COEN 242: Big Data
Spring 2022
Final Project Report

PAGERANK

Team 12
Jil Patel (W1607219)
Malavika S Tankala (W1628801)
Mansi Tandel (W1606463)
Shubham Singla (W1628734)

Table Of Contents

Table Of Figures

Table Of Tables

1

1.INTRODUCTION & MOTIVATION

1.1 Motivation

A search engine, particularly Google, has become an indispensable component of our daily life. Almost anything and everything we want to know can be answered by just entering our question as an input and receiving multiple relevant results in a matter of milliseconds. Consequently, knowledge has become far more accessible than it has ever been before. Behind this responsive technology is a jumble of algorithms, the most important of which is PageRank, that work together to deliver the results in response to the search query. As a result, we decided to pursue this topic for our final project because of its inherent value in everyday life.

Moreover, the data that had to be processed was extensive and provided us with an opportunity for implementing parallel computation. In contrast to the traditional paradigm of local computing, the benefits of a distributed computing paradigm and the use of Spark in coding the required program were needed, which motivated us to implement this project.

As the PageRank Algorithm was developed by Google, we were inspired to use the Google Web-Graph dataset which spans over 875713 nodes and 5105039 edges. Due to the abundant number of nodes and edges we believed that we could obtain an in-depth understanding of the PageRank Algorithm in its base form but with a real time capacity that can help us relate close to how Google produces accurate results instantaneously. Another fact that caught our eye to utilize the Google Web-Graph dataset is that it was released by Google as part of the Google programming contest.

1.2 Early Search engines

Before Google, there were other search engines that mostly worked by crawling the Web and creating an inverted index of the keywords (words or other strings of characters other than white space) found on each page. An inverted index is a data structure that makes it simple to find (points to) all the places where a term appears when given a term.

When a search query (a list of terms) was issued, the pages containing those terms were pulled from the inverted index and sorted according to how the phrases were used on the page. Thus, the existence of a term in a page's header made the page more relevant than the presence of the term in ordinary text, and large fonts made the page more relevant.

So many people took advantage of such a term and started fooling search for their individual benefit. So if you have your website published on the internet the only thing you care about is to attract more people to your website. Now, if you add more political words or more adventurous words then when people will search articles related to politics they will see your website as their first search result. **Term spam** is a technique for deceiving search engines into thinking your page is about something it isn't. The ease with which phrase spammers might operate rendered early search engines nearly unusable.

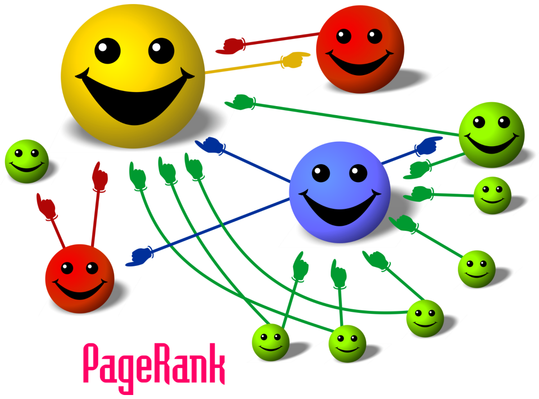


Figure 1: What is Google PageRank?

To overcome term spam, Google introduced two innovations:

1. **PageRank** was used to model where Web surfers would likely cluster if they started at a random website and followed randomly chosen outlinks from that page, and this process was allowed to iterate several times. Pages with a big number of visitors were deemed more "important" than pages with a small number of visitors. When it comes to determining which pages to show first in response to a search query, Google favors important pages over irrelevant pages.

2. **The content of a page was judged not only by the terms appearing on that page, but by the terms used in or near the links to that page.** Note that while it

is easy for a spammer to add false terms to a page they control, they cannot as easily get false terms added to the pages that link to their own page, if they do not control those pages.

1.3 PageRank Definition

The PageRank algorithm or Google algorithm was introduced by Larry Page and Sergey Brin, the founders of Google, to rank web pages and it is known as the heart of the Google search engine. The PageRank algorithm generates a probability distribution that is used to reflect the possibility of a random user clicking on links and ending up on a specific page. It can be used on any size document and the resulted distribution is initially evenly distributed among all pages in the collection.

The PageRank algorithm gives each page a rating of its importance, which is a recursively defined measure whereby a page becomes important if important pages link to it. This definition is recursive because the importance of a page refers back to the importance of other pages that link to it.

2

2.OBJECTIVE

- To understand the PageRank Algorithm.
- To implement the PageRank Algorithm using two methods:
    - Power Iteration
    - Naive Implementation.
- To compare the time taken to compute the page ranks of the node :
    - with partitions
    - without partitions
- To analyze if a node with most incoming nodes has the highest rank.
- To analyze the convergence value over multiple partitions and iterations of the Page Rank Algorithm.
- To calculate the top and bottom 10 web pages using the PageRank Algorithm.
- To analyze the impact of spider traps and dead ends.

3

3. ALGORITHM

For the pagerank algorithm we transform the outgoing links from each website into a webgraph. This webgraph will help us to calculate the page rank for each website. Here, each website represents a unique node in a web graph and links present in that website are the outgoing links from that node to the other node. We will look into type of implementation, One is the power iteration and other is the naive implementation.

3.1 Power Iteration:

In the power iteration we assign the probability to each unique node as 1/n where n is the total number of unique nodes in the webgraph. The reason behind this is we assume that each website gets equal probability and based on that we further calculate the probability (pagerank) until it reaches convergence value.

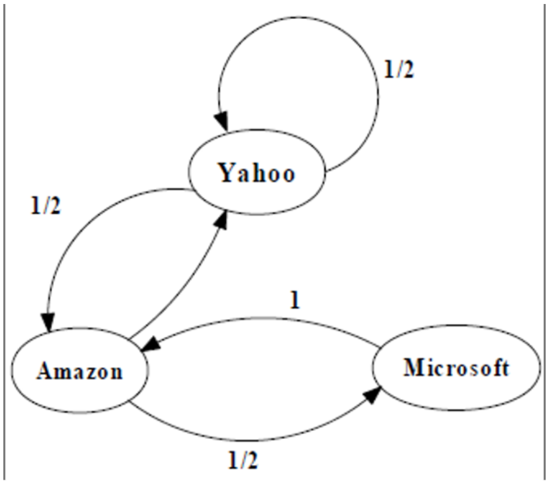


Figure 2: Page rank algorithm example

Consider the above webgraph, here the total number of unique nodes is 3 so we assign the probability (initial rank) to each node as ⅓. Now the initial matrix is generated as below

$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix}$$

Figure 3: Adjacency Matrix

Here the first column represents Yahoo and each row represents the different website. Now the values for the first column meaning, the M[0][0] represent the probability that a person visits Yahoo if he is on yahoo. So the ans is ½ as there are two outgoing links from the yahoo which includes one itself to yahoo and other to amazon. Now M[1][0] which is the probability that a person visits Amazon if he is on Yahoo is ½ . Now the probability that a person can get to Microsoft if he is on Yahoo is 0 as there are no outgoing links from Yahoo to Microsoft.

This way we calculate the initial matrix value for each webgraph.

Step 1: Create N x 1 matrix assigning probability to each node as 1/n (n= total number of unique nodes).

Step 2: Formulate a webgraph matrix N x N as explained above..

Step 3: Multiply N x N matrix with N x 1 matrix and get the rank.

Step 4: Calculate the convergence matrix. This is Output obtained from step 3 minus the initial rank matrix (which for the first time is 1/n).

Step 5: If the convergence value is less than 0.01 then stop the iteration else repeat from step 3 (Multiply the web matrix with new rank).

3.2 Teleporting:

There are two problems with the page rank one is the spider trap and other is the dead end.

Spider trap: When we encounter a webpage that refers to some other webpage, which again refers to that same webpage. Therefore, it will get into an infinite loop as there is no way out. This problem is called Spider trap. There can be many web pages on the web which may cause this situation so this problem needs some solution which is solved by teleporting.
Dead end: A web page that has no outgoing link causes a dead end. In the iterative computation, the occurrence of dead ends will cause the Page Rank of some or all pages to be set to 0, including pages that are not dead ends.

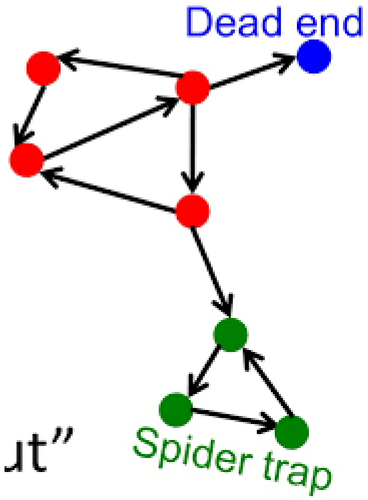


Figure 4: Spider trap and dead end example

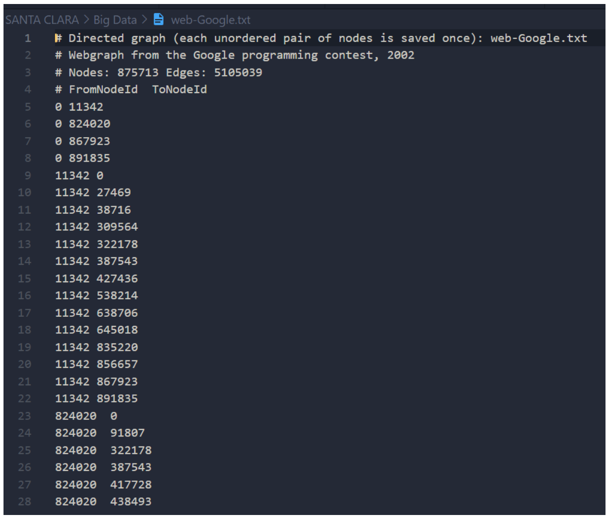
With teleporting we make an assumption that the probability that a person follows the actual path is β and the probability that it will jump to any node out of n node is (1-β). Here β is known as the damping factor which is taken as 0.85.

Now the rank is calculated as $r = \beta M.\ r + (1 - \beta)[\frac{1}{N}]_{NxN}$

4

4. IMPLEMENTATION

For the pagerank algorithm in our project we have used Google web graph dataset which contains 875713 nodes and 5105039 edges. This dataset is a text file where each line has two values from_node_id and to_node_id. This represents the outgoing links.



```
# Directed graph (each unordered pair of nodes is saved once): web-Google.txt
# Webgraph from the Google programming contest, 2002
# Nodes: 875713 Edges: 5105039
# FromNodeId  ToNodeId
0 11342
0 824020
0 867923
0 891835
11342 0
11342 27469
11342 38716
11342 309564
11342 322178
11342 387543
11342 427436
11342 538214
11342 638706
11342 645018
11342 835220
11342 856657
11342 867923
11342 891835
824020  0
824020  91807
824020  322178
824020  387543
824020  417728
824020  438493
```

Figure 5: Google web-graph dataset snippet

4.1 Data Transformation:

Firstly we transform this dataset into the spark RDD dataset with a key as the node id of the node from where the outgoing link starts and the value is node id of the node from where the outgoing link ends.

- links = sc.textFile(‘web-Google.txt’)
- links=links.map(lambdax:(x.split(‘\t’)[0], x.split(‘\t’)[1])).distinct().groupByKey().partitionBy(partition)
- N= links.count()
- ranks = links.map(lambda : url_neightbors : (url_neighbors[0],1.0/N))

- ranks=ranks.partitionBy(partitions)

Here the output of the code will be in the format of (node_id*, arr(node_ids))
arr(node_ids) is the all node id to which the outgoing link from node_id* points.

For the power iteration we assign the probability to each unique node as 1/N so in the next line we calculate the rank for each node.

4.2 Power Iteration:

To calculate the rank we need to multiply the matrix M with initial rank $r^0 = [1/n, 1/n, 1/n \ldots 1/n]^T$

So $r^{(t+1)} = M. r^{(t)}$
And then we calculate the Convergence as $|r^{(t+1)} - r^{(t)}| < \varepsilon$

```
for j in range(iteration) :
        ranks =ranks.join(links).flatMap(lambda x: [(i, float(x[1][1])/len(x[1][0]))    for    i    in
        x[1][0]]).reduceByKey(lambda x,y: x+y).partitionByKey(partition)

        ranks=ranks.sortByKey()
        r1= ranks.values().zipwithIndex().collect()
        Val = convergence(r0,r1)
```

The above for loop runs till we reach the convergence value. So firstly the join operation joins two array ranks and links. The output for the join will be (nodeid ,[[to_nodeids], 1/N)]). Then flat map runs for all the "nodeid" key and then for each nodeid it iterates the for loop for the array present in the value of "nodeid" and generates the new array with the value of the array as key and the probability as value.

The reduceByKey sums up the value for the same key and calculates the probability for that node. We then calculate the convergence value and check if it is less than zero or not. We will perform the iteration based on the convergence value. Here we partition the data into the partition number provided explicitly by the user.

