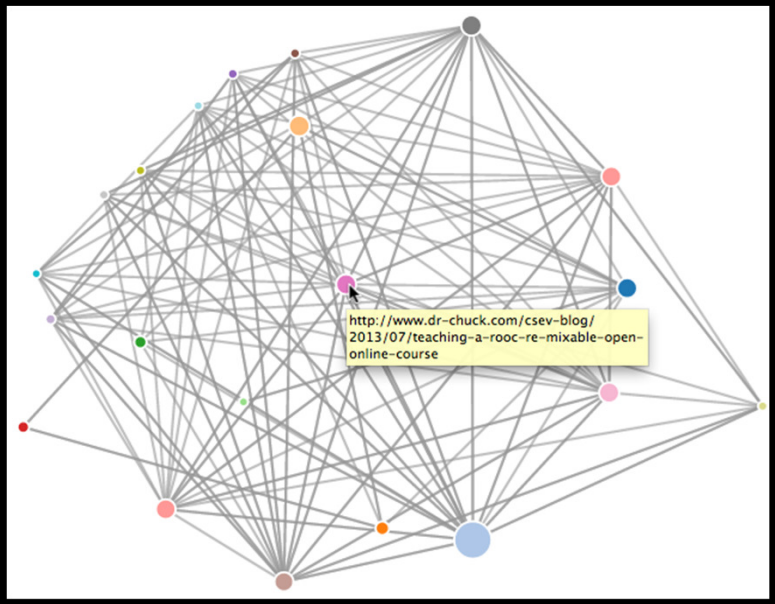
# Visualización de Datos con Python 3.7 y D3.js

Alfonso Neil Jiménez Casallas
Ingeniero de sistemas
Pontificia Universidad Javeriana

https://github.com/pochito427/
alfonso.jimenez@javeriana.edu.co
https://www.linkedin.com/in/alfonso-jimenez-3208a120/

# Caso de estudio: Page Rank

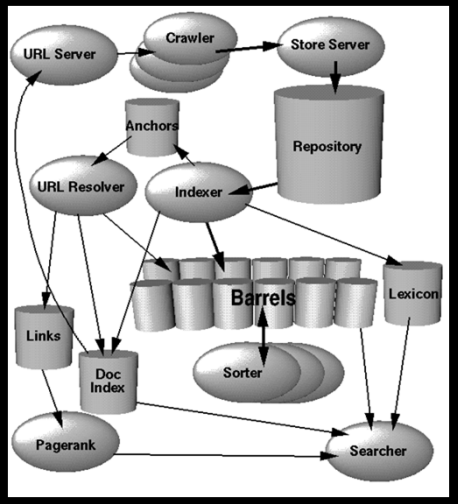
- Escribir un rastreador (crawler) simple de página Web
- Calcular una versión simple del algoritmo Page Rank de Google
- Visualizar la red resultante




http://www.py4e.com/code3/pagerank.zip

# Arquitectura de Motores de Búsqueda

- Rastreo Web (Web Crawling)
- Indexación (Index Building)
- Búsqueda (Searching)




http://infolab.stanford.edu/~backrub/google.html

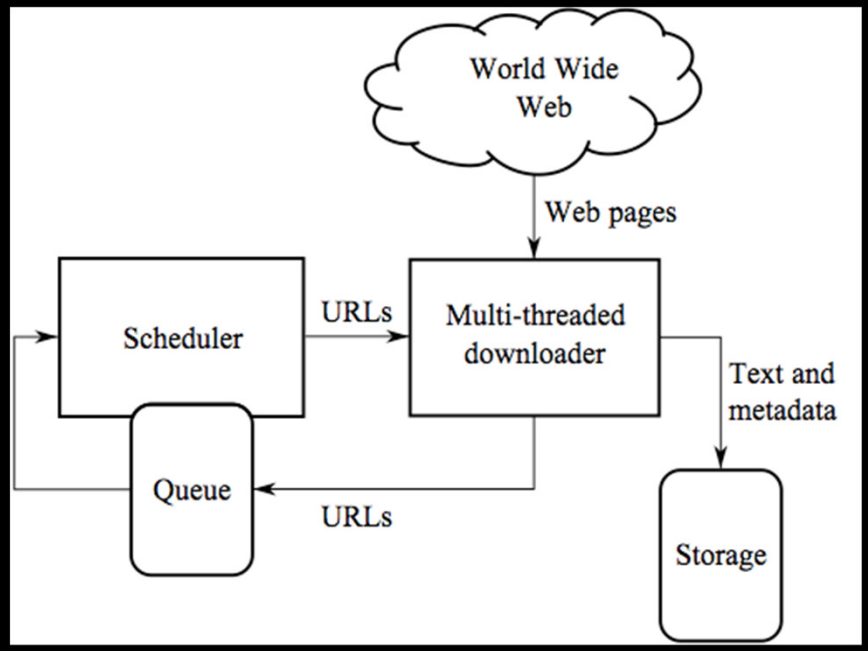
# Rastreador Web (Web Crawler)

