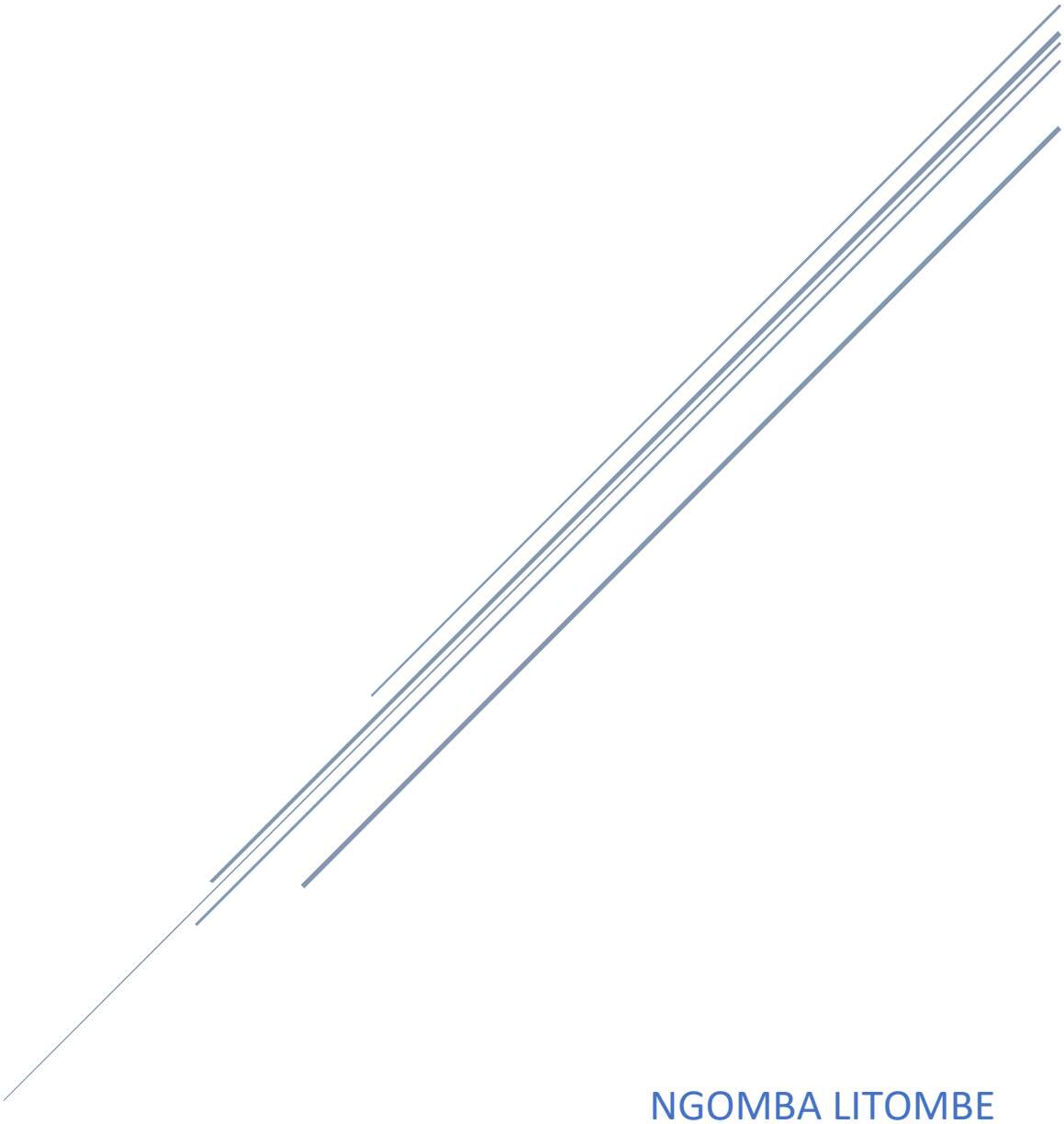
BIG DATA MANAGEMENT

PROJECT 3



NGOMBA LITOMBE
SN1199003

11/06/2020

Page Rank

Page rank is an algorithm developed by L. Page and S. Brin of google for ranking the importance of pages on the web. Generally speaking, the importance of a page is directly proportional to the importance of the pages pointing to it. This is the rationale behind this very important algorithm.