4.3 Naive Implementation:

The initial data preparation part remains the same as the power iteration. After we find the total number of unique nodes in the web graph we assign probability to each unique node as 1. The Naive implementation considers the damping factor to avoid the spider trap and dead end problems that occur in power iteration. So the new equation becomes 0.85 * probability to follow original path + 0.15 *probability that it goes to random nodes.

The reason behind this is whenever the person enters the spider trap it remains inside it for infinite time as there are no outgoing links and that node takes whole importance. So when a person reaches such a website he will get bored of visiting the same website continuously. Hence to avoid this google introduces a damping factor which means that there are 0.85% chances that it will remain to the same website and 0.15% chances that it will move to any random node.

- ranks = links.map(lambda : url_neightbors : (url_neighbors[0],1.0))
- ranks=ranks.partitionBy(partitions)

Iterations till convergence are as follows:

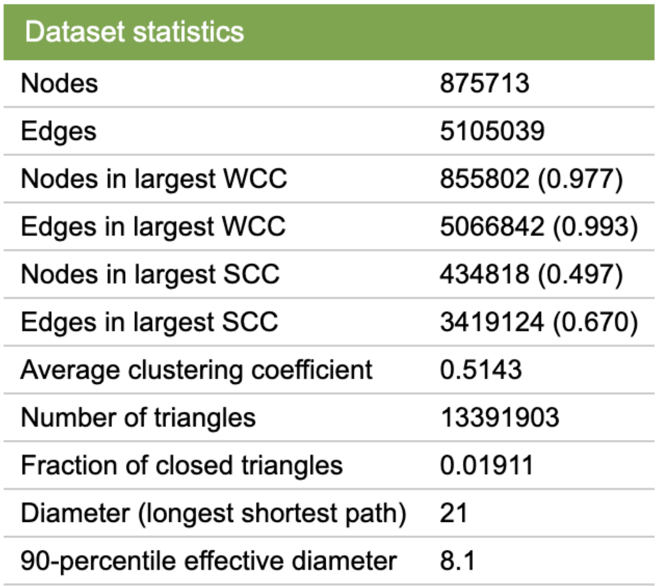
- contribution = links.join(ranks).flatMap(lambda x: [(i, float(x[1][1])/len(x[1][0])) for i in x[1][0]]).sortBy(lambda x: x[1], ascending= false)
- contribution = contribution.reduceByKey(add).mapValues(lambda score: score*0.85 + 0.15).partitionBy(partitions)
- r1=ranks.values().zipWithIndex().collect()
- ans= convergence(r0,r1,j,partitions,sc)

Convergence function is defined as

- rdd= sc.parallelize(r0).map(lambda x: (x[1],x[0]).partitionBy(partitions)
- rdd1= sc.parallelize(r1).map(lambda x: (x[1],x[0]).partitionBy(partitions)
- r=rdd.join(rdd1)
- ans= r.mapValues(lambda y: abs(y[1]-y[0])).values().reduce(lambda x,y: x+y)
- return ans

5

5. EXPERIMENT RESULTS AND ANALYSIS

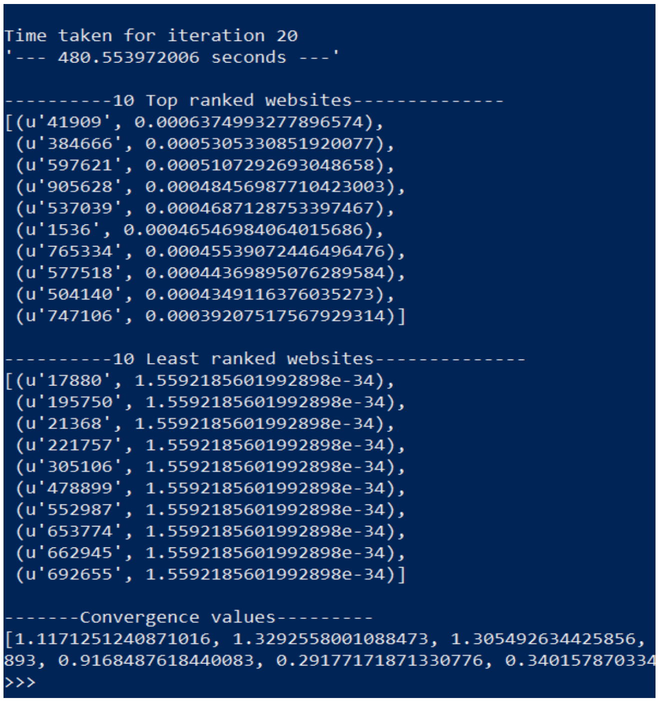


Dataset statistics	
Nodes	875713
Edges	5105039
Nodes in largest WCC	855802 (0.977)
Edges in largest WCC	5066842 (0.993)
Nodes in largest SCC	434818 (0.497)
Edges in largest SCC	3419124 (0.670)
Average clustering coefficient	0.5143
Number of triangles	13391903
Fraction of closed triangles	0.01911
Diameter (longest shortest path)	21
90-percentile effective diameter	8.1

Figure 6: Characteristics of the google web-graph dataset

5.1 PageRank Power Iteration with Partitions:

As given below, Figure 7 shows the results exhibiting Top 10 most ranked websites, 10 least ranked websites as well as convergence values with 20 partitions over 20 iterations using power iteration method. The time taken for execution is 480.553972006 seconds.



```
Time taken for iteration 20
'--- 480.553972006 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 0.0006374993277896574),
 (u'384666', 0.0005305330851920077),
 (u'597621', 0.0005107292693048658),
 (u'905628', 0.00048456987710423003),
 (u'537039', 0.0004687128753397467),
 (u'1536', 0.00046546984064015686),
 (u'765334', 0.00045539072446496476),
 (u'577518', 0.00044369895076289584),
 (u'504140', 0.0004349116376035273),
 (u'747106', 0.00039207517567929314)]

----------10 Least ranked websites--------------
[(u'17880', 1.5592185601992898e-34),
 (u'195750', 1.5592185601992898e-34),
 (u'21368', 1.5592185601992898e-34),
 (u'221757', 1.5592185601992898e-34),
 (u'305106', 1.5592185601992898e-34),
 (u'478899', 1.5592185601992898e-34),
 (u'552987', 1.5592185601992898e-34),
 (u'653774', 1.5592185601992898e-34),
 (u'662945', 1.5592185601992898e-34),
 (u'692655', 1.5592185601992898e-34)]

-------Convergence values---------
[1.1171251240871016, 1.3292558001088473, 1.305492634425856,
893, 0.9168487618440083, 0.29177171871330776, 0.340157870334
>>>
```

Figure 7: Result displaying time taken, top 10 ranked websites, 10 least ranked websites and convergence values with 20 partitions over 20 iterations by power iteration

Results of other observations

For better understanding of the power iteration method with partitions and to analyze the time taken for the computation, we have performed this algorithm with 10 and 20 partitions respectively with 20, 25 and 40 iterations. The observations are given in the following table.

Partitions	Iterations	Time taken in sec
10	20	514
	25	638.32
	40	995.166
20	20	480.55
	25	607.903
	40	954.177

Table 1: Results displaying time taken by power iteration method for execution with 10 and 20 partitions over 20,25 and 40 iterations

The bar graph represented below, compares time taken for the execution of the power iteration method with 10 partitions and 20 partitions over 20, 25 and 40 iterations respectively.

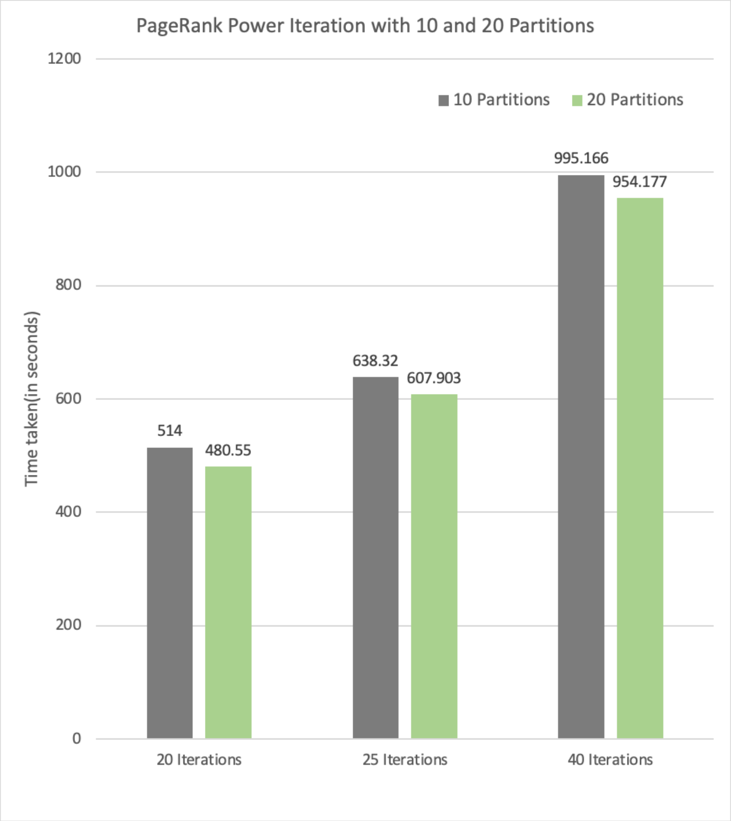
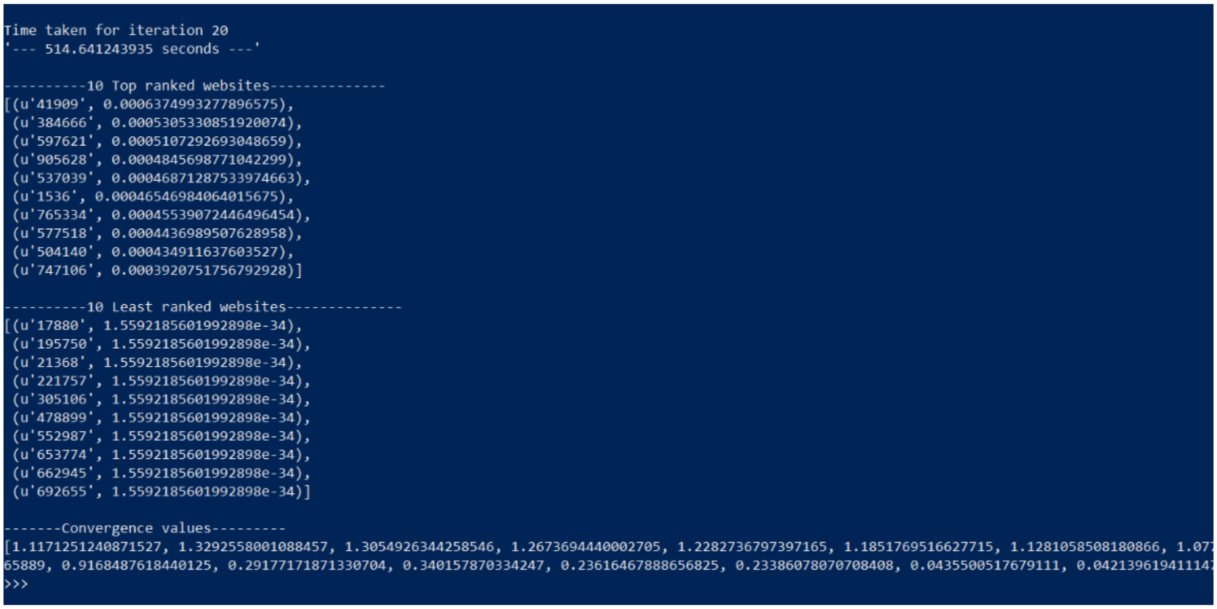


Figure 8 : Graph comparing the the time taken by the power iteration method with 10 and 20 partitions over 20, 25 and 40 iterations