Programa en computador que comprende la Internet (World Wide Web) de una manera metódica y automatizada. Es usado principalmente para crear una copia de todas las páginas visitadas para procesamiento posterior por un motor de búsquedas que indexará las páginas descargadas para proveer búsquedas rápidas.

http://en.wikipedia.org/wiki/Web_crawler

# Web Crawler

- Recupera una página
- Examina la página para encontrar enlaces
- Agrega los enlaces a una lista de sitios “a ser recuperados”
- Repite el proceso...




http://en.wikipedia.org/wiki/Web_crawler

# Políticas de Rastreo Web

- Una política de selección que establece cuáles páginas a descargar,
- Una política de re-visitas que establece cuándo verificar cambios a las páginas,
- Una política de cortesía que establece cómo evitar la sobrecarga de sitios Web, y
- Una política de paralelización que establece cómo coordinar rastreadores Web distibuídos

# robots.txt

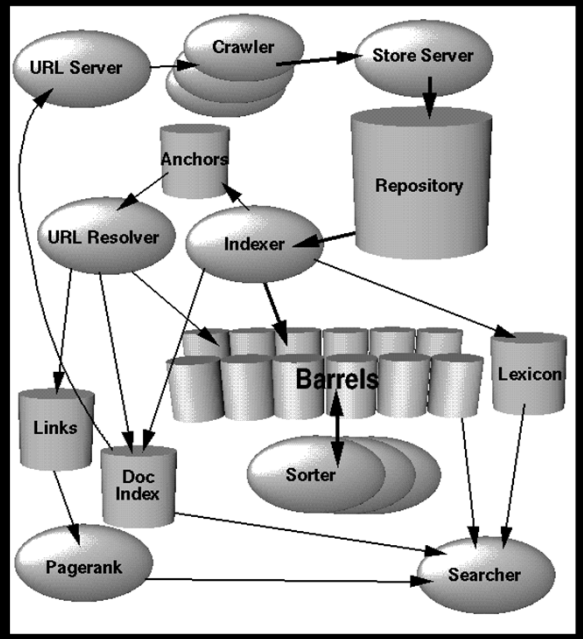
- Una vía para un sitio web de comunicar con los rastreadores web
- Un estándar informal y voluntario


```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /tmp/
Disallow: /private/
```

http://en.wikipedia.org/wiki/Robots_Exclusion_Standard
http://en.wikipedia.org/wiki/Spider_trap

# Arquitectura Google

- Web Crawling
- Index Building
- Searching




http://infolab.stanford.edu/~backrub/google.html

# Indexación de Búsquedas

La indexación en motores de búsqueda recopila, analiza, y almacena datos para facilitar la recuperación rápida y precisa de información. El propósito de almacenar un índice es optimizar velocidad y rendimiento en encontrar documentos relevantes para una consulta de búsqueda. Sin un índice, el motor de búsquedas exploraría cada documento en el cuerpo, lo cual requiriría un tiempo y poder de cómputo considerables.

http://en.wikipedia.org/wiki/Index_(search_engine)

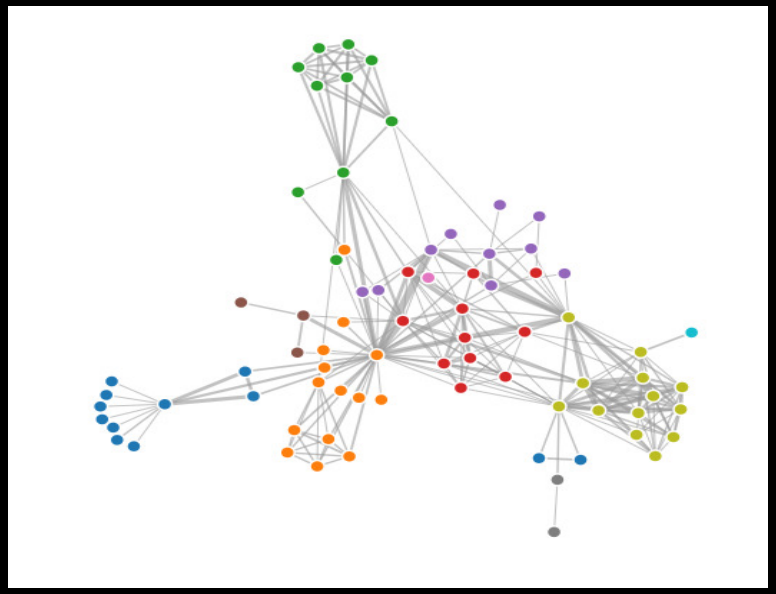
# D3.js

“D3.js es una biblioteca de JavaScript para manipular documentos basados en datos. D3 ayuda a dar vida a los datos usando HTML, SVG y CSS. El énfasis de D3 en los estándares web le brinda todas las capacidades de los navegadores modernos sin atarse a un marco propietario, que combina componentes de visualización potentes y un enfoque basado en datos para la manipulación DOM.”