As a motivation for this algorithm I present to you a part of the abstract from the paper “Bringing order to the Web” by L. Page in 1998:

” The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.”

In this project, I have made use of the web-Google directed graph dataset from the website https://snap.stanford.edu/data/web-Google.html .

a) Implementation of Simple Page rank version

After trying out the matrix implementation of the simple page rank and had a memory error as it appears my computer RAM was quite small for such an implementation, I went ahead with the adjacency list implementation. I made use of the following algorithm:

1. Initialize the page rank of all the pages (nodes in the graph) to 1.

2. For every page compute the corresponding page rank with the following formula

$$\text{Page rank Pi} = \sum_{j} \frac{\text{Page rank of inbound link page Pj}}{\text{number of out links of page Pj}}$$

3. Perform step 2, for 10 iterations and stop.

I obtained the following results

```
Here are the top 20 ranked pages
(Node id, Page rank)
(41909, 621.5753647305577)
(597621, 511.8590993857555)
(384666, 506.82871150133934)
(537039, 480.6446922251635)
(504140, 441.1817837381442)
(163075, 437.8928546254028)
(486980, 413.4784186204205)
(905628, 410.71341133984095)
(32163, 402.24175881814443)
(765334, 400.4295039294865)
(558791, 397.14369168630765)
(452291, 358.7973739951857)
(425770, 355.2838661923785)
(1536, 346.77439302801565)
(605856, 330.57718676575644)
(577518, 323.13527951025253)
(226374, 320.92041282846355)
(751384, 320.1385368541584)
(173976, 316.9250303512692)
(691633, 313.26850611581204)
```

```
Here are the bottom 20  ranked pages
(Node id, Page rank)
(916027, 0)
(916040, 0)
(916061, 0)
(916062, 0)
(916071, 0)
(916103, 0)
(916114, 0)
(518803, 0.0)
(916169, 0)
(720530, 0.0)
(916208, 0)
(916235, 0)
(122865, 0.0)
(916244, 0)
(208701, 0.0)
(916267, 0)
(916296, 0)
(916344, 0)
(916348, 0)
(916425, 0)
```

Bottom pages with zero-page ranks are justified because there are some pages which have no inbound links (also known as back links).

b) Improved version of Page Rank

I achieved this by replacing step 2 of part a, above with the following formula:

$$PR(u) = (1 - d) + d \times \sum \frac{PR(v)}{N(v)}$$

This is equivalent to the following in the matrix formulation(a is the same as d):

$$p = \alpha \cdot M \cdot p + (1-\alpha) \cdot I_N$$

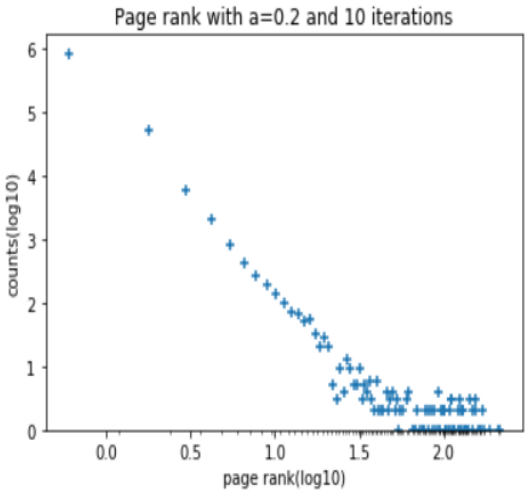
The variable a (same as d) is known as the damping factor and determines how quickly the page rank algorithm converges.

The following results were obtained for a = 0.2 still using 10 iterations as in part a.

The histogram shows the log number of pages within a particular page rank range. It is noticeable that there more pages with lower page ranks than with higher page ranks.