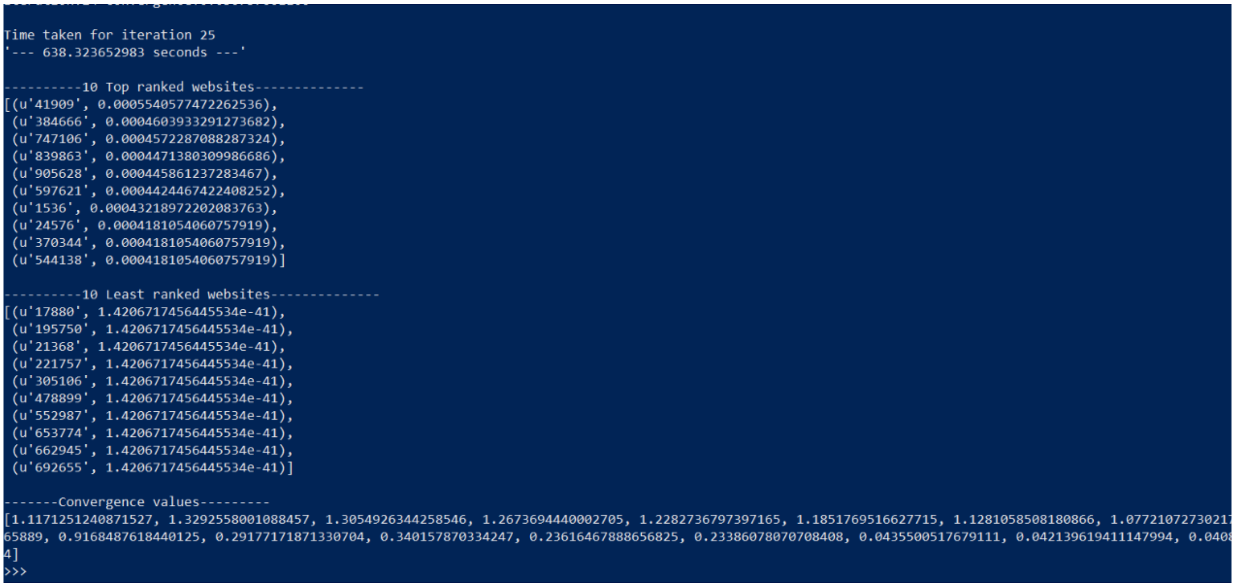
10 Partition 20 Iteration



```
Time taken for iteration 20
'--- 514.641243935 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 0.0006374993277896575),
 (u'384666', 0.0005305330851920074),
 (u'597621', 0.0005107292693048659),
 (u'905628', 0.0004845698771042299),
 (u'537039', 0.00046871287533974663),
 (u'1536', 0.00046546984064015675),
 (u'765334', 0.00045539072446496454),
 (u'577518', 0.0004436989507628958),
 (u'504140', 0.000434911637603527),
 (u'747106', 0.0003920751756792928)]

----------10 Least ranked websites--------------
[(u'17880', 1.5592185601992898e-34),
 (u'195750', 1.5592185601992898e-34),
 (u'21368', 1.5592185601992898e-34),
 (u'221757', 1.5592185601992898e-34),
 (u'305106', 1.5592185601992898e-34),
 (u'478899', 1.5592185601992898e-34),
 (u'552987', 1.5592185601992898e-34),
 (u'653774', 1.5592185601992898e-34),
 (u'662945', 1.5592185601992898e-34),
 (u'692655', 1.5592185601992898e-34)]

-------Convergence values---------
[1.1171251240871527, 1.3292558001088457, 1.3054926344258546, 1.2673694440002705, 1.2282736797397165, 1.1851769516627715, 1.1281058508180866, 1.07
65889, 0.9168487618440125, 0.29177171871330704, 0.340157870334247, 0.23616467888656825, 0.23386078070708408, 0.0435500517679111, 0.04213961941114
>>>
```

Figure 9: Results for execution of power iteration method with 10 partitions over 20 iterations

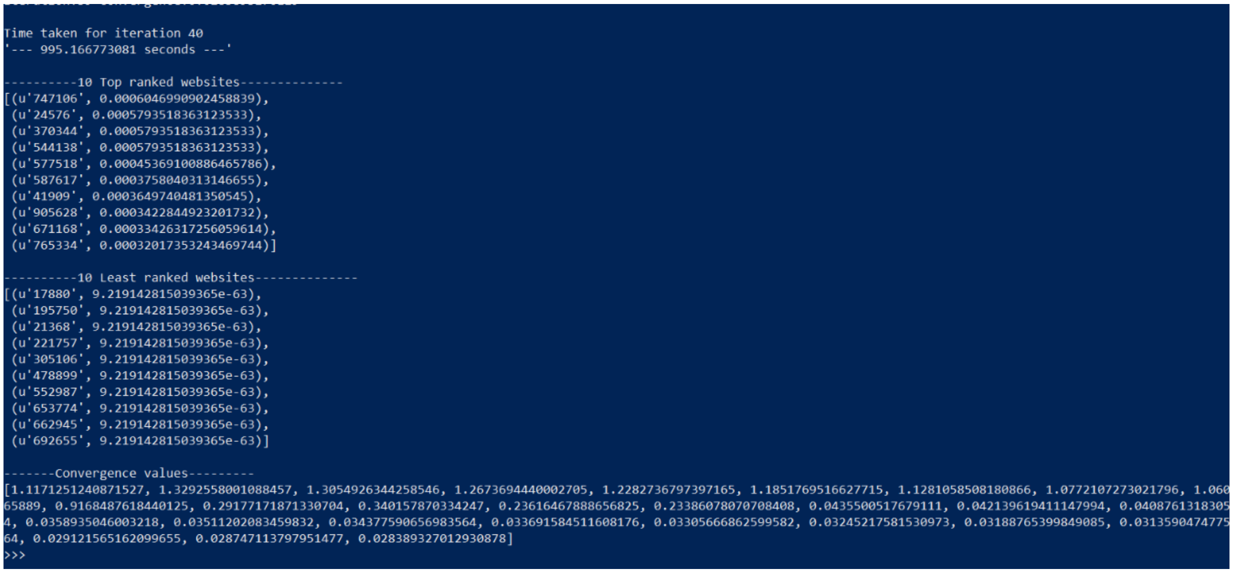
10 Partition 25 Iteration



```
Time taken for iteration 25
'--- 638.323652983 seconds ---'

----------10 Top ranked websites-------------
[(u'41909', 0.0005540577472262536),
 (u'384666', 0.0004603933291273682),
 (u'747106', 0.0004572287088287324),
 (u'839863', 0.0004471380309986686),
 (u'905628', 0.000445861237283467),
 (u'597621', 0.0004424467422408252),
 (u'1536', 0.00043218972202083763),
 (u'24576', 0.0004181054060757919),
 (u'370344', 0.0004181054060757919),
 (u'544138', 0.0004181054060757919)]

----------10 Least ranked websites-------------
[(u'17880', 1.4206717456445534e-41),
 (u'195750', 1.4206717456445534e-41),
 (u'21368', 1.4206717456445534e-41),
 (u'221757', 1.4206717456445534e-41),
 (u'305106', 1.4206717456445534e-41),
 (u'478899', 1.4206717456445534e-41),
 (u'552987', 1.4206717456445534e-41),
 (u'653774', 1.4206717456445534e-41),
 (u'662945', 1.4206717456445534e-41),
 (u'692655', 1.4206717456445534e-41)]

-------Convergence values---------
[1.1171251240871527, 1.3292558001088457, 1.3054926344258546, 1.2673694440002705, 1.2282736797397165, 1.1851769516627715, 1.1281058508180866, 1.07721072730217
65889, 0.9168487618440125, 0.29177171871330704, 0.340157870334247, 0.23616467888656825, 0.23386078070708408, 0.0435500517679111, 0.042139619411147994, 0.0408
4]
>>>
```

Figure 10: Results for execution of power iteration method with 10 partitions over 25 iterations

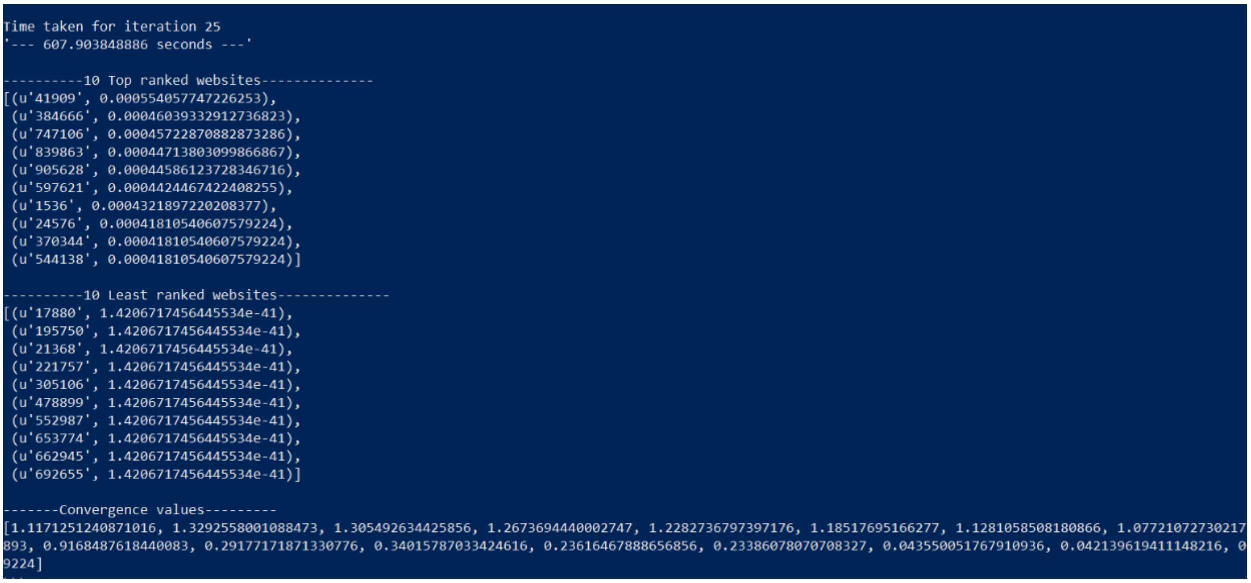
10 Partition 40 Iteration



```
Time taken for iteration 40
'--- 995.166773081 seconds ---'

----------10 Top ranked websites-------------
[(u'747106', 0.0006046990902458839),
 (u'24576', 0.0005793518363123533),
 (u'370344', 0.0005793518363123533),
 (u'544138', 0.0005793518363123533),
 (u'577518', 0.00045369100886465786),
 (u'587617', 0.0003758040313146655),
 (u'41909', 0.0003649740481350545),
 (u'905628', 0.0003422844923201732),
 (u'671168', 0.00033426317256059614),
 (u'765334', 0.00032017353243469744)]

----------10 Least ranked websites-------------
[(u'17880', 9.219142815039365e-63),
 (u'195750', 9.219142815039365e-63),
 (u'21368', 9.219142815039365e-63),
 (u'221757', 9.219142815039365e-63),
 (u'305106', 9.219142815039365e-63),
 (u'478899', 9.219142815039365e-63),
 (u'552987', 9.219142815039365e-63),
 (u'653774', 9.219142815039365e-63),
 (u'662945', 9.219142815039365e-63),
 (u'692655', 9.219142815039365e-63)]

-------Convergence values---------
[1.1171251240871527, 1.3292558001088457, 1.3054926344258546, 1.2673694440002705, 1.2282736797397165, 1.1851769516627715, 1.1281058508180866, 1.0772107273021796, 1.060
65889, 0.9168487618440125, 0.29177171871330704, 0.340157870334247, 0.23616467888656825, 0.23386078070708408, 0.0435500517679111, 0.042139619411147994, 0.0408761318305
4, 0.0358935046003218, 0.03511202083459832, 0.034377590656983564, 0.033691584511608176, 0.03305666862599582, 0.03245217581530973, 0.03188765399849085, 0.0313590474775
64, 0.029121565162099655, 0.028747113797951477, 0.028389327012930878]
>>>
```

Figure 11: Results for execution of power iteration method with 10 partitions over 40 iterations

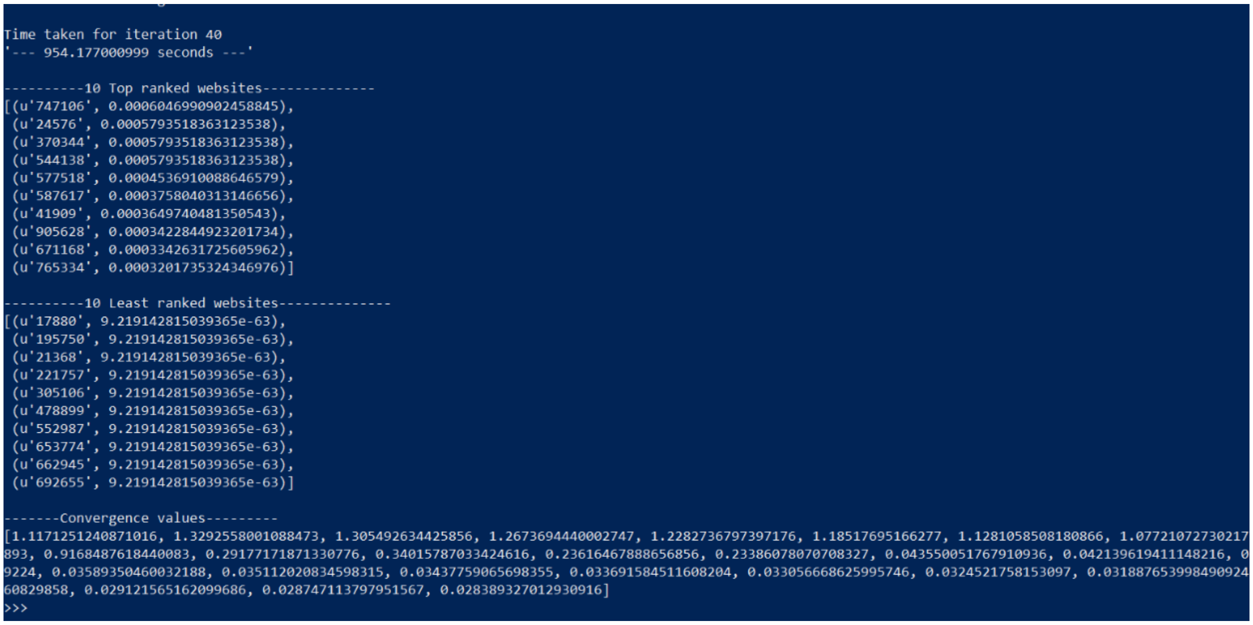
20 Partition 25 Iteration



```
Time taken for iteration 25
'--- 607.903848886 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 0.000554057747226253),
 (u'384666', 0.00046039332912736823),
 (u'747106', 0.00045722870882873286),
 (u'839863', 0.00044713803099866867),
 (u'905628', 0.00044586123728346716),
 (u'597621', 0.0004424467422408255),
 (u'1536', 0.0004321897220208377),
 (u'24576', 0.00041810540607579224),
 (u'370344', 0.00041810540607579224),
 (u'544138', 0.00041810540607579224)]

----------10 Least ranked websites--------------
[(u'17880', 1.4206717456445534e-41),
 (u'195750', 1.4206717456445534e-41),
 (u'21368', 1.4206717456445534e-41),
 (u'221757', 1.4206717456445534e-41),
 (u'305106', 1.4206717456445534e-41),
 (u'478899', 1.4206717456445534e-41),
 (u'552987', 1.4206717456445534e-41),
 (u'653774', 1.4206717456445534e-41),
 (u'662945', 1.4206717456445534e-41),
 (u'692655', 1.4206717456445534e-41)]

-------Convergence values---------
[1.1171251240871016, 1.3292558001088473, 1.305492634425856, 1.2673694440002747, 1.2282736797397176, 1.18517695166277, 1.1281058508180866, 1.07721072730217
893, 0.9168487618440083, 0.29177171871330776, 0.34015787033424616, 0.23616467888656856, 0.23386078070708327, 0.043550051767910936, 0.042139619411148216, 0
9224]
```