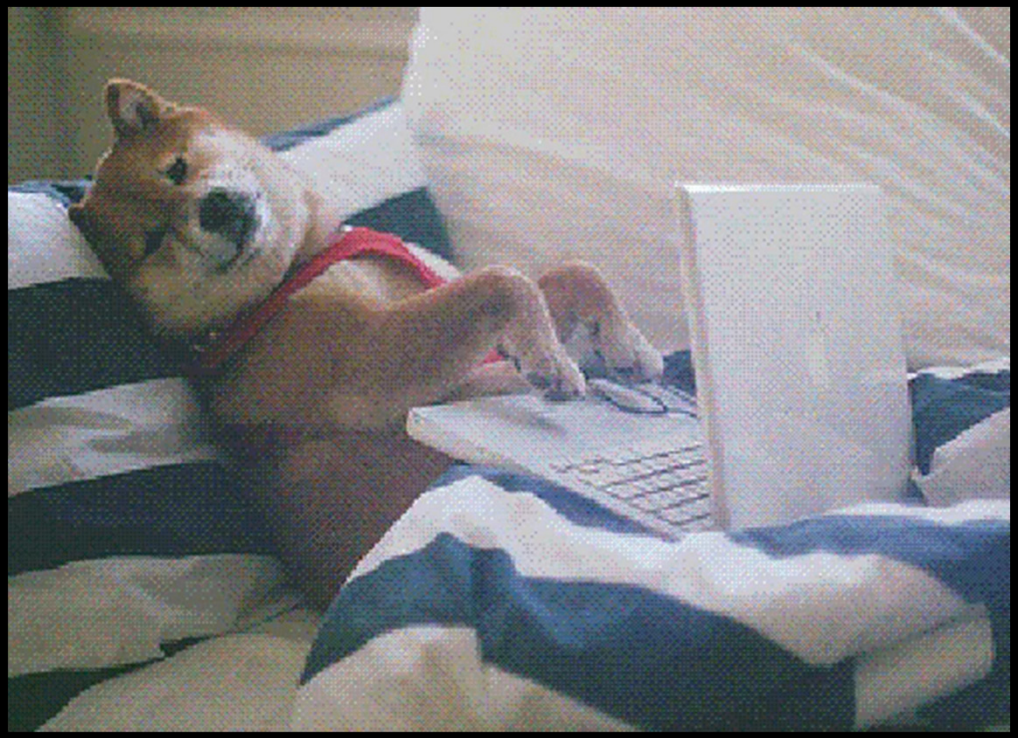
https://d3js.org/

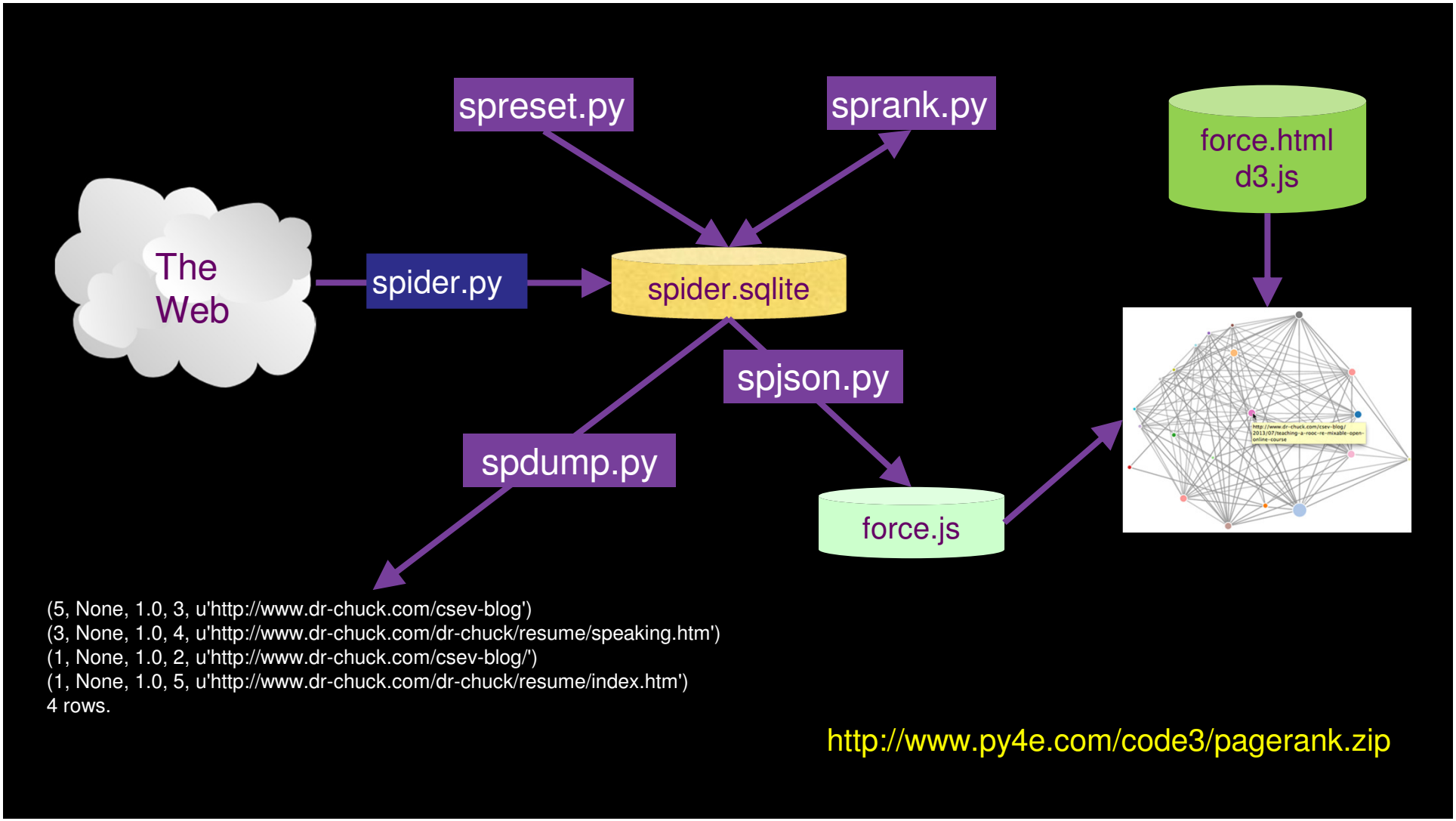
# Aplicaciones en grafos dirigidos

Red de co-ocurrencias entre caracteres en la novela “Los Miserables”



https://observablehq.com/@d3/force-directed-graph



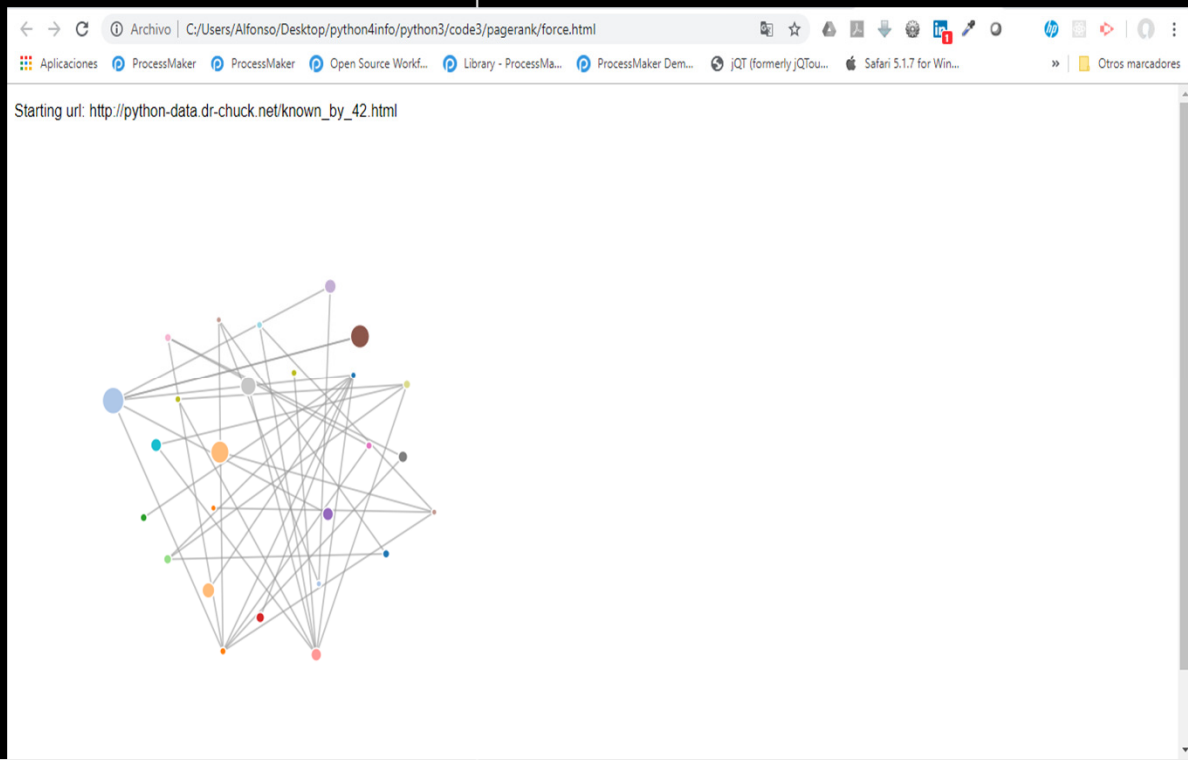



(5, None, 1.0, 3, u'http://www.dr-chuck.com/csev-blog')
(3, None, 1.0, 4, u'http://www.dr-chuck.com/dr-chuck/resume/speaking.htm')
(1, None, 1.0, 2, u'http://www.dr-chuck.com/csev-blog/')
(1, None, 1.0, 5, u'http://www.dr-chuck.com/dr-chuck/resume/index.htm')
4 rows.