Here are the top 20 ranked pages
(Node id, Page rank)
(163075, 239.7135440495773)
(537039, 215.12273130880982)
(597621, 209.36019026395223)
(605856, 186.45527729843033)
(885605, 173.7177568946275)
(819223, 171.01778041600554)
(551829, 170.92266595014627)
(751384, 166.3617121981291)
(908351, 164.08661028556196)
(504140, 162.40386363071443)
(32163, 156.29248653011942)
(837478, 155.58640871038648)
(558791, 153.7435501113081)
(765334, 153.1970002169146)
(173976, 152.9221813469781)
(41909, 149.38821902361704)
(828963, 147.85859169569963)
(213432, 147.73373376876648)
(384666, 144.53067326027374)
(7314, 143.70689519720952)

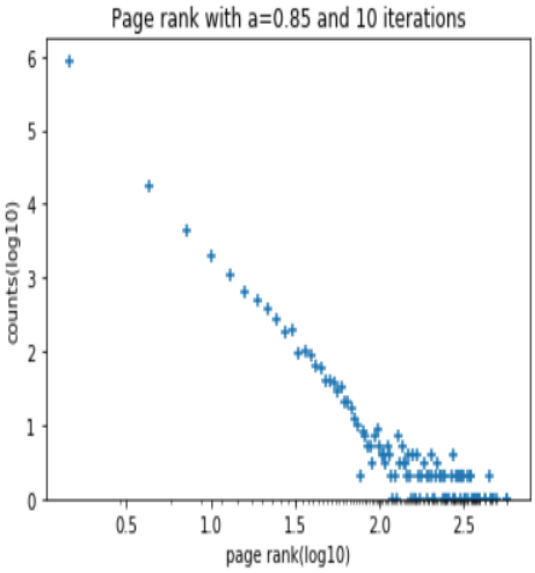
Here are the bottom 20 ranked pages
(Node id, Page rank)
(915960, 0.8)
(915964, 0.8)
(915987, 0.8)
(915990, 0.8)
(916027, 0.8)
(916040, 0.8)
(916061, 0.8)
(916062, 0.8)
(916071, 0.8)
(916103, 0.8)
(916114, 0.8)
(916169, 0.8)
(916208, 0.8)
(916235, 0.8)
(916244, 0.8)
(916267, 0.8)
(916296, 0.8)
(916344, 0.8)
(916348, 0.8)
(916425, 0.8)



The following results were obtained for a = 0.85

Here are the top 20 ranked pages
(Node id, Page rank)
(41909, 578.9930904657566)
(597621, 578.8650333658389)
(163075, 570.0201694513328)
(537039, 563.7648837381201)
(384666, 493.0157332133267)
(504140, 480.3695835056029)
(486980, 465.5881292579237)
(605856, 455.90287350110054)
(558791, 452.48266272684765)
(32163, 448.4184361287363)
(551829, 446.4818440538285)
(765334, 424.2569268602502)
(751384, 419.5439949902164)
(425770, 393.6115500968873)
(908351, 388.24637641149536)
(173976, 387.2074457942461)
(7314, 378.7485245528858)
(213432, 376.09586912177184)
(885605, 371.504217240319)
(691633, 369.0218505655483)

Here are the bottom 20 ranked pages
(Node id, Page rank)
(915960, 0.15)
(915964, 0.15)
(915987, 0.15)
(915990, 0.15)
(916027, 0.15)
(916040, 0.15)
(916061, 0.15)
(916062, 0.15)
(916071, 0.15)
(916103, 0.15)
(916114, 0.15)
(916169, 0.15)
(916208, 0.15)
(916235, 0.15)
(916244, 0.15)
(916267, 0.15)
(916296, 0.15)
(916344, 0.15)
(916348, 0.15)
(916425, 0.15)



The minimum page rank changes as there is now the (1-a) term in the formula.

c) Repeating parts a and b while noting the run time for iterations of 50,100 and 200.