Figure 12: Results for execution of power iteration method with 20 partitions over 25 iterations

20 Partition 40 Iteration

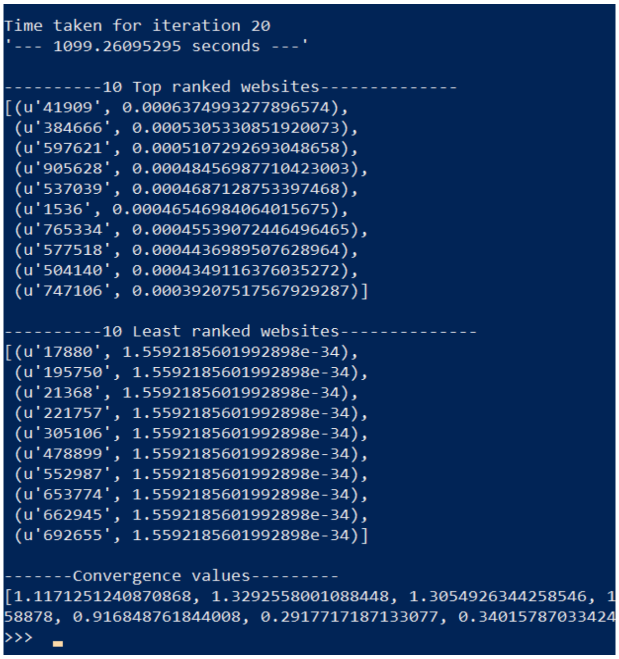


```
Time taken for iteration 40
'--- 954.177000999 seconds ---'

----------10 Top ranked websites--------------
[(u'747106', 0.0006046990902458845),
 (u'24576', 0.0005793518363123538),
 (u'370344', 0.0005793518363123538),
 (u'544138', 0.0005793518363123538),
 (u'577518', 0.0004536910088646579),
 (u'587617', 0.0003758040313146656),
 (u'41909', 0.0003649740481350543),
 (u'905628', 0.0003422844923201734),
 (u'671168', 0.0003342631725605962),
 (u'765334', 0.0003201735324346976)]

----------10 Least ranked websites--------------
[(u'17880', 9.219142815039365e-63),
 (u'195750', 9.219142815039365e-63),
 (u'21368', 9.219142815039365e-63),
 (u'221757', 9.219142815039365e-63),
 (u'305106', 9.219142815039365e-63),
 (u'478899', 9.219142815039365e-63),
 (u'552987', 9.219142815039365e-63),
 (u'653774', 9.219142815039365e-63),
 (u'662945', 9.219142815039365e-63),
 (u'692655', 9.219142815039365e-63)]

-------Convergence values---------
[1.1171251240871016, 1.3292558001088473, 1.305492634425856, 1.2673694440002747, 1.2282736797397176, 1.18517695166277, 1.1281058508180866, 1.07721072730217
893, 0.9168487618440083, 0.29177171871330776, 0.34015787033424616, 0.23616467888656856, 0.23386078070708327, 0.043550051767910936, 0.042139619411148216, 0
9224, 0.03589350460032188, 0.035112020834598315, 0.03437759065698355, 0.033691584511608204, 0.033056668625995746, 0.0324521758153097, 0.031887653998490924
60829858, 0.029121565162099686, 0.028747113797951567, 0.028389327012930916]
>>>
```

Figure 13: Results for execution of power iteration method with 20 partitions over 40 iterations

5.2 PageRank Power Iteration with No Partitions:

As given below, Figure 7 shows the results exhibiting Top 10 most ranked websites, 10 least ranked websites as well as convergence values without partitions over 20 iterations using power iteration method. The time taken for execution is 1099.26095295 seconds.



```
Time taken for iteration 20
'--- 1099.26095295 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 0.0006374993277896574),
 (u'384666', 0.0005305330851920073),
 (u'597621', 0.0005107292693048658),
 (u'905628', 0.00048456987710423003),
 (u'537039', 0.0004687128753397468),
 (u'1536', 0.00046546984064015675),
 (u'765334', 0.00045539072446496465),
 (u'577518', 0.0004436989507628964),
 (u'504140', 0.0004349116376035272),
 (u'747106', 0.00039207517567929287)]

----------10 Least ranked websites--------------
[(u'17880', 1.5592185601992898e-34),
 (u'195750', 1.5592185601992898e-34),
 (u'21368', 1.5592185601992898e-34),
 (u'221757', 1.5592185601992898e-34),
 (u'305106', 1.5592185601992898e-34),
 (u'478899', 1.5592185601992898e-34),
 (u'552987', 1.5592185601992898e-34),
 (u'653774', 1.5592185601992898e-34),
 (u'662945', 1.5592185601992898e-34),
 (u'692655', 1.5592185601992898e-34)]

-------Convergence values---------
[1.1171251240870868, 1.3292558001088448, 1.3054926344258546, 1
58878, 0.916848761844008, 0.2917717187133077, 0.34015787033424
>>>
```

Figure 14: Result displaying time taken, top 10 ranked websites, 10 least ranked websites and convergence values without partitions over 20 iterations by power iteration

Result at different iterations

To compare and analyze the results of the power iteration method with partitions and without partitions, we have executed the power iteration method without partitions over 20, 25 and 40 iterations respectively. The table given below shows the time taken without partitions over different numbers of iterations.

Iteration	Time taken in sec
20	1099.26
25	1364.24
40	2143.364

Table 2: Time taken by power iteration method without partitions over 20, 25 and 40 iterations.

The bar graph demonstrated below, compares time taken for the execution of power iteration method without partitions over 20, 25 and 40 iterations respectively.

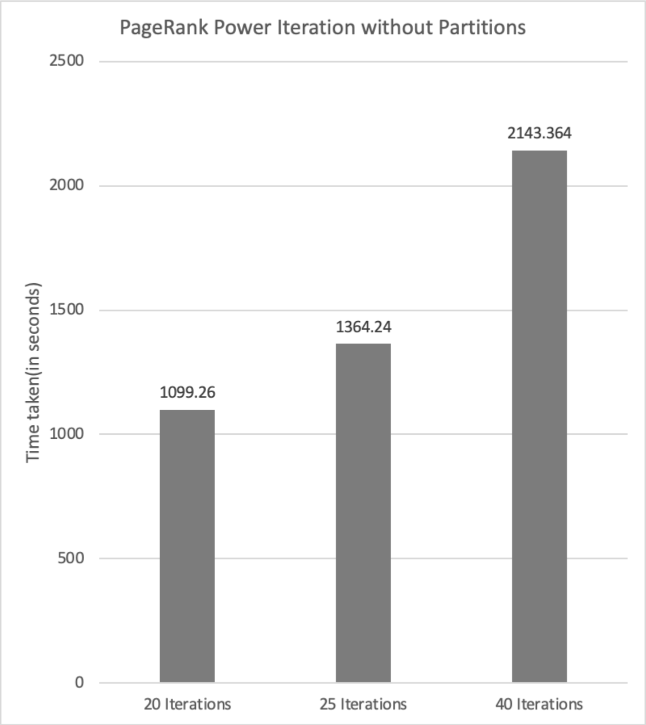
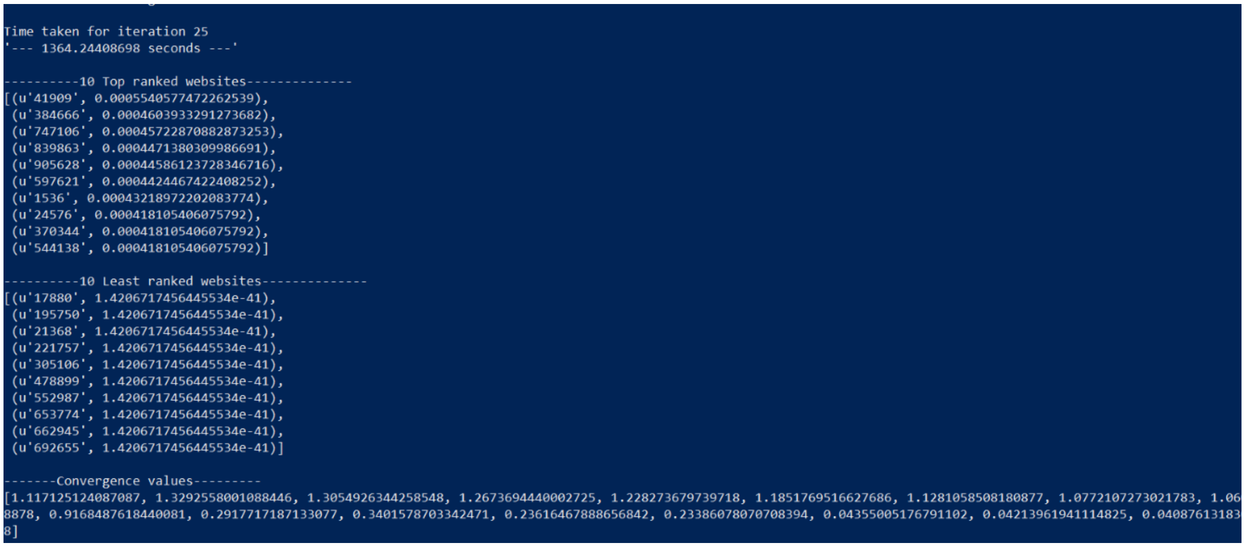


Figure 15: Graph comparing the time taken by the power iteration method without partitions over 20, 25 and 40 iterations

