Ferran Pérez, U160396
Rosa Alós, U150080
P201

Assignment 3
Page Rank

1. The list of Top-20 sites by PageRank

```
##########top 20 Sites Page Rank##########
0.0015168643782398274 www.opsi.gov.uk
0.0014182182459715198 www.adobe.co.uk
0.0009654563807161581 www.ico.gov.uk
0.00089555331065544 www.dti.gov.uk
0.0008937889977065464 www.defra.gov.uk
0.000780473245157769 news.bbc.co.uk
0.0007209475362666682 www.direct.gov.uk
0.000697529301060852 www.dfes.gov.uk
0.0006817074015284916 www.fsa.gov.uk
0.0006581577785798681 www.nationalrail.co.uk
0.0006554311467163489 www.communities.gov.uk
0.000648287488l544456 www.bbc.co.uk
0.0006028805944996983 www.google.co.uk
0.0005906484974062024 www.dh.gov.uk
0.0005818417136696459 www.hmso.gov.uk
0.0005757813160594587 www.hse.gov.uk
0.000540229700172051 www.fco.gov.uk
0.0005155226994647975 www.nationaltrust.org.uk
0.000483592734339035 www.homeoffice.gov.uk
0.0004584891772400869 mysite.wanadoo-members.co.uk
```

2. The list of Top-20 sites by PageRank without spam.

```
##########top 20 Sites Page Rank w/o Spam##########
0.0015163269912291798 www.opsi.gov.uk
0.0014177550573844725 www.adobe.co.uk
0.0009642175795585692 www.ico.gov.uk
0.0008953203325559037 www.dti.gov.uk
0.0008925273921990404 www.defra.gov.uk
0.0007798118062589196 news.bbc.co.uk
0.0007204877150489438 www.direct.gov.uk
0.0006978022269123422 www.dfes.gov.uk
0.0006801444750391456 www.fsa.gov.uk
0.0006574169744247544 www.nationalrail.co.uk
0.0006552851670602633 www.communities.gov.uk
0.0006481317727919905 www.bbc.co.uk
0.0006023073035861512 www.google.co.uk
0.0005910116655537075 www.dh.gov.uk
0.0005818574145054327 www.hmso.gov.uk
0.000575379388422 79 www.hse.gov.uk
0.0005390379973780 92 www.fco.gov.uk
0.0005154578899466747 www.nationaltrust.org.uk
0.0004833514113988187 www.homeoffice.gov.uk
0.0004575516290926249 4 mysite.wanadoo-members.co.uk
```

3. Comment the results

Primeramente, vemos como el efecto del spam se puede obviar. Los valores cambian ínfimamente, lo que es señal que una lista de 300 páginas spam no tiene mucho efecto en una base de datos del orden de 100.000 páginas web. Si entramos al detalle, vemos como las páginas que probablemente usen sitios de spam para promocionarse como mysite.wanadoo-members.co.uk pierden un poco de su puntuación en favor de páginas que no usen esas técnicas, como dfes.gov.uk (web de educación británica).

Además, por lo general todas las páginas pierden un poco de puntuación, señal que todas las páginas de spam que ya no computamos las incluían para ganar popularidad dentro del sistema y que las páginas de spam no se quedan a puntuación 0, sinó a la puntuación inicial de 1/cantidad de páginas.

4. The list of top-20 sites containing .co.uk by PageRank

```
##########top 20 Sites Page Rank with co.uk domain ##########
0.0014182182459715198 www.adobe.co.uk
0.000780473245157769 news.bbc.co.uk
0.0006581577785798681 www.nationalrail.co.uk
0.000648282874881544456 www.bbc.co.uk
0.0006028805944996983 www.google.co.uk
0.00045848917724008694 mysite.wanadoo-members.co.uk
0.0004268162067476992 www.actinic.co.uk
0.0003640398196569995 www.networkrail.co.uk
0.00032710620860882734 www.caa.co.uk
0.00032325517863542096 www.erolonline.co.uk
0.00031455563140291595 www.punterlink.co.uk
0.00030441339496852974 www.streetmap.co.uk
0.0003031045384251774 4 www.tso.co.uk
0.0002926680894983065 www.kelkoo.co.uk
0.00028086861863899906 www.guardian.co.uk
0.0002781833278394606 www.rac.co.uk
0.00026384699236390 58 www.event-management-uk.co.uk
0.00024662598481841865 www.telegraph.co.uk
0.00023703672636871383 www.investorsinpeople.co.uk
0.00021834116672779356 www.business-directory-uk.co.uk
0.0002080517504550283 www.infotex.co.uk
```