Repeating part a

Number of iterations = 50

Run time = 117.50s

```
Here are the top 20 ranked pages
(Node id, Page rank)
(747106, 497.32010287512907)
(24576, 483.3484274247061)
(370344, 483.3484274247061)
(544138, 483.3484274247061)
(577518, 337.6675506696786)
(587617, 294.1477329673495)
(671168, 261.55556369857743)
(873996, 227.85707111642571)
(914474, 227.4999999999999)
(791675, 217.36909297453136)
(905628, 211.34752178766087)
(41909, 204.68365450823222)
(765334, 197.8294362101575)
(714416, 194.04124510719666)
(627251, 193.04124510719666)
(1536, 185.71007864960612)
(699629, 165.6024748940495)
(232639, 164.63167820182838)
(384666, 163.97025916792066)
(597621, 160.848095081745)
```

```
Here are the bottom 20  ranked pages
(Node id, Page rank)
(916027, 0)
(916040, 0)
(916061, 0)
(916062, 0)
(916071, 0)
(916103, 0)
(916114, 0)
(518803, 0.0)
(916169, 0)
(720530, 0.0)
(916208, 0)
(916235, 0)
(122865, 0.0)
(916244, 0)
(208701, 0.0)
(916267, 0)
(916296, 0)
(916344, 0)
(916348, 0)
(916425, 0)
```

Iterations = 100

Run time = 252.18s

```
Here are the top 20 ranked pages
(Node id, Page rank)
(747106, 611.4868050782524)
(24576, 608.138576723503)
(370344, 608.138576723503)
(544138, 608.138576723503)
(577518, 341.7061878912347)
(587617, 316.46305659311344)
(671168, 281.30959151804467)
(791675, 269.992458040024)
(873996, 246.04591734264295)
(714416, 245.4843958566759)
(627251, 244.4843958566759)
(914474, 227.4999999999999)
(699629, 190.78845417426024)
(463515, 174.77905023444495)
(756523, 170.55974421530186)
(131655, 165.83144153908538)
(908695, 157.7604120912928)
(43792, 151.41315789473683)
(203907, 151.41315789473683)
(375897, 151.41315789473683)
```

```
Here are the bottom 20  ranked pages
(Node id, Page rank)
(916027, 0)
(916040, 0)
(916061, 0)
(916062, 0)
(916071, 0)
(916103, 0)
(916114, 0)
(518803, 0.0)
(916169, 0)
(720530, 0.0)
(916208, 0)
(916235, 0)
(122865, 0.0)
(916244, 0)
(208701, 0.0)
(916267, 0)
(916296, 0)
(916344, 0)
(916348, 0)
(916425, 0)
```

Iterations = 200
Run time = 483.52s

```
(Node id, Page rank)
(747106, 648.9105763238387)
(24576, 648.6045794050914)
(370344, 648.6045794050914)
(544138, 648.6045794050914)
(577518, 342.6451628696995)
(587617, 320.3107291086494)
(791675, 286.8083518026193)
(671168, 284.72140387843297)
(714416, 262.0976231015253)
(627251, 261.09762310152536)
(873996, 249.12297032174737)
(914474, 227.4999999999999)
(699629, 202.54313581671292)
(463515, 187.47842214884312)
(756523, 183.4945383664572)
(131655, 177.34535756992406)
(908695, 163.65166356934577)
(802206, 157.65181603179053)
(877461, 155.97565358617499)
(385039, 155.28931335459723)
```

```
(Node id, Page rank)
(916027, 0)
(916040, 0)
(916061, 0)
(916062, 0)
(916071, 0)
(916103, 0)
(916114, 0)
(518803, 0.0)
(916169, 0)
(720530, 0.0)
(916208, 0)
(916235, 0)
(122865, 0.0)
(916244, 0)
(208701, 0.0)
(916267, 0)
(916296, 0)
(916344, 0)
(916348, 0)
(916425, 0)
```