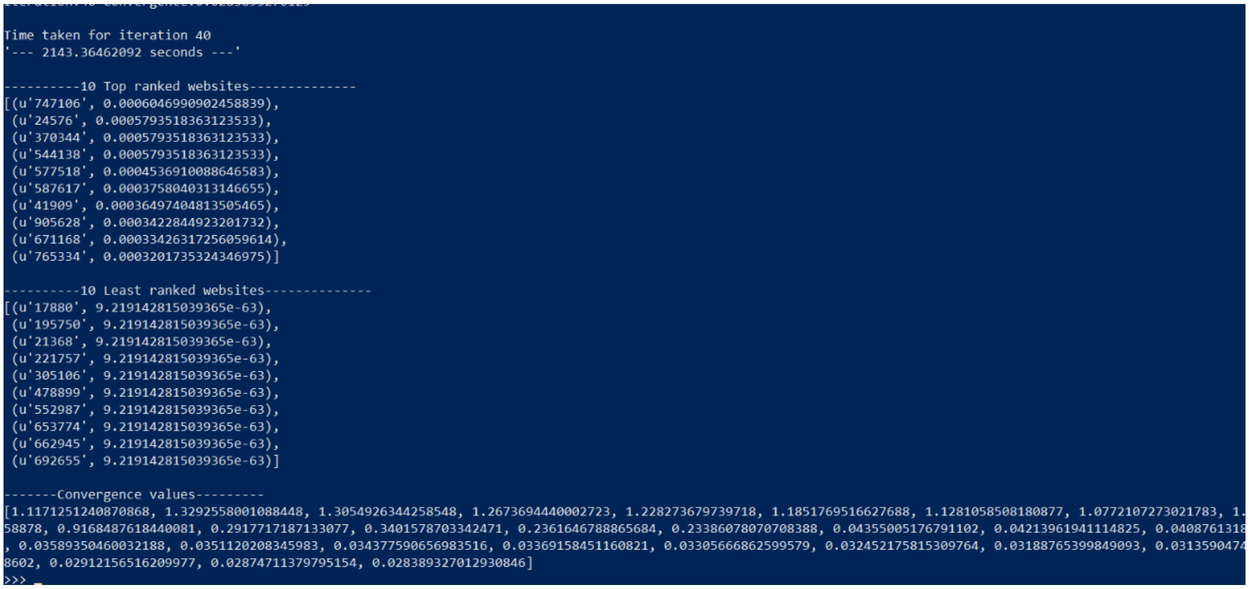
25 Iterations



```
Time taken for iteration 25
'--- 1364.24408698 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 0.0005540577472262539),
 (u'384666', 0.0004603933291273682),
 (u'747106', 0.00045722870882873253),
 (u'839863', 0.0004471380309986691),
 (u'905628', 0.00044586123728346716),
 (u'597621', 0.0004424467422408252),
 (u'1536', 0.00043218972202083774),
 (u'24576', 0.000418105406075792),
 (u'370344', 0.000418105406075792),
 (u'544138', 0.000418105406075792)]

----------10 Least ranked websites--------------
[(u'17880', 1.4206717456445534e-41),
 (u'195750', 1.4206717456445534e-41),
 (u'21368', 1.4206717456445534e-41),
 (u'221757', 1.4206717456445534e-41),
 (u'305106', 1.4206717456445534e-41),
 (u'478899', 1.4206717456445534e-41),
 (u'552987', 1.4206717456445534e-41),
 (u'653774', 1.4206717456445534e-41),
 (u'662945', 1.4206717456445534e-41),
 (u'692655', 1.4206717456445534e-41)]

-------Convergence values---------
[1.117125124087087, 1.3292558001088446, 1.3054926344258548, 1.2673694440002725, 1.228273679739718, 1.1851769516627686, 1.1281058508180877, 1.0772107273021783, 1.06
8878, 0.9168487618440081, 0.2917717187133077, 0.3401578703342471, 0.23616467888656842, 0.23386078070708394, 0.04355005176791102, 0.04213961941114825, 0.04087613183
8]
```

Figure 16:Results for execution of power iteration method without partitions over 25 iterations

40 Iteration



```
Time taken for iteration 40
'--- 2143.36462092 seconds ---'

----------10 Top ranked websites--------------
[(u'747106', 0.0006046990902458839),
 (u'24576', 0.0005793518363123533),
 (u'370344', 0.0005793518363123533),
 (u'544138', 0.0005793518363123533),
 (u'577518', 0.0004536910088646583),
 (u'587617', 0.0003758040313146655),
 (u'41909', 0.00036497404813505465),
 (u'905628', 0.0003422844923201732),
 (u'671168', 0.00033426317256059614),
 (u'765334', 0.0003201735324346975)]

----------10 Least ranked websites--------------
[(u'17880', 9.219142815039365e-63),
 (u'195750', 9.219142815039365e-63),
 (u'21368', 9.219142815039365e-63),
 (u'221757', 9.219142815039365e-63),
 (u'305106', 9.219142815039365e-63),
 (u'478899', 9.219142815039365e-63),
 (u'552987', 9.219142815039365e-63),
 (u'653774', 9.219142815039365e-63),
 (u'662945', 9.219142815039365e-63),
 (u'692655', 9.219142815039365e-63)]

-------Convergence values---------
[1.1171251240870868, 1.3292558001088448, 1.3054926344258548, 1.2673694440002723, 1.228273679739718, 1.1851769516627688, 1.1281058508180877, 1.0772107273021783, 1.
58878, 0.9168487618440081, 0.2917717187133077, 0.3401578703342471, 0.2361646788865684, 0.23386078070708388, 0.04355005176791102, 0.04213961941114825, 0.0408761318
, 0.03589350460032188, 0.0351120208345983, 0.034377590656983516, 0.03369158451160821, 0.03305666862599579, 0.032452175815309764, 0.03188765399849093, 0.0313590474
8602, 0.02912156516209977, 0.02874711379795154, 0.028389327012930846]
>>> _
```

Figure 17:Results for execution of power iteration method without partitions over 40 iterations

5.3 Graphical Analysis of Power Iteration With And Without Partitions:

A bar graph as given below by Figure 18 shows the analysis of time taken by the algorithm to get page ranks with 20 partitions and without partition over 20 and 25 iterations. As per the graph we can conclude that the time taken to compute the page ranks with partitions is lesser than the time taken for the computation of page ranks without partitions. So it is clear that the usage of partitions will increase the parallelization and decrease shuffling operations which results in speeding up the jobs.

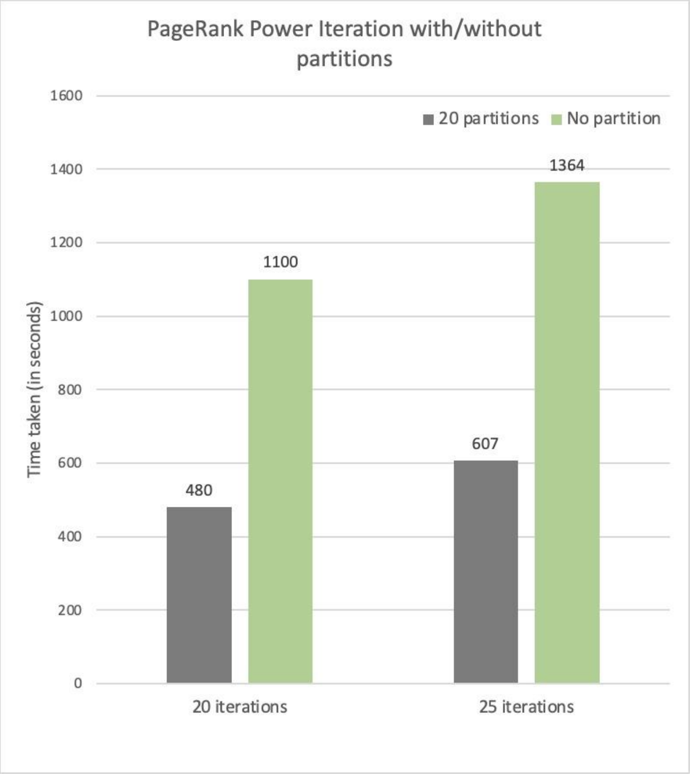


Figure 18: Time taken by power iteration method with 20 partition and without partition over 20 and 25 iterations

The line graph as given in Figure 19 shows the convergence values for 90 iterations over 40 partitions. The convergence value that we have considered is with the criteria of 0.0001. At the 90th iteration, the observed convergence value was around 0.0210.

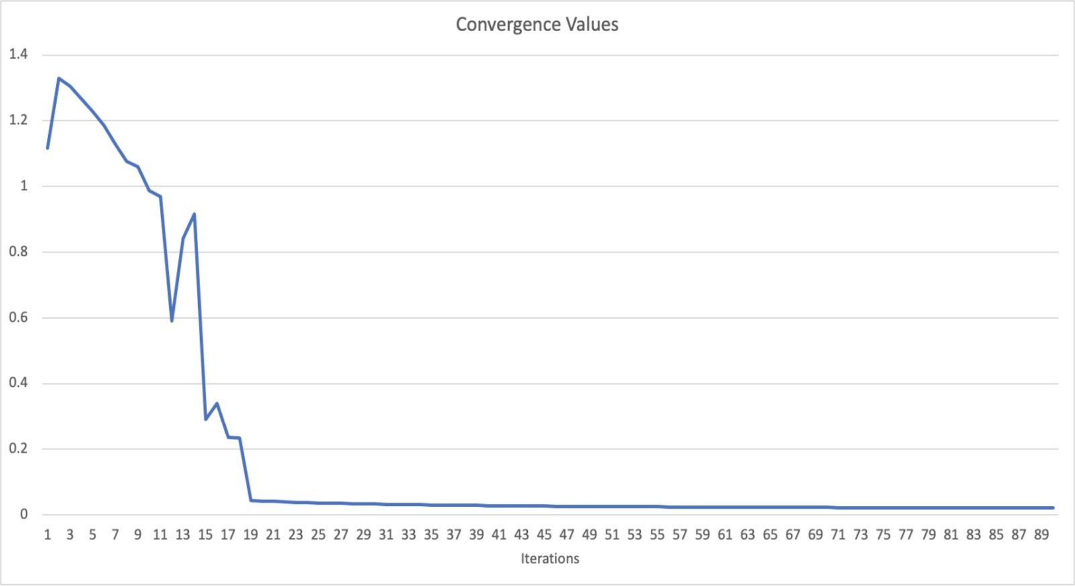
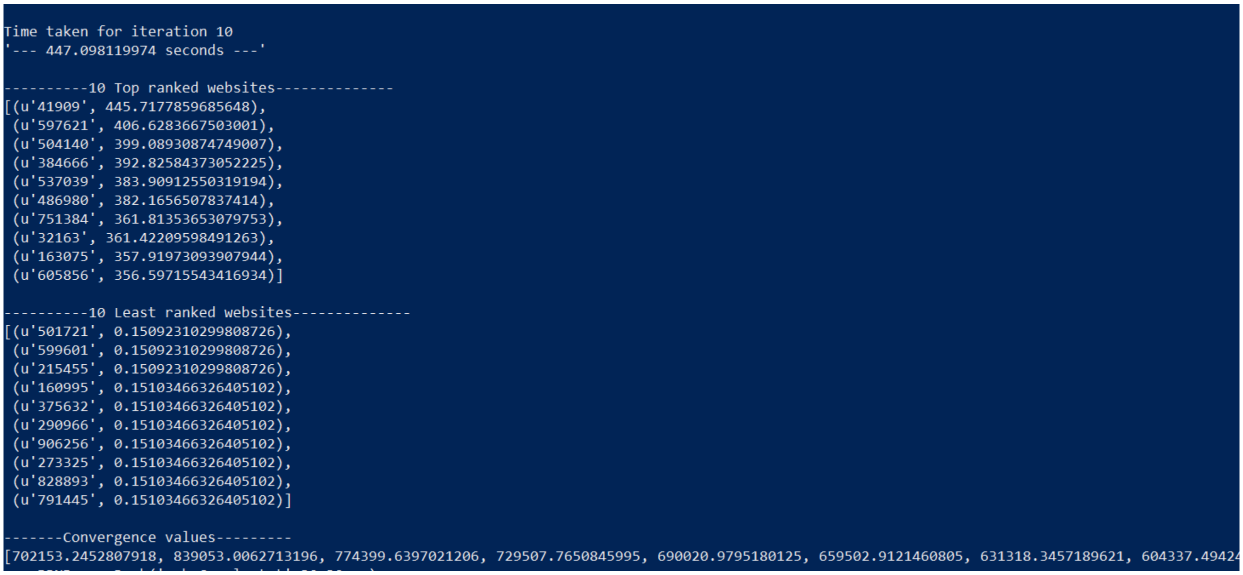


Figure 19: Convergence values for 90 iterations over 40 partitions

5.4 PageRank Naive With Partitions:

Below Figure shows the statistics for Naive implementation running for 10 Iterations with 10 Partitions. Comparing the result of Naive with Power iteration the website that ranks first is 41909 while the top 10 rank websites in both are the same, just the order is different.



```
Time taken for iteration 10
'--- 447.098119974 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 445.7177859685648),
 (u'597621', 406.6283667503001),
 (u'504140', 399.08930874749007),
 (u'384666', 392.82584373052225),
 (u'537039', 383.90912550319194),
 (u'486980', 382.1656507837414),
 (u'751384', 361.81353653079753),
 (u'32163', 361.42209598491263),
 (u'163075', 357.91973093907944),
 (u'605856', 356.59715543416934)]

----------10 Least ranked websites--------------
[(u'501721', 0.15092310299808726),
 (u'599601', 0.15092310299808726),
 (u'215455', 0.15092310299808726),
 (u'160995', 0.1510346632640510 2),
 (u'375632', 0.15103466326405102),
 (u'290966', 0.15103466326405102),
 (u'906256', 0.15103466326405102),
 (u'273325', 0.15103466326405102),
 (u'828893', 0.15103466326405102),
 (u'791445', 0.15103466326405102)]

-------Convergence values---------
[702153.2452807918, 839053.0062713196, 774399.6397021206, 729507.7650845995, 690020.9795180125, 659502.9121460805, 631318.3457189621, 604337.49424
```

Figure 20: Naive implementation for 10 iteration and 10 partition

Partition	Iteration	Time taken in sec
10	10	447.098
	20	862.989
20	10	413.9737
	20	799.2325

Table 3: Time taken by naive implementation method with 10 and 20 partitions over 10 and 20 iterations.

A bar graph as given below by Figure 21 shows the analysis of time taken by the naive algorithm to get page ranks with 10 and 20 partitions over 10 and 20 iterations. As per the graph we can conclude that the time taken to compute the page ranks decreases as the number of partitions increases.