5. The list of top-20 sites containing co.uk by No-spam-PageRank

```
##########top 20 Sites Page Rank with co.uk domain w/o Spam ##########
0.0014177550573844725 www.adobe.co.uk
0.0007798118062589196 news.bbc.co.uk
0.0006574169744247544 www.nationalrail.co.uk
0.000648131772791990 5 www.bbc.co.uk
0.0006023073035861512 www.google.co.uk
0.00045755162909262494 mysite.wanadoo-members.co.uk
0.0004275822265788012 www.actinic.co.uk
0.0003635781173426 81 www.networkrail.co.uk
0.00032670701629315484 www.caa.co.uk
0.0003173640738929826 www.erolonline.co.uk
0.0003165194895798746 6 www.punterlink.co.uk
0.0003042495106094141 6 www.streetmap.co.uk
0.0003028006825034482 www.tso.co.uk
0.0002928870966451433 www.kelkoo.co.uk
0.0002805098457623399 6 www.guardian.co.uk
0.0002779493305391587 www.rac.co.uk
0.0002659336091578067 www.event-management-uk.co.uk
0.0002461311067730707 www.telegraph.co.uk
0.0002369173091153006 7 www.investorsinpeople.co.uk
0.0002190358778869699 3 www.business-directory-uk.co.uk
0.0002076140042528688 7 www.infotex.co.uk
```

6. Comment the results

Nuevamente tenemos unos resultados donde la eliminación del spam no supone una gran diferencia. Esta vez somos capaces de ver a resultados menos relevantes del dominio co.uk, y como es un dominio heterogéneo podemos observar una replica bastante exacta de los anteriores resultados: las páginas que son más susceptibles a beneficiarse del spam presentan caídas mayores que las páginas más respetables, pero casi todas las páginas sufren pérdidas derivadas de las páginas de spam vinculadas a ellas.

7. The list of top-20 sites containing .gov.uk by PageRank

```
##########top 20 Sites Page Rank with gov.uk domain ##########
0.0015168643782398274 www.opsi.gov.uk
0.0009654563807161581 www.ico.gov.uk
0.00089555331065544 www.dti.gov.uk
0.0008937889977065464 www.defra.gov.uk
0.0007209475362666682 www.direct.gov.uk
0.000697529301060852 www.dfes.gov.uk
0.0006817074015284916 www.fsa.gov.uk
0.0006554311467163489 www.communities.gov.uk
0.0005906484974062024 www.dh.gov.uk
0.0005818417136696459 www.hmso.gov.uk
0.0005757813160594587 www.hse.gov.uk
0.000540229700172051 www.fco.gov.uk
0.0004836592734339035 www.homeoffice.gov.uk
0.0004572991098763979 www.dft.gov.uk
0.0004455162026640717 4 www.dataprotection.gov.uk
0.0004375347957066734 www.dwp.gov.uk
0.0004196267367553457 www.legislation.hmso.gov.uk
0.0003958947431418644 www.informationcommissioner.gov.uk
0.0003718424401708976 6 www.statistics.gov.uk
0.0003704741179414513 www.hm-treasury.gov.uk
0.0003391975486385459 www.tfl.gov.uk
```

8. The list of top-20 sites containing gov.uk by No-spam-PageRank

```
##########top 20 Sites Page Rank with gov.uk domain w/o spam##########
0.0015163269912291798www.opsi.gov.uk
0.0009642175795585692www.ico.gov.uk
0.0008953203325559037www.dti.gov.uk
0.0008925273921990404www.defra.gov.uk
0.0007204877150489438www.direct.gov.uk
0.0006978022269123422www.dfes.gov.uk
0.0006801444750391456www.fsa.gov.uk
0.0006552851670602633www.communities.gov.uk
0.0005910116655537075www.dh.gov.uk
0.0005818574145054327www.hmso.gov.uk
0.000575537938842279www.hse.gov.uk
0.0005390379973780 92www.fco.gov.uk
0.0004833514113988187www.homeoffice.gov.uk
0.0004571642600167578 3www.dft.gov.uk
0.0004445341940639843www.dataprotection.gov.uk
0.000437352163838 5658www.dwp.gov.uk
0.0004194319700341 24www.legislation.hmso.gov.uk
0.0003956975759919554www.informationcommissioner.gov.uk
0.0003717872266842816 4www.statistics.gov.uk
0.0003700056345806674 4www.hm-treasury.gov.uk
0.000339140521697311 9www.tfl.gov.uk
```

9. Comment the results

Esta vez si que tenemos unos resultados diferentes a los apartados anteriores, puesto que ahora solo contemplamos páginas web procedentes del gobierno, así que serán respetables por antonomasia.