Repeating part b

For a = 0.2

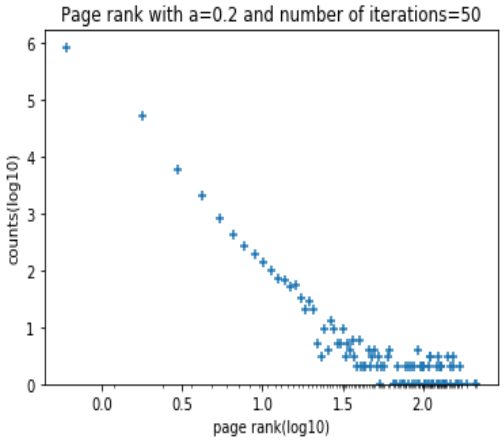
Iteration = 50

Run time = 276.83s

The run time in this section is the total time for making the page rank and the histogram.

```
Here are the top 20 ranked pages
(Node id, Page rank)
(163075, 239.71354349968868)
(537039, 215.1227308851778)
(597621, 209.3601898802077)
(605856, 186.45527674192718)
(885605, 173.71775638460414)
(819223, 171.0177801875008)
(551829, 170.92266534051276)
(751384, 166.36171169633457)
(908351, 164.08661003181385)
(504140, 162.40386329091695)
(32163, 156.2924861513348)
(837478, 155.5864087125079)
(558791, 153.74354950737126)
(765334, 153.19700003458033)
(173976, 152.9221809133189)
(41909, 149.38821860471074)
(828963, 147.8585912652099)
(213432, 147.73373338415036)
(384666, 144.5306729900653)
(7314, 143.7068948066325)
```

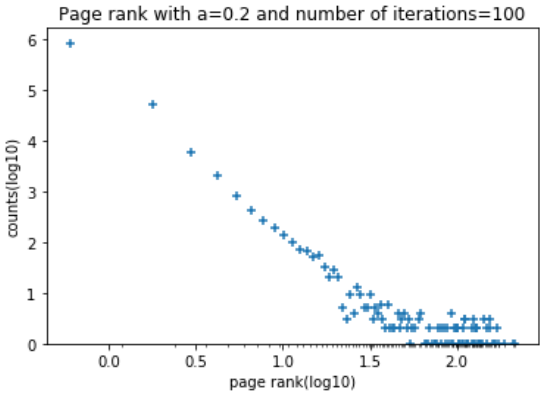
```
Here are the bottom 20 ranked pages
(Node id, Page rank)
(915960, 0.8)
(915964, 0.8)
(915987, 0.8)
(915990, 0.8)
(916027, 0.8)
(916040, 0.8)
(916061, 0.8)
(916062, 0.8)
(916071, 0.8)
(916103, 0.8)
(916114, 0.8)
(916169, 0.8)
(916208, 0.8)
(916235, 0.8)
(916244, 0.8)
(916267, 0.8)
(916296, 0.8)
(916344, 0.8)
(916348, 0.8)
(916425, 0.8)
```



Iteration = 100

Run time =518.88s

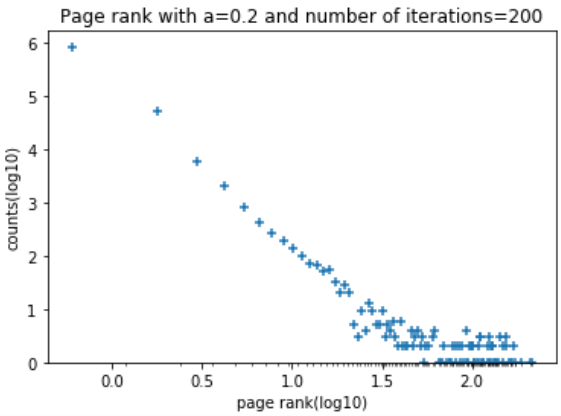
```
Here are the top 20 ranked pages Here are the bottom 20 ranked pages
(Node id, Page rank)             (Node id, Page rank)
(163075, 239.71354349968868)     (915960, 0.8)
(537039, 215.1227308851778)      (915964, 0.8)
(597621, 209.3601898802077)      (915987, 0.8)
(605856, 186.45527674192718)     (915990, 0.8)
(885605, 173.71775638460414)     (916027, 0.8)
(819223, 171.0177801875008)      (916040, 0.8)
(551829, 170.92266534051276)     (916061, 0.8)
(751384, 166.36171169633457)     (916062, 0.8)
(908351, 164.08661003181385)     (916071, 0.8)
(504140, 162.40386329091695)     (916103, 0.8)
(32163, 156.2924861513348)       (916114, 0.8)
(837478, 155.5864087125079)      (916169, 0.8)
(558791, 153.74354950737126)     (916208, 0.8)
(765334, 153.19700003458033)     (916235, 0.8)
(173976, 152.9221809133189)      (916244, 0.8)
(41909, 149.38821860471074)      (916267, 0.8)
(828963, 147.8585912652099)      (916296, 0.8)
(213432, 147.73373338415036)     (916344, 0.8)
(384666, 144.5306729900653)      (916348, 0.8)
(7314, 143.7068948066325)        (916425, 0.8)
```