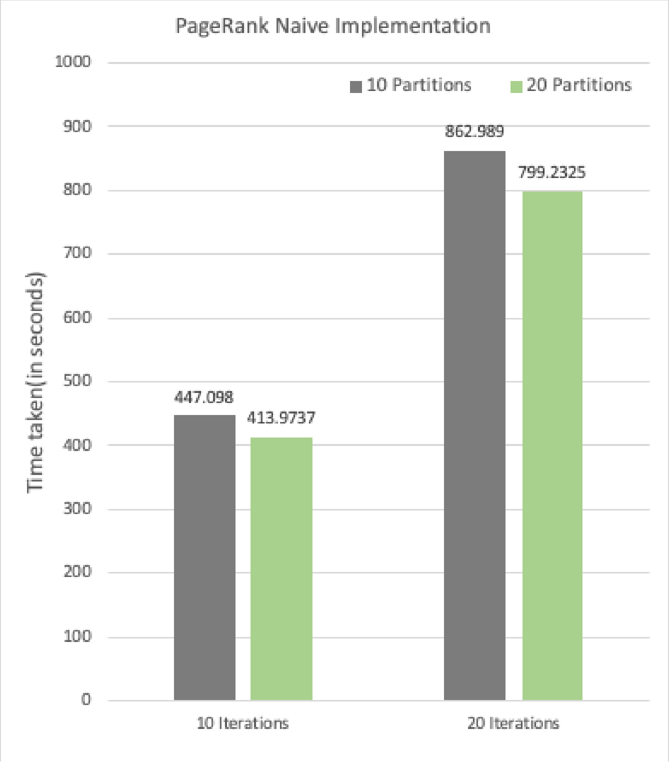


Figure 21: Time taken by naive implementation method with 10 partition and 20 partition over 10 and 20 iterations

10 Partition 20 Iteration

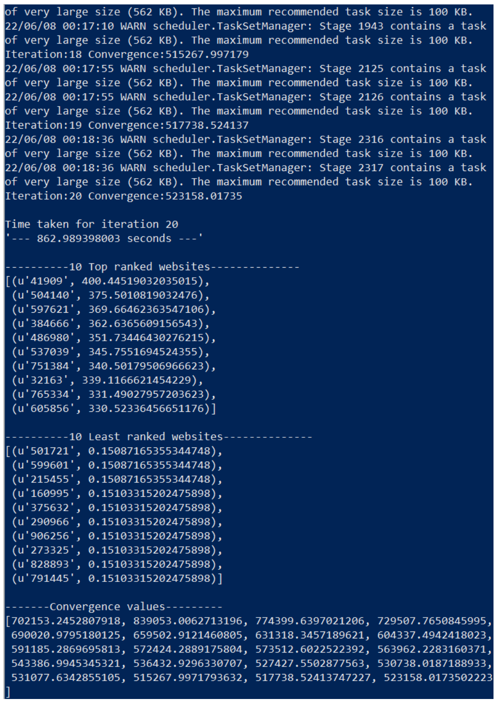
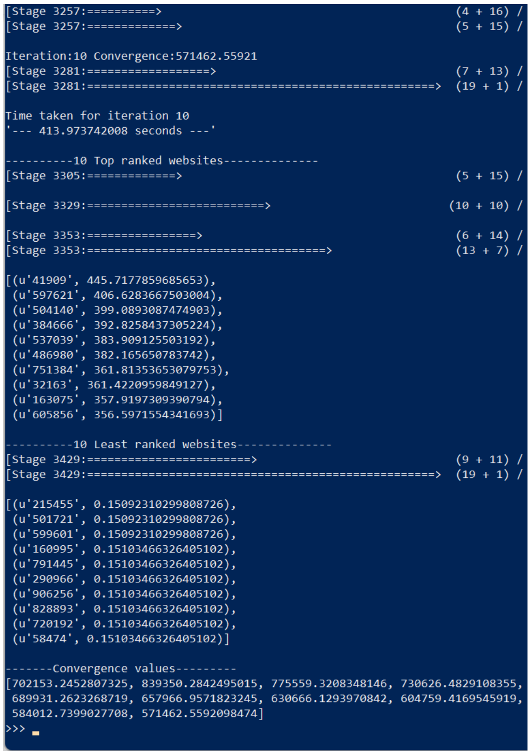

```
of very large size (562 KB). The maximum recommended task size is 100 KB.
22/06/08 00:17:10 WARN scheduler.TaskSetManager: Stage 1943 contains a task
of very large size (562 KB). The maximum recommended task size is 100 KB.
Iteration:18 Convergence:515267.997179
22/06/08 00:17:55 WARN scheduler.TaskSetManager: Stage 2125 contains a task
of very large size (562 KB). The maximum recommended task size is 100 KB.
22/06/08 00:17:55 WARN scheduler.TaskSetManager: Stage 2126 contains a task
of very large size (562 KB). The maximum recommended task size is 100 KB.
Iteration:19 Convergence:517738.524137
22/06/08 00:18:36 WARN scheduler.TaskSetManager: Stage 2316 contains a task
of very large size (562 KB). The maximum recommended task size is 100 KB.
22/06/08 00:18:36 WARN scheduler.TaskSetManager: Stage 2317 contains a task
of very large size (562 KB). The maximum recommended task size is 100 KB.
Iteration:20 Convergence:523158.01735

Time taken for iteration 20
'--- 862.989398003 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 400.44519032035015),
 (u'504140', 375.5010819032476),
 (u'597621', 369.66462363547106),
 (u'384666', 362.6365609156543),
 (u'486980', 351.73446430276215),
 (u'537039', 345.7551694524355),
 (u'751384', 340.50179506966623),
 (u'32163', 339.1166621454229),
 (u'765334', 331.49027957203623),
 (u'605856', 330.52336456651176)]

----------10 Least ranked websites--------------
[(u'501721', 0.15087165355344748),
 (u'599601', 0.15087165355344748),
 (u'215455', 0.15087165355344748),
 (u'160995', 0.15103315202475898),
 (u'375632', 0.15103315202475898),
 (u'290966', 0.15103315202475898),
 (u'906256', 0.15103315202475898),
 (u'273325', 0.15103315202475898),
 (u'828893', 0.15103315202475898),
 (u'791445', 0.15103315202475898)]

-------Convergence values---------
[702153.2452807918, 839053.0062713196, 774399.6397021206, 729507.7650845995,
 690020.9795180125, 659502.9121460805, 631318.3457189621, 604337.4942418023,
 591185.2869695813, 572424.2889175804, 573512.6022522392, 563962.2283160371,
 543386.9945345321, 536432.9296330707, 527427.5502877563, 530738.0187188933,
 531077.6342855105, 515267.9971793632, 517738.52413747227, 523158.0173502223
]
```

Figure 22: Results for execution of naive implementation with 10 partitions over 20 iterations

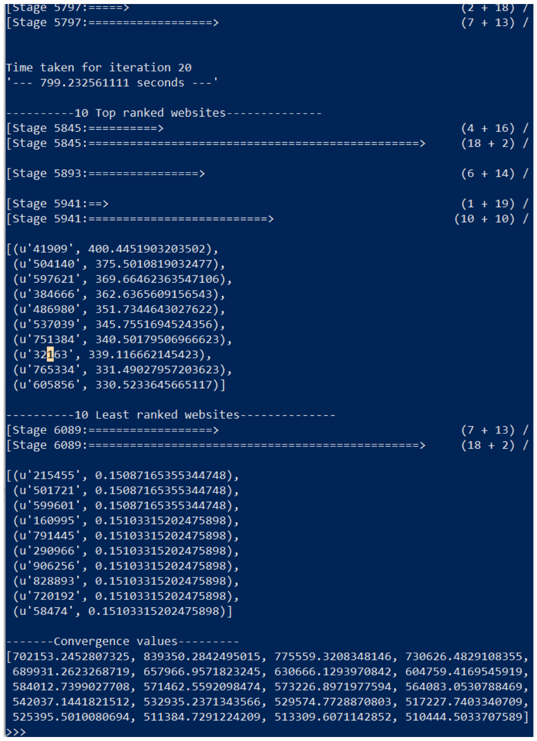
20 Partition 10 Iteration



```
[Stage 3257:==========>                                            (4 + 16) /
[Stage 3257:=============>                                         (5 + 15) /

Iteration:10 Convergence:571462.55921
[Stage 3281:==================>                                    (7 + 13) /
[Stage 3281:===================================================>   (19 + 1) /

Time taken for iteration 10
'--- 413.973742008 seconds ---'

----------10 Top ranked websites--------------
[Stage 3305:=============>                                         (5 + 15) /

[Stage 3329:==========================>                           (10 + 10) /

[Stage 3353:================>                                      (6 + 14) /
[Stage 3353:===================================>                   (13 + 7) /

[(u'41909', 445.7177859685653),
 (u'597621', 406.6283667503004),
 (u'504140', 399.0893087474903),
 (u'384666', 392.8258437305224),
 (u'537039', 383.909125503192),
 (u'486980', 382.165650783742),
 (u'751384', 361.81353653079753),
 (u'32163', 361.4220959849127),
 (u'163075', 357.9197309390794),
 (u'605856', 356.5971554341693)]

----------10 Least ranked websites--------------
[Stage 3429:========================>                              (9 + 11) /
[Stage 3429:===================================================>   (19 + 1) /

[(u'215455', 0.15092310299808726),
 (u'501721', 0.15092310299808726),
 (u'599601', 0.15092310299808726),
 (u'160995', 0.15103466326405102),
 (u'791445', 0.15103466326405102),
 (u'290966', 0.15103466326405102),
 (u'906256', 0.15103466326405102),
 (u'828893', 0.15103466326405102),
 (u'720192', 0.15103466326405102),
 (u'58474', 0.15103466326405102)]

-------Convergence values---------
[702153.2452807325, 839350.2842495015, 775559.3208348146, 730626.4829108355,
 689931.2623268719, 657966.9571823245, 630666.1293970842, 604759.4169545919,
 584012.7399027708, 571462.5592098474]
>>>
```

Figure 23: Results for execution of naive implementation with 20 partitions over 10 iterations.

20 Partition 20 Iteration

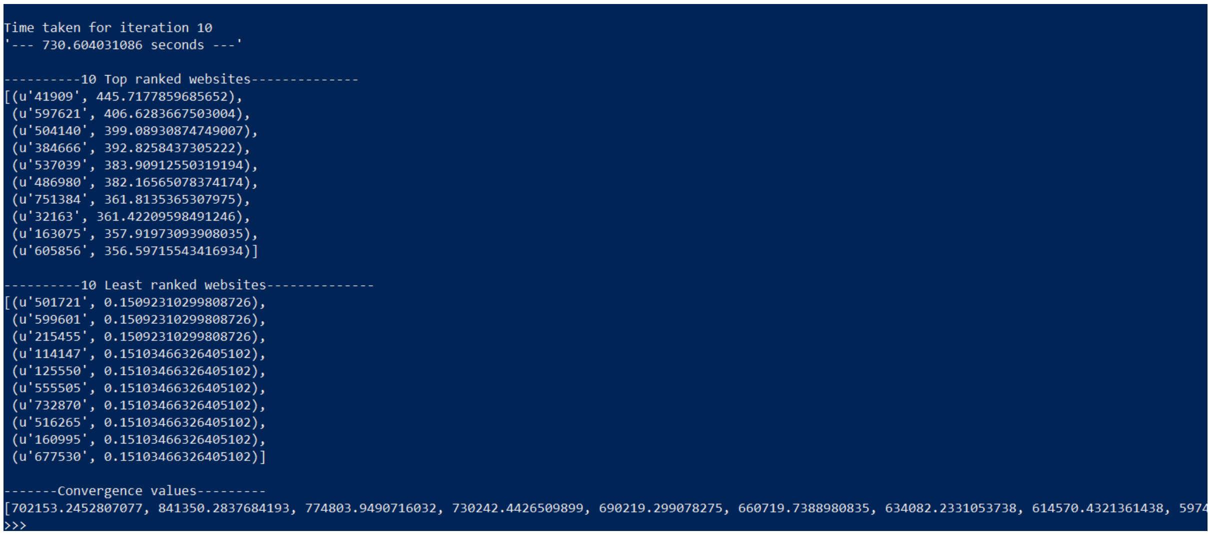


```
[Stage 5797:=====>                                                 (2 + 18) /
[Stage 5797:=================>                                     (7 + 13) /


Time taken for iteration 20
'--- 799.232561111 seconds ---'

----------10 Top ranked websites--------------
[Stage 5845:==========>                                            (4 + 16) /
[Stage 5845:===============================================>       (18 + 2) /

[Stage 5893:================>                                      (6 + 14) /

[Stage 5941:==>                                                    (1 + 19) /
[Stage 5941:==========================>                           (10 + 10) /

[(u'41909', 400.4451903203502),
 (u'504140', 375.5010819032477),
 (u'597621', 369.66462363547106),
 (u'384666', 362.6365609156543),
 (u'486980', 351.7344643027622),
 (u'537039', 345.7551694524356),
 (u'751384', 340.50179506966623),
 (u'32163', 339.116662145423),
 (u'765334', 331.49027957203623),
 (u'605856', 330.5233645665117)]

----------10 Least ranked websites--------------
[Stage 6089:=================>                                     (7 + 13) /
[Stage 6089:===============================================>       (18 + 2) /

[(u'215455', 0.15087165355344748),
 (u'501721', 0.15087165355344748),
 (u'599601', 0.15087165355344748),
 (u'160995', 0.15103315202475898),
 (u'791445', 0.15103315202475898),
 (u'290966', 0.15103315202475898),
 (u'906256', 0.15103315202475898),
 (u'828893', 0.15103315202475898),
 (u'720192', 0.15103315202475898),
 (u'58474', 0.15103315202475898)]

-------Convergence values---------
[702153.2452807325, 839350.2842495015, 775559.3208348146, 730626.4829108355,
 689931.2623268719, 657966.9571823245, 630666.1293970842, 604759.4169545919,
 584012.7399027708, 571462.5592098474, 573226.8971977594, 564083.0530788469,
 542037.1441821512, 532935.2371343566, 529574.7728870803, 517227.7403340709,
 525395.5010080694, 511384.7291224209, 513309.6071142852, 510444.5033707589]
>>>
```

Figure 24: Results for execution of naive implementation with 20 partitions over 20 iterations.