http://www.py4e.com/code3/pagerank.zip

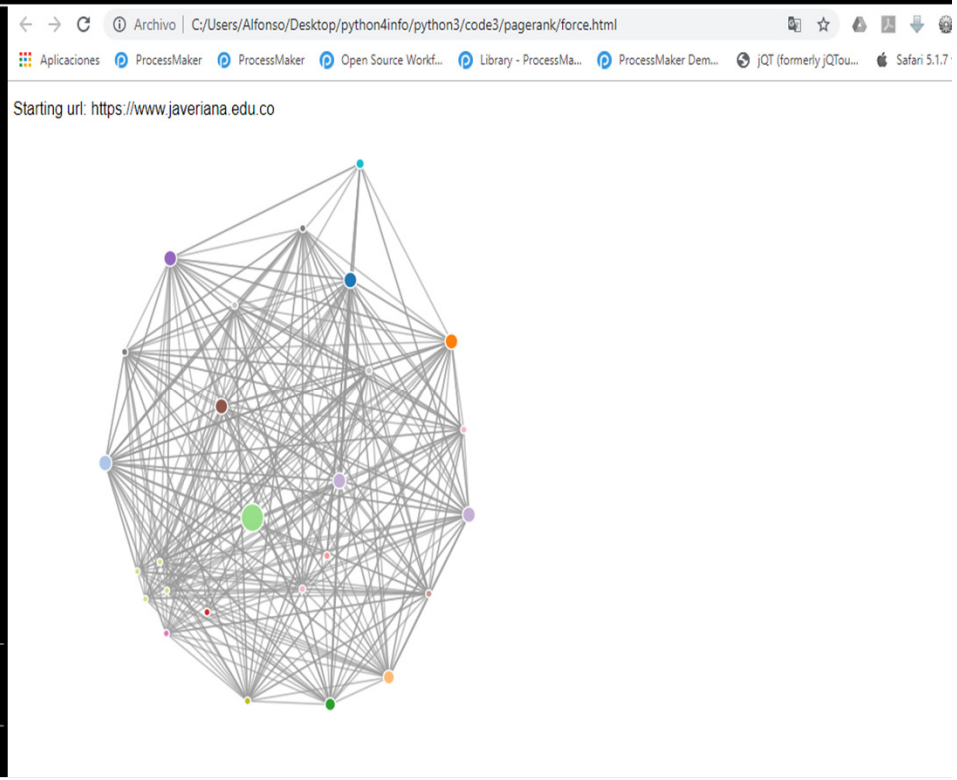
# Prueba del PageRank para la URL http://python-data.dr-chuck.net

```
99 1.6076249242715756e-17
100 9.480864938011857e-18
[(1, 0.07935445138366898), (7, 0.1587089276733796), (20, 0.8699724595084117), (137, 0.20363623131344208), (187, 0.20363623131344208)]

C:\Users\Alfonso\Desktop\python4info\python3\code3\pagerank>python3 spdump.py
(6, 2.497974849443514, 2.4979748494434912, 167, 'http://python-data.dr-chuck.net/known_by_Caitaidh.html')
(6, 2.0063914094917608, 2.006391409491774, 217, 'http://python-data.dr-chuck.net/known_by_Peebles.html')
(5, 1.4480960457637058, 1.4480960457636758, 8, 'http://python-data.dr-chuck.net/known_by_Miranne.html')
(5, 3.158203172498137 4, 3.15820317249812, 545, 'http://python-data.dr-chuck.net/known_by_Joojo.html')
(4, 1.2335786943101776, 1.2335786943101805, 424, 'http://python-data.dr-chuck.net/known_by_Oata.html')
(4, 1.6480236818689624, 1.6480236818689495, 730, 'http://python-data.dr-chuck.net/known_by_Charly.html')
(4, 2.835276303045442, 2.835276303045417, 987, 'http://python-data.dr-chuck.net/known_by_Josese.html')
(4, 2.2881799417389597, 2.2881799417389543, 1282, 'http://python-data.dr-chuck.net/known_by_Constance.html')
(4, 4.3395634447773513, 4.3395634447773419, 1342, 'http://python-data.dr-chuck.net/known_by_Ziya.html')
(4, 1.9922835032639936, 1.9922835032639936, 2093, 'http://python-data.dr-chuck.net/known_by_Denise.html')
(4, 1.8624241595669817, 1.8624241595566972, 2206, 'http://python-data.dr-chuck.net/known_by_Ismaeel.html')
(3, 0.3444103408649136, 0.3444103408649123, 149, 'http://python-data.dr-chuck.net/known_by_Abbeygail.html')
(3, 1.6811212196759728, 1.6811212196760015, 164, 'http://python-data.dr-chuck.net/known_by_Khyla.html')
(3, 3.085221427230209, 3.085221427230229, 307, 'http://python-data.dr-chuck.net/known_by_Terry.html')
(3, 1.5113028237495718, 1.5113028237495914, 321, 'http://python-data.dr-chuck.net/known_by_Jaymi.html')
(3, 0.6091310439689915, 0.6091310439689912, 354, 'http://python-data.dr-chuck.net/known_by_Patricia.html')
(3, 1.4981198898879757, 1.4981198898880081, 357, 'http://python-data.dr-chuck.net/known_by_Miller.html')
(3, 1.5695226580653214, 1.569522658065301, 488, 'http://python-data.dr-chuck.net/known_by_Jing.html')
(3, 1.9086265377924811, 1.9086265377924471, 507, 'http://python-data.dr-chuck.net/known_by_Samanta.html')
(3, 0.8687066123785291, 0.8687066123785212, 538, 'http://python-data.dr-chuck.net/known_by_Primrose.html')
(3, 0.7384613274892983, 0.7384613274892806, 595, 'http://python-data.dr-chuck.net/known_by_Shawnpaul.html')
(3, 0.7502507040641068, 0.7502507040641102, 598, 'http://python-data.dr-chuck.net/known_by_Aliza.html')
(3, 1.6342186170851623, 1.6342186170851405, 695, 'http://python-data.dr-chuck.net/known_by_Ioanna.html')