Iteration = 200

Run time = 1012.80s

```
Here are the top 20 ranked pages Here are the bottom 20 ranked pages
(Node id, Page rank)             (Node id, Page rank)
(163075, 239.71354349968868)     (915960, 0.8)
(537039, 215.1227308851778)      (915964, 0.8)
(597621, 209.3601898802077)      (915987, 0.8)
(605856, 186.45527674192718)     (915990, 0.8)
(885605, 173.71775638460414)     (916027, 0.8)
(819223, 171.0177801875008)      (916040, 0.8)
(551829, 170.92266534051276)     (916061, 0.8)
(751384, 166.36171169633457)     (916062, 0.8)
(908351, 164.08661003181385)     (916071, 0.8)
(504140, 162.40386329091695)     (916103, 0.8)
(32163, 156.2924861513348)       (916114, 0.8)
(837478, 155.5864087125079)      (916169, 0.8)
(558791, 153.74354950737126)     (916208, 0.8)
(765334, 153.19700003458033)     (916235, 0.8)
(173976, 152.9221809133189)      (916244, 0.8)
(41909, 149.38821860471074)      (916267, 0.8)
(828963, 147.8585912652099)      (916296, 0.8)
(213432, 147.73373338415036)     (916344, 0.8)
(384666, 144.5306729900653)      (916348, 0.8)
(7314, 143.7068948066325)        (916425, 0.8)
```



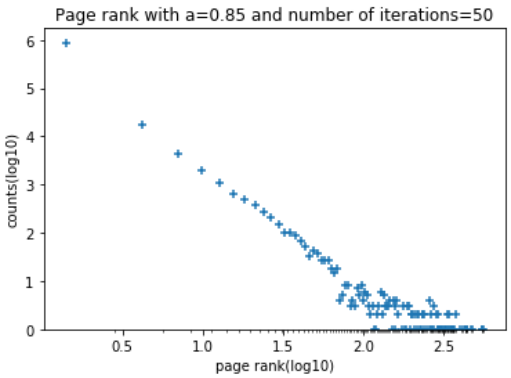
For a=0.85

Iteration = 50

Run time = 273.35s

```
 For iteration number = 50
Here are the top 20 ranked pages
(Node id, Page rank)
(597621, 564.2769578256215)
(41909, 562.6942690300694)
(163075, 552.2279634888794)
(537039, 549.0701276677146)
(384666, 480.69128858973625)
(504140, 467.38759292132585)
(486980, 442.84226756230595)
(605856, 438.5752490926591)
(32163, 435.2892454299863)
(558791, 433.2191060074905)
(551829, 428.8431468095391)
(765334, 417.2185866860103)
(751384, 403.90633077633936)
(425770, 380.58139048753264)
(908351, 379.2078282329449)
(173976, 372.10698355560396)
(7314, 365.65977707439606)
(213432, 363.6752640677462)
(885605, 358.6250414004875)
(819223, 355.6986731438794)
```

```
Here are the bottom 20 ranked pages
(Node id, Page rank)
(915960, 0.15)
(915964, 0.15)
(915987, 0.15)
(915990, 0.15)
(916027, 0.15)
(916040, 0.15)
(916061, 0.15)
(916062, 0.15)
(916071, 0.15)
(916103, 0.15)
(916114, 0.15)
(916169, 0.15)
(916208, 0.15)
(916235, 0.15)
(916244, 0.15)
(916267, 0.15)
(916296, 0.15)
(916344, 0.15)
(916348, 0.15)
(916425, 0.15)
```