5.5 PageRank Naive Without Partitions:

Below figure shows the statistics for Naive implementation running for 10 Iterations without partitions.



```
Time taken for iteration 10
'--- 730.604031086 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 445.7177859685652),
 (u'597621', 406.6283667503004),
 (u'504140', 399.08930874749007),
 (u'384666', 392.8258437305222),
 (u'537039', 383.90912550319194),
 (u'486980', 382.16565078374174),
 (u'751384', 361.8135365307975),
 (u'32163', 361.42209598491246),
 (u'163075', 357.91973093908035),
 (u'605856', 356.59715543416934)]

----------10 Least ranked websites--------------
[(u'501721', 0.15092310299808726),
 (u'599601', 0.15092310299808726),
 (u'215455', 0.15092310299808726),
 (u'114147', 0.15103466326405102),
 (u'125550', 0.15103466326405102),
 (u'555505', 0.15103466326405102),
 (u'732870', 0.15103466326405102),
 (u'516265', 0.15103466326405102),
 (u'160995', 0.15103466326405102),
 (u'677530', 0.15103466326405102)]

-------Convergence values---------
[702153.2452807077, 841350.2837684193, 774803.9490716032, 730242.4426509899, 690219.299078275, 660719.7388980835, 634082.2331053738, 614570.4321361438, 5974
>>>
```

Figure 25: Results for naive implementation method for 10 iterations and no partitions.

Also by comparing the result of with and without partition for both Power Iteration and Naive Implementation we are getting the same result for the top 10 ranked website and least 10 rank website so it shows that consistency of algorithm.

Iteration	Time taken in sec
10	733.0633
20	1324.95

Table 4: Time taken by by the naive implementation without partitions over 20, 25 and 40 iterations

The bar graph demonstrated below, compares time taken for the execution of naive method without partitions over 20, 25 and 40 iterations.

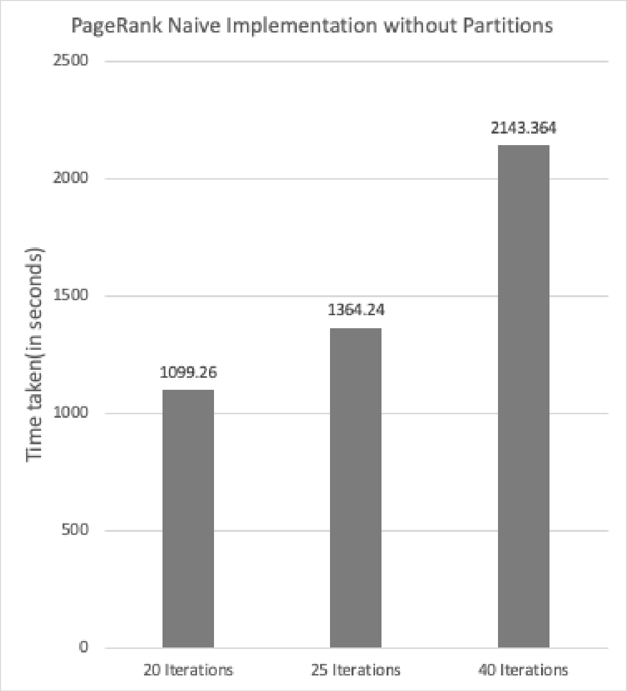
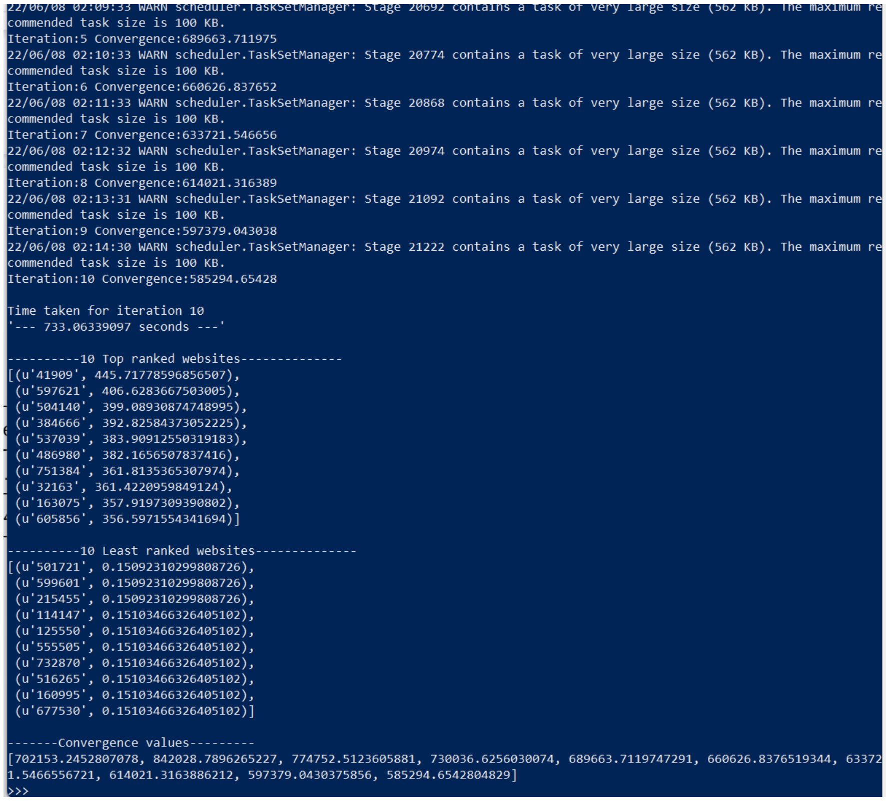


Figure 26: Graph comparing the time taken by the naive implementation without partitions over 20, 25 and 40 iterations

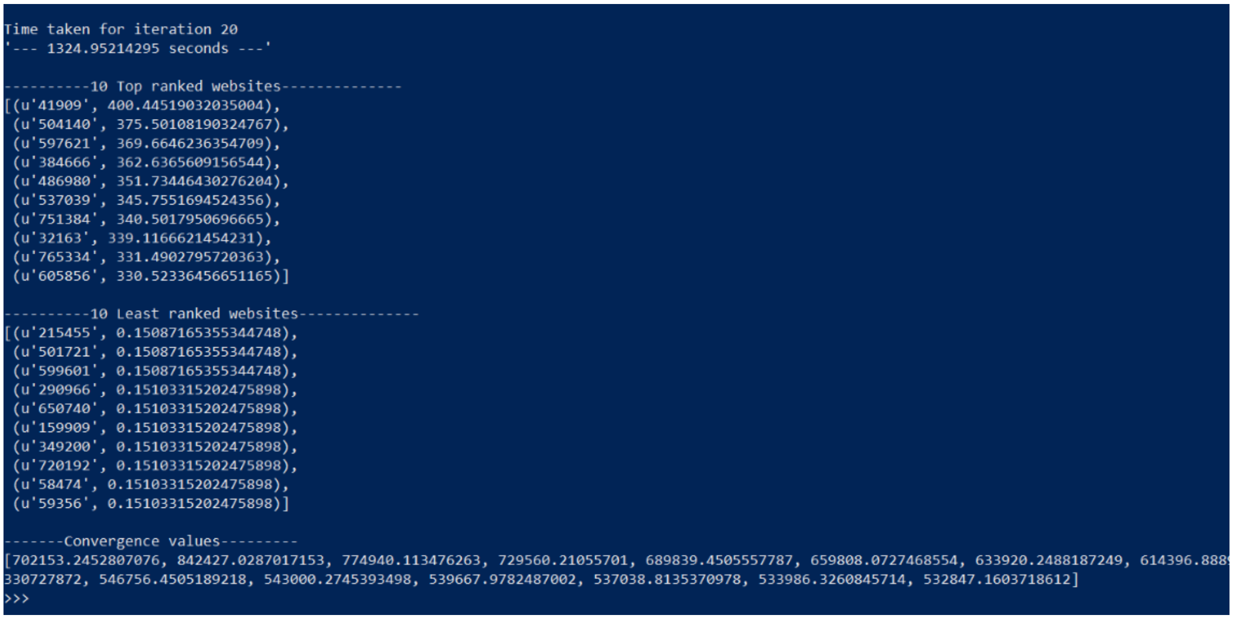
10 Iterations



```
22/06/08 02:09:33 WARN scheduler.TaskSetManager: Stage 20692 contains a task of very large size (562 KB). The maximum re
commended task size is 100 KB.
Iteration:5 Convergence:689663.711975
22/06/08 02:10:33 WARN scheduler.TaskSetManager: Stage 20774 contains a task of very large size (562 KB). The maximum re
commended task size is 100 KB.
Iteration:6 Convergence:660626.837652
22/06/08 02:11:33 WARN scheduler.TaskSetManager: Stage 20868 contains a task of very large size (562 KB). The maximum re
commended task size is 100 KB.
Iteration:7 Convergence:633721.546656
22/06/08 02:12:32 WARN scheduler.TaskSetManager: Stage 20974 contains a task of very large size (562 KB). The maximum re
commended task size is 100 KB.
Iteration:8 Convergence:614021.316389
22/06/08 02:13:31 WARN scheduler.TaskSetManager: Stage 21092 contains a task of very large size (562 KB). The maximum re
commended task size is 100 KB.
Iteration:9 Convergence:597379.043038
22/06/08 02:14:30 WARN scheduler.TaskSetManager: Stage 21222 contains a task of very large size (562 KB). The maximum re
commended task size is 100 KB.
Iteration:10 Convergence:585294.65428

Time taken for iteration 10
'--- 733.06339097 seconds ---'

----------10 Top ranked websites-------------
[(u'41909', 445.71778596856507),
 (u'597621', 406.6283667503005),
 (u'504140', 399.08930874748995),
 (u'384666', 392.82584373052225),
 (u'537039', 383.90912550319183),
 (u'486980', 382.1656507837416),
 (u'751384', 361.8135365307974),
 (u'32163', 361.4220959849124),
 (u'163075', 357.9197309390802),
 (u'605856', 356.5971554341694)]

----------10 Least ranked websites-------------
[(u'501721', 0.15092310299808726),
 (u'599601', 0.15092310299808726),
 (u'215455', 0.15092310299808726),
 (u'114147', 0.15103466326405102),
 (u'125550', 0.15103466326405102),
 (u'555505', 0.15103466326405102),
 (u'732870', 0.15103466326405102),
 (u'516265', 0.15103466326405102),
 (u'160995', 0.15103466326405102),
 (u'677530', 0.15103466326405102)]

-------Convergence values---------
[702153.2452807078, 842028.7896265227, 774752.5123605881, 730036.6256030074, 689663.7119747291, 660626.8376519344, 63372
1.5466556721, 614021.3163886212, 597379.0430375856, 585294.6542804829]
>>>
```

Figure 27:Results for naive implementation method for 10 iterations and no partitions.