(3, 1.2935831329843372, 1.2935831329843266, 1186, 'http://python-data.dr-chuck.net/known_by_Clio.html')
(3, 1.3828509371082147, 1.3828509371082198, 1238, 'http://python-data.dr-chuck.net/known_by_Kristoffer.html')
(3, 3.503973288820543, 3.5039732888205166, 1426, 'http://python-data.dr-chuck.net/known_by_Chelsey.html')
(3, 0.925661434404678, 0.9256614344046692, 1529, 'http://python-data.dr-chuck.net/known_by_Saiba.html')
(3, 1.4194292180343904, 1.4194292180834385, 1567, 'http://python-data.dr-chuck.net/known_by_Kadin.html')
(3, 1.910817490097222, 1.910817490097201, 1644, 'http://python-data.dr-chuck.net/known_by_Dilano.html')
(3, 0.9951473409301005, 0.9951473409301028, 1842, 'http://python-data.dr-chuck.net/known_by_Tristain.html')
(3, 0.9499792280923292, 0.9499792280923127, 2449, 'http://python-data.dr-chuck.net/known_by_Kodi.html')
(3, 2.589854837527781, 2.5898548375278367, 2609, 'http://python-data.dr-chuck.net/known_by_Luella.html')
(3, 1.3289246970699113, 1.3289246970699053, 3165, 'http://python-data.dr-chuck.net/known_by_Ander.html')
(3, 2.855752325228588, 2.855752325228568, 3296, 'http://python-data.dr-chuck.net/known_by_Kori.html')
(3, 2.6457840142542155, 2.6457840142542572, 3524, 'http://python-data.dr-chuck.net/known_by_Tyllor.html')
(3, 1.5719400605373084, 1.5719400605372968, 4015, 'http://python-data.dr-chuck.net/known_by_Daisy.html')
(2, 0.8699724595084274, 0.8699724595084117, 20, 'http://python-data.dr-chuck.net/known_by_Loudon.html')
(2, 0.6534634660675016, 0.6534634660675178, 125, 'http://python-data.dr-chuck.net/known_by_Harris.html')
(2, 0.31475061428195, 0.31475061428195, 159, 'http://python-data.dr-chuck.net/known_by_Caie.html')
(2, 0.14161727908655408, 0.14161727908655533, 273, 'http://python-data.dr-chuck.net/known_by_Haillie.html')
(2, 0.15825819673809394, 0.15825819673809266, 280, 'http://python-data.dr-chuck.net/known_by_Millie.html')
(2, 0.2823981327066626, 0.2823981327066666, 454, 'http://python-data.dr-chuck.net/known_by_Suvi.html')
(2, 0.5865834096152125, 0.5865834096152523, 492, 'http://python-data.dr-chuck.net/known_by_Harry.html')
(2, 0.6383540272340914, 0.6383540272340098, 542, 'http://python-data.dr-chuck.net/known_by_Mehmet.html')
(2, 0.8530121861453319, 0.8530121861453168, 841, 'http://python-data.dr-chuck.net/known_by_Forbes.html')
(2, 0.7377150405356286, 0.7377150405356301, 870, 'http://python-data.dr-chuck.net/known_by_Charlize.html')
(2, 0.6183719249417226, 0.6183719249417226, 887, 'http://python-data.dr-chuck.net/known_by_Shilpa.html')
(2, 1.7183690761606418, 1.7183690761606265, 1004, 'http://python-data.dr-chuck.net/known_by_Kalen.html')
(2, 1.4675261269292803, 1.4675261269292803, 1024, 'http://python-data.dr-chuck.net/known_by_Avinash.html')
(2, 0.2339474326191034, 0.2339474326190962, 1104, 'http://python-data.dr-chuck.net/known_by_Fatiha.html')
100 rows.
```



