Lab4: Pagerank

Q1 2022-23

Integrants:
Pablo Montón Gimeno
Cristian Sánchez Estapé

Característiques de la implementació

Es fan les següents consideracions:

- S'ha realitzat una reinterpretació de l'esquelet del programa la qual bàsicament es redueix a la naturalesa de la classe *Edge* i *Airport*: la primera actua com a indicador del nombre total d'arestes que surten del mateix node (funcionalitat equivalent a l'atribut *outweight* de la classe *Airport*); la segona fa un ús particular dels atributs *routes* i *routeHash*, ja que aquests emmagatzemen les rutes de les quals l'aeroport mateix n'és destinatari (o cosa que és el mateix: sent *j* un aeroport qualsevol, el seu *routeHash* emmagatzema, per tot aeroport *i* tal que existeix una ruta *(i, j)*, el nombre de rutes *(i, j)* existents).
- La implementació del mètode *readRoutes*: fa una lectura del fitxer *routes.txt* en què, si detecta un aeroport que no figura al fitxer *airports.txt*, els elimina. A més a més, construeix (en el sentit prèviament descrit) tant els objectes *Edge* com els atributs *routes* i *routeHash* dels deguts aeroports.
- La implementació del mètode *computePageRanks*, que comprèn diversos aspectes: en primer lloc, s'ha optat pels diccionaris com a estructura de dades per emmagatzemar la informació computada del *pagerank*, cosa que garanteix l'eficiència computacional del càlcul i que alhora es veu reflectida en els objectes *P* i *Q*; en segon lloc, s'ha optat per la implementació del procediment auxiliar *getVals* que, donat un aeroport i un diccionari de *pagerank* global *P*, fa el còmput corresponent (vegeu el codi) i retorna una llista de valors associada; en tercer lloc, i com a aspecte més important, destaquem el posterior procediment de normalització dels càlculs de *pagerank*: per cada iteració (i especialment a la primera), el sumatori del *pagerank* de cadascun dels aeroports no necessàriament ha de ser 1, amb la qual cosa, en el nostre cas, hem optat per normalitzar, a cada iteració del mètode, el valor del *pagerank* associat a cada node (o aeroport), amb la qual cosa garantim que, per una banda, la suma sempre sigui (virtualment) 1, alhora que no incorrem en incoherències associades a la modificació o assignació de valors de pagerank anòmals. Matissem aquesta darrera declaració: en lloc de normalitzar els valors de *pagerank*, una cosa que es podria haver fet és diferenciar entre nodes *pou* (o *sinks*) i nodes *no-pou*, de tal manera que, mentre pels que no són pous computem el seu degut valor de *pagerank* d'acord a la fórmula definida, pels nodes *pou* podríem haver decidit, donada la seva naturalesa, assignar-los, per exemple, un valor intermedi (en aquest cas, equivalent al pagerank mig dels nodes *no-pou*), la qual cosa ens hagués conduït a una mala solució (els nodes *pou* tindrien més *pagerank* que aquells que només tinguessin 1 aresta de sortida, per exemple). Entenem que aquesta estratègia no és única, però sí que apunta a un càlcul correcte del *pagerank* que li pertoca a cada aeroport. Tot i això, les conseqüències d'optar per aquest enfocament es comentaran a la següent secció. En quart lloc, definim la condició de parada en funció de la quantitat de decimals del *pagerank* total que han canviat entre dues iteracions. Finalment, el procediment retorna una tupla formada pel nombre d'iteracions que ha executat el mètode, juntament amb el diccionari de *pageranks*.
- La implementació del mètode outputPageRanks ha implicat la modificació de l'estructura original del mètode: ara rep un paràmetre corresponent al diccionari de *pageranks*, i mostra per pantalla, de manera formatada, els *pageranks* ordenats de manera descendent.

Observacions

Damping factor i límit

Al mateix enunciat de la pràctica, se'ns indica que el *damping factor* més popular acostuma a ser entre 0,8 i 0,9. Nosaltres hem volgut observar què és el que passa quan triem valors molt petits o molt grans (sobre el rang [0, 1]) per identificar quin és el valor més adequat per la nostra implementació.

A banda d'això, el límit que hem utilitzat a la nostra implementació final ha estat triat d'acord amb el nombre de decimals de diferència entre dues iteracions subsegüents, d'acord a què l'hem considerat el criteri més coherent amb el resultat desitjat del mètode. Vam considerar altres opcions, com un nombre fix d'iteracions, o fins i tot respecte al nombre de decimals de diferència del valor mitjà del pagerank entre dues iteracions subsegüents, però cap d'aquests altres criteris resultava millor, almenys en la mesura en què no es basen en el pagerank (núm. fix d'iteracions), o no prou coherentment (valor mitjà del pagerank, que podria ser qualsevol valor, a diferència de la suma de tots els pageranks, que sabem que és 1). A partir de l'experimentació respecte al límit es veuen les conseqüències (prèviament esmentades) de la normalització que fem del pagerank: la diferència del pagerank observat entre dues iteracions acostuma a ser ínfima, i si el límit no és molt petit, la convergència es produeix gairebé de manera immediata, motiu pel qual, si es vol fixar un límit, aquest hauria de ser petit si es vol garantir que el mètode faci més de 2 iteracions.