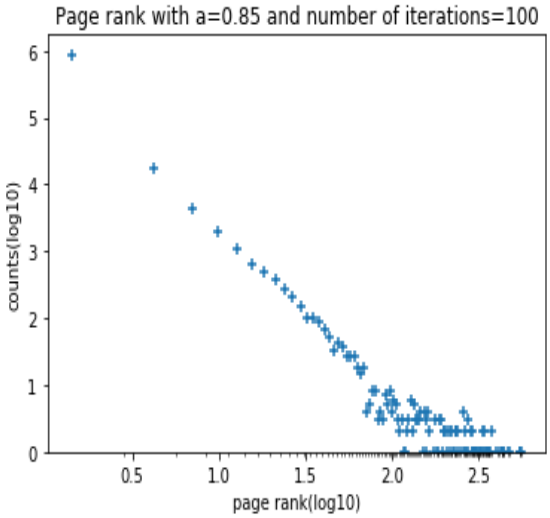


Iteration =100

Run time =521.73s

```
Here are the top 20 ranked pages
(Node id, Page rank)
(597621, 564.2705964625867)
(41909, 562.6861948488671)
(163075, 552.2240353559567)
(537039, 549.0642643603892)
(384666, 480.68450355701225)
(504140, 467.3820842916298)
(486980, 442.8401547545888)
(605856, 438.57324291128646)
(32163, 435.2847062514206)
(558791, 433.2163723471044)
(551829, 428.84158405298734)
(765334, 417.21319895341844)
(751384, 403.9040348934426)
(425770, 380.5776920000236)
(908351, 379.20434415911643)
(173976, 372.1049198919437)
(7314, 365.65701174339256)
(213432, 363.6723078403875)
(885605, 358.6246020131767)
(819223, 355.69581076930956)
```

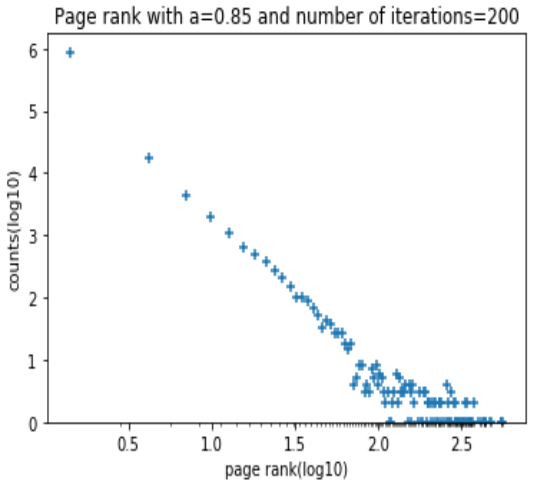
```
Here are the bottom 20 ranked pages
(Node id, Page rank)
(915960, 0.15)
(915964, 0.15)
(915987, 0.15)
(915990, 0.15)
(916027, 0.15)
(916040, 0.15)
(916061, 0.15)
(916062, 0.15)
(916071, 0.15)
(916103, 0.15)
(916114, 0.15)
(916169, 0.15)
(916208, 0.15)
(916235, 0.15)
(916244, 0.15)
(916267, 0.15)
(916296, 0.15)
(916344, 0.15)
(916348, 0.15)
(916425, 0.15)
```



Iteration = 200
Run time = 1020.70s

Here are the top 20 ranked pages
(Node id, Page rank)
(597621, 564.2705960061253)
(41909, 562.6861942533999)
(163075, 552.2240351484243)
(537039, 549.0642639464701)
(384666, 480.6845031027829)
(504140, 467.3820839345839)
(486980, 442.8401546995526)
(605856, 438.5732428633009)
(32163, 435.2847059205641)
(558791, 433.21637226218616)
(551829, 428.8415840009773)
(765334, 417.2131983052951)
(751384, 403.904034814504)
(425770, 380.5776918412439)
(908351, 379.20434390219225)
(173976, 372.1049198240531)
(7314, 365.657011622004)
(213432, 363.672307706025)
(885605, 358.6246019993965)
(819223, 355.69581055326654)