# Prueba del PageRank para la URL https://www.javeriana.edu.co

```
[(1, 0.0013867186293981485), (24, 0.0001270344231980925 8), (2, 0.0014083354121141832), (101, 0.0001601821500392914 2), (48, 0.0013543417236851405)]

C:\Users\Alfonso\Desktop\python4info\python3\code3\pagerank>python3 spdump.py
146, 0.006514140741757064, 0.0014083354121141832, 2, 'https://www.javeriana.edu.co/educon')
100, 23.886879451437462, 23.901383663703 68, 246, 'https://www.javeriana.edu.co/pesquisa/perfiles/huellas')
100, 21.232792244882974, 21.245676644001068, 247, 'https://www.javeriana.edu.co/pesquisa/perfiles/entrevistas')
94, 0.006414154073833761, 0.0013867186293981485, 1, 'https://www.javeriana.edu.co')
87, 0.006321199151121728, 0.0013666220864186203, 46, 'https://www.javeriana.edu.co/programas/posgrados')
74, 0.004277408870995728, 0.0009247614726249122, 15, 'https://www.javeriana.edu.co/extension')
65, 0.006205473717999223, 0.0013416026353487 86, 4, 'https://www.javeriana.edu.co/editorial')
60, 0.004869242742863759, 0.0018527139736375471, 13, 'https://www.javeriana.edu.co/institucional/financiacion-becas')
60, 0.006264397333517037, 0.0013543417236851405, 48, 'https://www.javeriana.edu.co/programas/asignaturas-de-libre-escogencia')
60, 0.006264397333517037, 0.0013543417236851405, 49, 'https://www.javeriana.edu.co/programas/listado-de-cursos-preuniversitarios')
60, 0.006265704725462106, 0.0013282459309945168, 51, 'https://www.javeriana.edu.co/institucional/centros-institutos-y-observatorios')
53, 4.5499758631111828, 4.552664843442692, 273, 'https://www.javeriana.edu.co/pesquisa/hipopotamos-en-colombia-un-problema-de-enormes-dimensiones')
51, 23.886879451437466, 23.901383663703 68, 244, 'https://www.javeriana.edu.co/pesquisa/perfiles')
40, 0.000744155441761935 9, 0.0001608819142907155, 104, 'https://www.javeriana.edu.co/educon/politica-y-sociedad')
35, 0.000740910413458462 4, 0.0001601823500392914 2, 108, 'https://www.javeriana.edu.co/educon/comunicacion-y-lenguaje')
34, 0.000740910413458462 4, 0.0001601823500392914 2, 101, 'https://www.javeriana.edu.co/educon/artes')
32, 0.000736982497591842 9, 0.0001593331477837 45, 89, 'https://www.javeriana.edu.co/educon/salud')
32, 0.000736982497591842 9, 0.0001593331477837 45, 98, 'https://www.javeriana.edu.co/educon/estudios-ambientales-y-rurales')
30, 0.000732130783959335 8, 0.0001582842235450656 4, 88, 'https://www.javeriana.edu.co/educon/ciencias-juridicas')
30, 0.000732130783959335 7, 0.0001582842235450656 4, 112, 'https://www.javeriana.edu.co/educon/region-oriente')
29, 0.000734691353067356 7, 0.0001588781001072622, 70, 'https://www.javeriana.edu.co/web/educon/programacion')
26, 0.002503703651652 98, 0.0005412923444426421, 130, 'https://www.javeriana.edu.co/web/programas/especializaciones')
25, 0.002503703651652 98, 0.0005412923444426421, 131, 'https://www.javeriana.edu.co/web/programas/maestrias')
25, 0.001893291405137 6848, 0.0004093232610513 5, 159, 'https://www.javeriana.edu.co/programas/especializaciones')
23, 0.002524714776862 336, 0.0005458348793455247, 127, 'https://www.javeriana.edu.co/web/programas/home')
21, 0.042517402174198 433, 0.0042797470223 23405, 261, 'https://www.javeriana.edu.co/pesquisa/category/multimedia')
18, 0.021900404088632 724, 0.0219063749241 8162, 640, 'https://www.javeriana.edu.co/pesquisa/tag/medio-ambiente')
16, 0.000579071969653 0814, 0.0001251934204947 4528, 165, 'https://www.javeriana.edu.co/admisiones/admisiones-paso-a-paso-2')
15, 0.117400263438678 3, 0.1174708808203761, 547, 'https://www.javeriana.edu.co/pesquisa/tag/peces')