Per fixar un *Damping Factor* i un *límit*, i a la vegada veure com varien les iteracions i temps de computació dels pageranks en funció d'aquests, hem fet una variació del script anomenat *PageRankDF.py.* És degut a la normalització dels *pagerank* que hem de disminuir molt el nombre de dígits de diferència del límit per tal de veure canvis en les iteracions i, per tant, en el temps de computació.

Limit	0.0001

DampFact	Iterations	CompTime
0.1	2	0.024343490600585938
0.2	2	0.0245361328125
0.4	2	0.022649526596069336
0.6	2	0.022733688354492188
0.8	2	0.02654552459716797
0.85	2	0.023172616958618164
0.88	2	0.022887706756591797
0.9	2	0.02338719367980957
0.99	2	0.023308753967285156

Limit	1e-16

DampFact	Iterations	CompTime
0.1	11	0.13338685035705566
0.2	20	0.24706363677978516
0.4	36	0.4293804168701172
0.6	73	0.8689470291137695
0.8	144	1.6732077598571777
0.85	196	2.2907662391662598
0.88	179	2.0796165466308594
0.9	134	1.5591044425964355
0.99	1056	12.695318698883057

Limit	1e-08

DampFact	Iterations	CompTime
0.1	2	0.024393081665039062
0.2	2	0.025931835174560547
0.4	2	0.023682594299316406
0.6	2	0.023371458053588867
0.8	2	0.02318096160888672
0.85	2	0.02342987060546875
0.88	2	0.02334308624267578
0.9	2	0.023443937301635742
0.99	2	0.023305892944335938

Limit	1e-32

DampFact	Iterations	CompTime
0.1	11	0.1221468448638916
0.2	20	0.21446752548217773
0.4	36	0.3870055675506592
0.6	73	0.7904336452484131
0.8	144	1.549682378768921
0.85	196	2.137850046157837
0.88	179	1.9292707443237305
0.9	134	1.4402389526367188
0.99	1056	11.524076461791992

Còmput del pagerank segons el límit

Com veiem, el *damping factor* òptim seria el de 0,9, ja que traiem un temps i iteracions que és una mena de mínim local, però dona resultats adients. Ja tenint un *damping factor* fix, és interessant realitzar una petita experimentació sobre com varia el pagerank segons el límit. Podem veure com els resultats varien segons aquest o com hi ha aeroports que no surten fent menys iteracions.

Limit | 0.0001

IATA CODE	OUTWEIGHT	PAGERANK
LAX	507	0.00580501
LHR	504	0.00529528
SIN	393	0.00514855
DEN	459	0.00493836
CDG	506	0.00482394
FRA	494	0.0047727
ORD	534	0.00454391
SYD	218	0.00438644
DXB	356	0.00404986
JFK	424	0.00395663
AMS	464	0.00391999

Limit | 1e-08

IATA CODE	OUTWEIGHT	PAGERANK
LAX	507	0.00580501
LHR	504	0.00529528
SIN	393	0.00514855
DEN	459	0.00493836
CDG	506	0.00482394
FRA	494	0.0047727
ORD	534	0.00454391
SYD	218	0.00438644
DXB	356	0.00404986
JFK	424	0.00395663
AMS	464	0.00391999

Limit | 1e-16

IATA CODE	OUTWEIGHT	PAGERANK
ORD	534	0.00676342
LAX	507	0.00675319
DEN	459	0.00632729
LHR	504	0.00574408
CDG	506	0.00555672
PEK	515	0.00550613
FRA	494	0.00543275
SIN	393	0.00507591
ATL	398	0.00503974
AMS	464	0.00498408
JFK	424	0.0049619

Limit | 1e-32

IATA CODE	OUTWEIGHT	PAGERANK
ORD	534	0.00676342
LAX	507	0.00675319
DEN	459	0.00632729
LHR	504	0.00574408
CDG	506	0.00555672
PEK	515	0.00550613
FRA	494	0.00543275
SIN	393	0.00507591
ATL	398	0.00503974
AMS	464	0.00498408
JFK	424	0.0049619

Veient els resultats d'aquests experiments, veiem que la normalització del pagerank, juntament amb un *damping* factor fix, té un clar efecte sobre els aeroports amb més pagerank, posat que aquests clarament varien no solament en termes d'ordre, sinó també en la resposta donada (i.e. vegeu que, pels dos casos inferiors, hi apareixen els aeroports *PEK* i *ATL*, els quals no apareixen als superiors). Amb això, ens aventurem a assumir l'existència d'un límit $\mathfrak{L}$ a partir del qual la solució donada pel programa esdevé igual. Nogensmenys cal afegir que, si bé és possible que existeixi un límit $\mathfrak{L}'$ a partir del qual la solució torni a veure's modificada (que per motius d'extensió no hem intentat trobar), també cal afegir que no s'ha d'assumir que, a menor límit, major temps de còmput (i.e. vegeu que, en les figures de l'apartat anterior, per un *damping factor* = 0,9 es tarda més amb un límit de 1e-16 que un de 1e-32) i, per tant, valdria la pena definir aquest paràmetre d'acord amb una experimentació més exhaustiva que la presentada en aquest document.