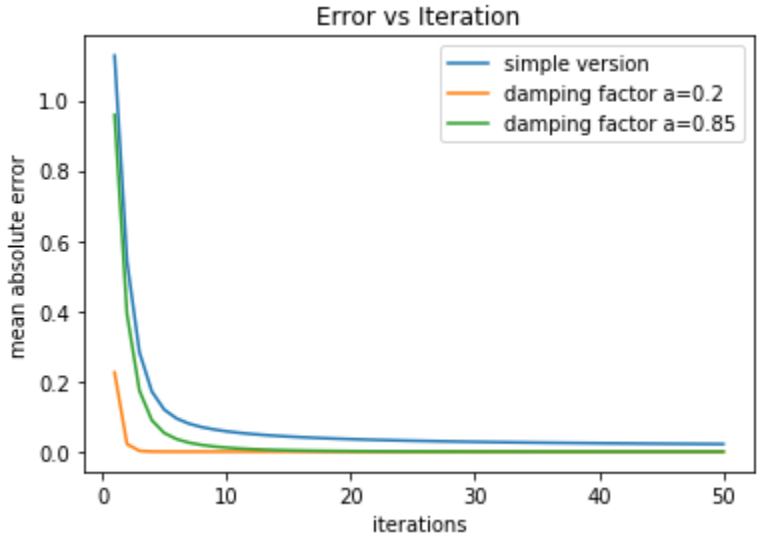
Here are the bottom 20 ranked pages
(Node id, Page rank)
(915960, 0.15)
(915964, 0.15)
(915987, 0.15)
(915990, 0.15)
(916027, 0.15)
(916040, 0.15)
(916061, 0.15)
(916062, 0.15)
(916071, 0.15)
(916103, 0.15)
(916114, 0.15)
(916169, 0.15)
(916208, 0.15)
(916235, 0.15)
(916244, 0.15)
(916267, 0.15)
(916296, 0.15)
(916344, 0.15)
(916348, 0.15)
(916425, 0.15)



Observation

As the number of iterations increases, the page rank scores do not change much and graph almost looks the same.

d) What is the point that you could stop the iterations?



Observations

1. The smaller the damping factor, the earlier convergence is achieved.
2. The simple version without damping converges at a higher error margin than the others with damping.

Observing that the iterations don't change much after a certain point, I could stop the iterations when the mean absolute differences in the page rank scores between two successive iterations (represented as error in the graph above) is below a certain threshold. That is, the algorithm converges.
For part a, which can also be considered as a model with damping factor a=1, the values continue to decrease generally more slowly than the other cases.
For part b, with damping a = 0.85, a sharper decrease is observed towards convergence. The model with a=0.2 has the sharpest decrease.

Based these observations, I set the threshold value for part a to 2e-2 and that for part b to 1e-3 as conditions for convergence.

How would you detect that in your code so you could stop the computations earlier?

I implemented this in my code by replacing step 3. of the main algorithm described above in part a and b with:

Step 3: if mean absolute error between page ranks of previous iteration and the current iteration is less than threshold, stop the algorithm.

An example for part b is shown in the little piece of code below.

```
old_rank=[k[1] for k in p_old.items()]
new_rank=[k[1] for k in p.items()]
if sum([abs(x-y) for x,y in zip(old_rank,new_rank)]) * (1/len(new_rank)) <= 1e-3:
    print('Converges at iteration number {}'.format(i))
    return p
```

With this error margin of 1e-3 for part b:
Using a=0.2, model converges after 3 iterations.
Using a=0.85, model converges after 21 iterations.

With error margin set to 2e-2 for part a, the model converged after 61 iterations

It is worth mentioning that despite the fact that the damping factor plays a great role in the convergence of the algorithm, a =0.85 is widely considered a sweet spot for the damping factor as ,a , also represents the probability that a random web surfer would follow the hyperlinks of a page instead of teleporting to another web page.