15, 0.097445430581966, 0.0975054851703533 9, 898, 'https://www.javeriana.edu.co/pesquisa/tag/ecosistema')
14, 0.000377436441100242 25, 8.160049451044834e-05, 480, 'https://www.javeriana.edu.co/web/institucional/pos-icetex')
14, 0.017698416199915 686, 0.0177091419737547 93, 711, 'https://www.javeriana.edu.co/pesquisa/tag/memoria')
13, 0.075384690294638 44, 0.0753687835880208, 275, 'https://www.javeriana.edu.co/pesquisa/tag/ingenieria')
12, 15.925098868658797, 15.934842128752 19, 301, 'https://www.javeriana.edu.co/pesquisa/mas-alla-de-lo-evidente')
9, 0.003624571054974624 7, 0.0031613077707121308, 18, 'https://www.javeriana.edu.co/pesquisa/antes-de-votar-revise-a-que-se-pueden-comprometer-los-candidatos')
9, 0.000429679737539109 24, 9.2895328234739666e-05, 78, 'https://www.javeriana.edu.co/educon/apocaliptica-judeo-cristiana-y-de-san-juan')
9, 0.001099208937351513, 0.0002376452909429327, 147, 'https://www.javeriana.edu.co/vicerrectoria-academica/home')
9, 7.151611965209598e-05, 1.546154555784559 3e-05, 204, 'https://www.javeriana.edu.co/web/educon/google-marketing-platform')
9, 7.151611965209598e-05, 1.546154555784559 3e-05, 212, 'https://www.javeriana.edu.co/web/educon/proyectos-de-ficcion-y-no-ficcion')
9, 7.151611965209598e-05, 1.546154555784559 3e-05, 214, 'https://www.javeriana.edu.co/web/educon/aplicacion-de-las-rda-en-bibliotecas')
8, 0.008376094033473214, 0.0083743130351262 33, 643, 'https://www.javeriana.edu.co/pesquisa/la-era-de-la-hegemonia-cuantica')
7, 0.000468577813792847 1, 0.0001012919822498043 8, 190, 'https://www.javeriana.edu.co/admisiones/registro')
7, 3.599345266822328 5e-05, 7.781663923396486e-06, 492, 'https://www.javeriana.edu.co/web/educon/acompanamiento-en-el-duelo')
7, 3.408526982635822e-05, 7.369121183583553 46e-06, 496, 'https://www.javeriana.edu.co/web/educon/politica-y-sociedad?p_p_id=101_INSTANCE_641NTVqGhBK8&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_pos=2&p_p_col_count=3&p_r_p_564233524_categoryId=5707352')
7, 0.001036990367263726, 0.0010342733475146866, 569, 'https://www.javeriana.edu.co/pesquisa/los-inborrables-anos-70')
7, 0.000863152190077905, 0.0008635248930369 7, 582, 'https://www.javeriana.edu.co/pesquisa/tag/presupuesto')
7, 0.008376094033473214, 0.0083743130351262 33, 652, 'https://www.javeriana.edu.co/pesquisa/medicina-e-ingenieria-se-unen-para-salvar-vidas')
7, 0.002528330143636 51026, 0.0025290747 5783776598, 835, 'https://www.javeriana.edu.co/pesquisa/tag/campesinos')
6, 3.743084982511144e-05, 8.092424374828674e-06, 210, 'https://www.javeriana.edu.co/web/educon/comunicacion-y-lenguaje?p_p_id=101_INSTANCE_641NTVqGhBK8&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_pos=1&p_p_col_count=2&p_r_p_564233524_categoryId=5707352')
6, 0.458554477664786114, 0.4588287184248231 5, 388, 'https://www.javeriana.edu.co/pesquisa/tag/agrocadena')
100 rows.
```



# Referencias y recursos adicionales

- https://www.py4e.com/html3/16-viz
- http://www.dr-chuck.com/
- https://www.youtube.com/playlist?list=PLIRFEj9H3Oj7Bp8-DfGpfAfDBiblRfl5p
- http://infolab.stanford.edu/~backrub/google.html
- https://d3js.org/
- https://github.com/d3/d3-force
- https://observablehq.com/@d3/force-directed-graph
- https://bl.ocks.org/mbostock/ad70335eeef6d167bc36fd3c04378048
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- https://www.statisticshowto.datasciencecentral.com/absolute-error/

# Acknowledgements / Contributions



Initial Development: Charles Severance, University of Michigan School of Information