Es por esto por lo que vemos que todos los resultados se mantienen muy parecidos y no se registran pérdidas. En algunos sube un poco, y en otros baja un poco, pero es negligible.

10. The list of top-20 sites by spam gain

```
33.14277368425596 www.escortnet.co.uk
29.058929855290366 www.missionfish.org.uk
17.898491711374145 www.statistics.006.free-counter.co.uk
13.6422600947175 www.uk-shoponline.co.uk
10.800429380479825 www.shop.co.uk
10.417142543549902 www.geordie-girls.co.uk
10.353438590154301 www.into.demon.co.uk
10.069001846752684 www.computerarts.co.uk
9.320732629825939 www.aili.co.uk
8.869703491243296 connect4fun.co.uk
8.452918216407495 www.kompass.co.uk
8.003510769421466 www.mercurywd.co.uk
7.8800664219899685 www.theshopping-centre.co.uk
7.824324020347563 www.markwarner.co.uk
7.7418348004899045 www.suppliersnearby.co.uk
7.67197459115261 www.quality-site-finder.co.uk
7.531583765512691 www.hertfordshiremobilediscos.co.uk
6.78796705669567 www.eastwoodtoday.co.uk
6.666666666666665 www.jlc.me.uk
5.9640753129903645 www.ideas21.co.uk
```

11. Comment the results

Aquí los resultados son mucho más esclarecedores que en apartados anteriores. Como era de esperar, observamos el conjunto de páginas más sospechosa de ejercer prácticas de spam o de beneficiarse de ellas. Hemos entrado en algunas de ellas y hemos podido comprobar que se dedican a la promoción de artículos o servicios y que siguen el patrón de las páginas que utilizan técnicas de spam.

12. A brief description of your variant of PageRank

Nuestra variante de PageRank funciona de manera parecida al PageRank original, pero con la diferencia que en vez de dividir cada valor por el outlink degree lo dividimos por la raíz cuadrada del outlink degree. Esto provoca que las puntuaciones de las páginas con muchas conexiones se diluyan menos entre cada uno de sus edges. Si por ejemplo una página como google que en cada iteración tiene mucha puntuación, pero muchas conexiones entre las que repartirla, ahora dará un valor mayor a cada una de esas conexiones. Eso provoca un desequilibrio que nos hace tener que normalizar cada vez que hacemos una iteración.

13. The list of top-20 sites by your variant of PageRank

```
##########top 20 Sites Page Rank##########
0.0044283353851542135 www.dataprotection.gov.uk
0.003529314947193229 www.libdems.org.uk
0.0035153398624564874 www.prai.co.uk
0.003508030404579182 islington-libdems.org.uk
0.0035023627094074306 warwick-leamington-libdems.org.uk
0.003502302101910261 6 libdems4london.org.uk
0.003502176469459167 4 montlibdems.org.uk
0.003502167058670223 2 chichesterlibdems.org.uk
0.003502151950023693 6 surreyheathlibdems.org.uk
0.003502150155631655 bobrussell.org.uk
0.00350205251233454 stevegoddard.org.uk
0.003502027306607252 6 emilygasson.org.uk
0.003502027306607252 6 jameskeeley.org.uk
0.003502919381198015 bracknell-libdems.org.uk
0.003502871135787527 darren4streatham.org.uk
0.003502871135787527 friendsofstoneymiddletonschool.org.uk
0.003502871135787527 garylawson.org.uk
0.003502871135787527 jamesquinlanforparliament.org.uk
0.003502871135787527 liberty-network.org.uk
0.003502871135787527 lizleffman.org.uk
```

14. Comment the results

Nuestra variante de PageRank premia a las páginas poco populares, principalmente páginas de organizaciones .org. Esto deducimos que se debe al motivo que hemos comentado antes. Estas páginas tienen muchas conexiones con muchas páginas importantes, y ahora que cada una de esas les pasa una mayor cantidad de puntuación, ocurre que la puntuación con la que acaban es mayor que la propia de la página importante.