 준비

- 하이퍼링크 그래프 다운로드 받기
  - example\_index: 노드 정보
  - o example\_arc: 에지 정보

```
# http://webdatacommons.org/hyperlinkgraph
with open("example_index", "wb") as f:
    r = requests.get("http://webdatacommons.org/hyperlinkgraph/data/example_index")
    f.write(r.content)
with open("example_arcs", "wb") as f:
    r = requests.get("http://webdatacommons.org/hyperlinkgraph/data/example_arcs")
    f.write(r.content)
```

## 데이터 준비

• 하이퍼링크 그래프 불러오기

```
nodes = np.loadtxt("example_index", dtype=object)[:,0]
num_nodes = nodes.shape[0]

edges = np.loadtxt("example_arcs", dtype=int)
num_edges = edges.shape[0]
```

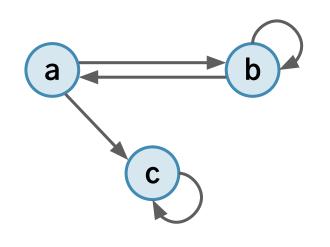
## 데이터 살펴보기

```
import networkx as nx
import matplotlib.pyplot as plt
plt.figure(dpi=300)
G = nx.DiGraph()
for e in edges:
     G.add edge(nodes[e[0]], nodes[e[1]])
pos = nx.spring_layout(G, k=1, iterations=200)
                                                                                                                                     ook.com
nx.draw(G, pos, with labels=True, node size=10, font size=5, width=0.1)
                                                                                                                                     icle 7.com
imclacom
                                                                                                                                         minimalist.co
                                           amazon o.uk
                                                                                                                                           twitten.com
                                       1000notes.com
                                                                                                                                              am>20n.ca
                                      beautybar.com
                                                                wiktionary.org
                                                                                                                                              diapers.com
                                                                                 amazon co.jp
                                   creativecommons.org
                                                                                                                                             locke@.com
                                                                                                    zoomblog.com downthis ideo.com
sapo.pt mesvilaweb.cat
blogia.com
                                      audible om
                                                                                                                                               gmpo.org
                                                                                                                                               vodeod.com
                                        flickepet
                                                                                                             kenyaunlimited.com
                                                                           animationplayhouse.com
                                    hockeyfigkts.com
                                                                                                             goodle.compenad.com
                                                                                                                                             techneati.com
                                       automattic.com
                                                                                                                                                histas.com
                                                                                                            eltangoysusinvitados.com
                                        cypri€ .fr
                                                                                         gwika com
                                                                                                                                              carambatv.ru
                                                                                                  pimedia.commemeorandum.com
                                        nytimescom
                                                                                                                               over-blog.com canablog.com
                                                                                                classicalvalues.comunu
                                          fabricom
                                                                                                                                            instagr.am
                                            tinyur.com
                                                                                   flickcom
                                                                                                                                        wikir edia.org
                                              disqus.com
                                                                                                                                        livejounal.com
                                                  amazon,fr
                                                                                              blogalaxia.com
                                                                                                                                    vaho com.au
                                                                                                            vahoo com
                                                                                                                                   drdrum.com
                                                    bookdepository.com
100500.tv
                                                                                                                              smallparts.com
                                                                                                 realestate.com.au
                                                               ioblabe books.com
                                                                                                           linkwithin com
rea-group.com
irjavanide
```

# **Google Matrix**

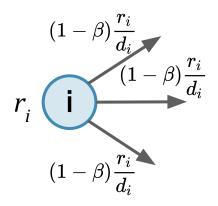
Google Matrix는 Dense 함
 ⇒ 연산이 과도하게 많음

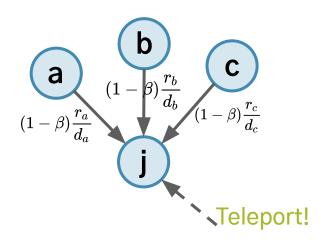
$$A = \beta M + (1 - \beta) \left[ \frac{1}{N} \right]_{N \times N}$$



#### 보다 효율적인 구현 방법

- Adjacency List 형식으로 데이터를 구성
- 각 노드마다, 그 노드의 점수를 out neighbor들에 골고루 나눠줌
- 각 노드마다, 받은 점수를 합산 함





# pagerank()

- edges: 에지 목록
- beta: teleport 확률
- threshold: r 벡터 변화량이 이 값보다 작을경우 계산 중단
- epochs: 최대 반복 계산 횟수

```
def pagerank(edges, beta, threshold=10^-20, epochs=100):

# Adjacency List 만들기
neighbors = [[] for _ in range(num_nodes)]

for e in edges:
    neighbors[e[0]].append(e[1])

# 벡터 초기화: 모든 값을 (1/n)으로 설정
r = [1/num_nodes] * num_nodes
```

# pagerank()

return r

```
for epoch in range(epochs):
   # 업데이트된 값을 저장할 벡터 초기화
   r next = [0] * num_nodes
   # 모든 노드에서...
   for u in range(num nodes):
      # 그 노드가 가리키는 이웃 노드에게...
      for v in neighbors[u]:
          # 현재 자신의 점수 r[u]에 (1-beta)을 곱하고, 등분하여 나눠줌
          r next[v] += (1 - beta) * r[u] / len(neighbors[u])
   # 1-(r next의 값들의 합)은 teleport에 의해 각 노드에 도착할 확률임
   # 그러므로, 1-(r next의 값들의 합)을 모든 노드에게 등분하여 나눠줌
   teleport prob = 1-sum(r next)
   for u in range(num nodes):
       r next[u] += teleport prob/num nodes
   # 이전 벡터 r과 새로운 벡터 r next 사이의 변화량을 합산
   delta = sum(abs(a-b) for a, b in zip(r, r next))
   # 원래 벡터를 새로 계산한 벡터로 교체
   r = r next
   if delta < threshold:</pre>
       break
```

#### 계산 및 출력

(0.016119) youtube.com
(0.015904) tumblr.com
(0.015904) google.com

(0.015600) wikimedia.org (0.015058) rea-group.com

(0.013151) yahoo.com

```
r = pagerank(edges, 0.15)
for score, node in sorted(zip(r, nodes), reverse=True)[:10]:
    print(f"({score:.6f}) {node}")

(0.085426) blogspot.com
(0.024220) creativecommons.org
(0.021582) wikipedia.org
(0.017731) canalblog.com
```

## Questions?