20 Iterations

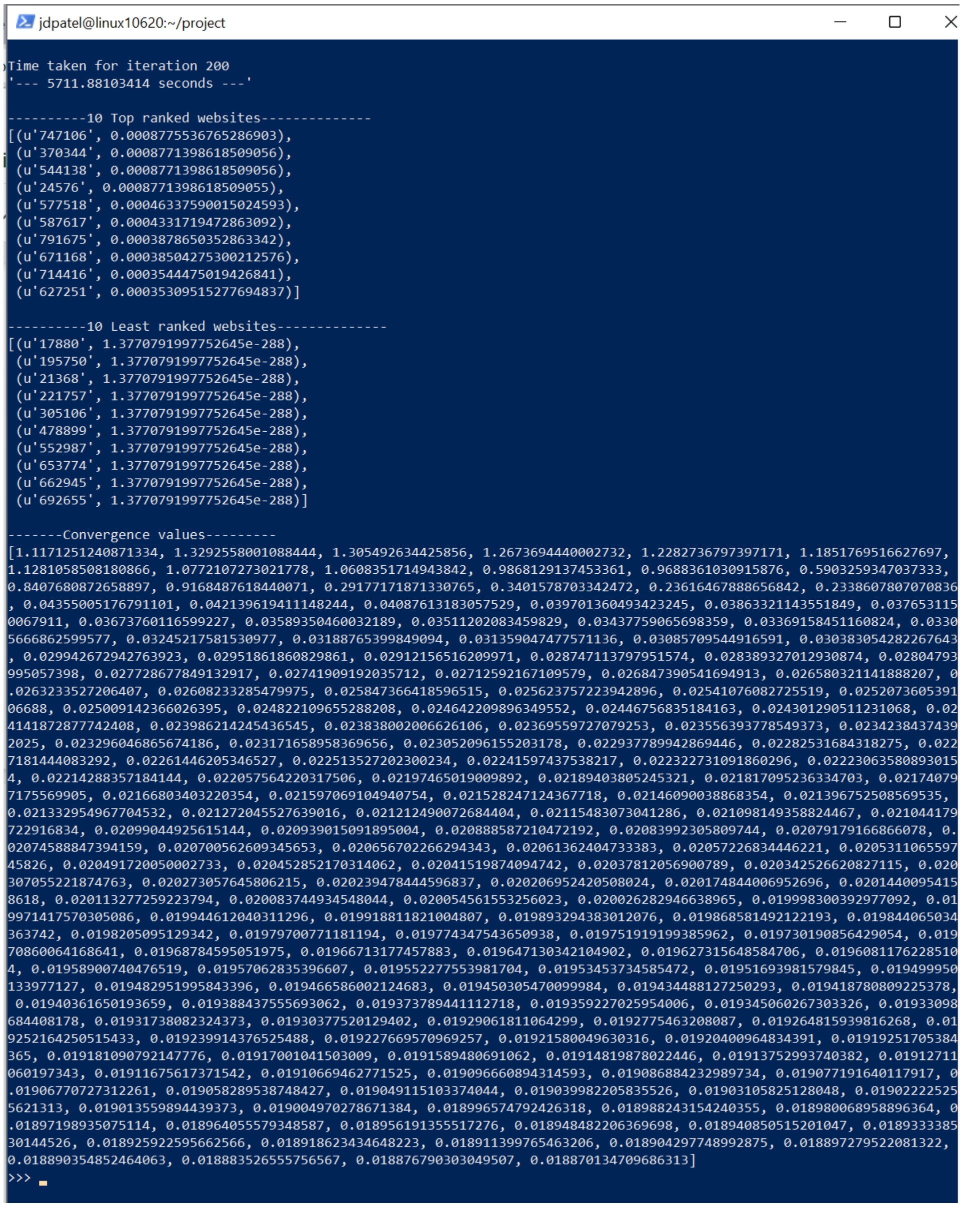


```
Time taken for iteration 20
'--- 1324.95214295 seconds ---'

----------10 Top ranked websites--------------
[(u'41909', 400.44519032035004),
 (u'504140', 375.50108190324767),
 (u'597621', 369.6646236354709),
 (u'384666', 362.6365609156544),
 (u'486980', 351.73446430276204),
 (u'537039', 345.7551694524356),
 (u'751384', 340.5017950696665),
 (u'32163', 339.1166621454231),
 (u'765334', 331.4902795720363),
 (u'605856', 330.52336456651165)]

----------10 Least ranked websites--------------
[(u'215455', 0.15087165355344748),
 (u'501721', 0.15087165355344748),
 (u'599601', 0.15087165355344748),
 (u'290966', 0.15103315202475898),
 (u'650740', 0.15103315202475898),
 (u'159909', 0.15103315202475898),
 (u'349200', 0.15103315202475898),
 (u'720192', 0.15103315202475898),
 (u'58474', 0.15103315202475898),
 (u'59356', 0.15103315202475898)]

-------Convergence values---------
[702153.2452807076, 842427.0287017153, 774940.113476263, 729560.21055701, 689839.4505557787, 659808.0727468554, 633920.2488187249, 614396.888
330727872, 546756.4505189218, 543000.2745393498, 539667.9782487002, 537038.8135370978, 533986.3260845714, 532847.1603718612]
>>>
```

Figure 28:Results for naive implementation method for 20 iterations and no partitions.

5.6 PageRank Power Iteration To Observe Spider Trap:

Below is the screenshot for 70 partitions and 200 Iterations. This experiment was done to observe the spider trap. It takes almost 5711.88 seconds which is 95 minutes to run this iteration. The convergence has almost reached 0.018 which means on doing further more iteration we might get 0 values for all the least ranked websites.



```
Time taken for iteration 200
'--- 5711.88103414 seconds ---'

----------10 Top ranked websites--------------
[(u'747106', 0.0008775536765286903),
 (u'370344', 0.0008771398618509056),
 (u'544138', 0.0008771398618509056),
 (u'24576', 0.0008771398618509055),
 (u'577518', 0.0004633759001502459 3),
 (u'587617', 0.00043317194728630 92),
 (u'791675', 0.0003878650352863342),
 (u'671168', 0.00038504275300212576),
 (u'714416', 0.0003544475019426841),
 (u'627251', 0.00035309515277694837)]

----------10 Least ranked websites--------------
[(u'17880', 1.3770791997752645e-288),
 (u'195750', 1.3770791997752645e-288),
 (u'21368', 1.3770791997752645e-288),
 (u'221757', 1.3770791997752645e-288),
 (u'305106', 1.3770791997752645e-288),
 (u'478899', 1.3770791997752645e-288),
 (u'552987', 1.3770791997752645e-288),
 (u'653774', 1.3770791997752645e-288),
 (u'662945', 1.3770791997752645e-288),
 (u'692655', 1.3770791997752645e-288)]

-------Convergence values---------
[1.1171251240871334, 1.3292558001088444, 1.305492634425856, 1.2673694440002732, 1.2282736797397171, 1.1851769516627697,
1.1281058508180866, 1.0772107273021778, 1.0608351714943842, 0.9868129137453361, 0.9688361030915876, 0.5903259347037333,
0.8407680872658897, 0.9168487618440071, 0.29177171871330765, 0.3401578703342472, 0.23616467888656842, 0.2338607807070836
, 0.04355005176791101, 0.04213961941114 8244, 0.04087613183057529, 0.03970136049342 3245, 0.03863321143551849, 0.037653115
0067911, 0.03673760116599227, 0.03589350460032189, 0.03511202083459829, 0.03437759065698359, 0.03369158451160824, 0.0330
5666862599577, 0.03245217581530977, 0.03188765399849094, 0.03135904747757 1136, 0.03085709544916591, 0.03038305428226764 3
, 0.02994267294276392 3, 0.02951861860829861, 0.02912156516209971, 0.02874711379795157 4, 0.02838932701293087 4, 0.02804793
995057398, 0.02772867784913291 7, 0.02741909192035712, 0.02712592167109579, 0.02684739054169491 3, 0.02658032114188820 7, 0
.026323352720640 7, 0.02608233285479975, 0.02584736641859651 5, 0.02562375722394289 6, 0.02541076082725519, 0.0252073605391
06688, 0.02500914236602639 5, 0.02482210965528820 8, 0.02464220989634955 2, 0.02446756835184163, 0.02430129051123106 8, 0.02
414187287774240 8, 0.02398621424543654 5, 0.02383800200662610 6, 0.02369559727079253, 0.02355639377854937 3, 0.0234238437439
2025, 0.02329604686567418 6, 0.02317165895836965 6, 0.02305209615520317 8, 0.02293778994286944 6, 0.02282531684318275, 0.022
7181444083292, 0.02261446205346527, 0.02251352720230023 4, 0.02241597437538217, 0.02232273109186029 6, 0.0222306358089301 5
4, 0.02214288357184144, 0.02205756422031750 6, 0.02197465019009892, 0.02189403805245321, 0.02181709523633470 3, 0.02174079
7175569905, 0.02166803403220354, 0.02159706910494075 4, 0.02152824712436771 8, 0.02146090038868354, 0.02139675250856953 5,
0.02133295496770453 2, 0.02127204552763901 6, 0.02121249007268440 4, 0.02115483073041286, 0.02109814935882446 7, 0.021044179
722916834, 0.02099044925615144, 0.02093901509189500 4, 0.02088858721047219 2, 0.02083992305809744, 0.02079179166866078, 0.
02074588847394159, 0.02070056260934565 3, 0.02065670226629434 3, 0.02061362404733383, 0.02057226834446221, 0.020531106559 7
45826, 0.02049172005000273 3, 0.02045285217031406 2, 0.02041519874094742, 0.02037812056900789, 0.02034252662082711 5, 0.020
307055221874763, 0.02027305764580621 5, 0.02023947844459683 7, 0.02020695242050802 4, 0.02017484400695269 6, 0.0201440095415
8618, 0.02011327725922379 4, 0.02008374493454804 4, 0.02005456155325602 3, 0.02002628294663896 5, 0.01999830039297709 2, 0.01
9971417570305086, 0.01994461204031129 6, 0.01991881182100480 7, 0.01989329438301207 6, 0.01986858149212219 3, 0.01984406503 4
363742, 0.019820509512934 2, 0.01979700771181194, 0.01977434754365093 8, 0.01975191919938596 2, 0.01973019085642905 4, 0.019
70860064168641, 0.01968784595051975, 0.01966713177457883, 0.01964713034210490 2, 0.01962731564858470 6, 0.0196081176228510
4, 0.01958900740476519, 0.01957062835396607, 0.01955227755398170 4, 0.01953453734585472, 0.01951693981579845, 0.019499950
133977127, 0.01948295199584339 6, 0.01946658600212468 3, 0.01945030547009998 4, 0.01943448812725029 3, 0.01941878080922537 8,
 0.01940361650193659, 0.01938843755569306 2, 0.01937378944111271 8, 0.01935922702595400 6, 0.01934506026730332 6, 0.01933098
684408178, 0.01931738082324373, 0.01930377520129402, 0.01929061811064299, 0.01927754632080 87, 0.01926481593981626 8, 0.01
9252164250515433, 0.01923991437652548 8, 0.01922766957096925 7, 0.01921580049630316, 0.01920400964834391, 0.01919251705384
365, 0.01918109079214777 6, 0.01917001041503009, 0.01915894806910 62, 0.01914819878022446, 0.01913752993740382, 0.01912711
060197343, 0.01911675617371542, 0.01910669462771525, 0.01909666089431459 3, 0.01908688423298973 4, 0.01907719164011791 7, 0
.0190677072731226 1, 0.01905828953874842 7, 0.01904911510337404 4, 0.01903998220583552 6, 0.01903105825128048, 0.0190222525
5621313, 0.01901355989443937 3, 0.01900497027867138 4, 0.01899657479242631 8, 0.01898824315424035 5, 0.01898006895889636 4, 0
.01897198935075114, 0.01896405557934858 7, 0.01895619135551727 6, 0.01894848220636969 8, 0.01894085051520104 7, 0.0189333385
30144526, 0.01892592259566256 6, 0.01891862343464822 3, 0.01891139976546320 6, 0.01890429774899287 5, 0.01889727952208132 2,
0.01889035485246406 3, 0.01888352655575656 7, 0.01887679030304950 7, 0.01887013470968631 3]
>>> 
```

Figure 29: Results of power iteration method over 200 iterations with 70 partitions.

6

6. Outcomes/Results

1. How parallelizing huge data using Mapper and reducer can help us to complete the task efficiently.

   We implemented two variations of page rank algorithm i.e Power iteration and Naive implementation both of them using partition and without partition. The google web graph dataset consists of 5 million of edges and to run the power iteration algorithm for 20 iterations it takes 1099.26 seconds. However, processing such data in parallel using 10 partitions takes 514 seconds, which determines that using parallel processing the task can be completed quickly.

2. Spider Trap and dead end observation in webgraph

   Spider Trap and dead end problems occur in power iteration where few websites at the end consume all the importance. Meaning when a random walker follows the path then after a certain amount of time he might visit the same website again and again and he might get bored. To observe the spider trap we did an experiment to run power iteration for 200 iterations where we can see that the convergence value almost reaches to 0.018 and the rank of the last 10 websites is near to 0 which shows that a random walker will not be able to reach those website.

3. While obtaining the top 10 pageranks, which webpage can be the most visited webpage?

   The webpage that has the most number of incoming links should be at the topmost part of the rankings. So for that, we observed the pageranks over different iterations with partitions as well as without partitions. From the acquired results, we can say that the webpage ‘41909’ should have the most number of outgoing links as the webpage ‘41909’ tops the rankings for power iteration method with 10 partitions as well as 20 partitions over 20 and 25 iterations respectively. The same observations were made for the power iteration method without partitions too. If we consider the naive implementation method, the webpage ‘41909’ ranks as the topmost webpage for 10 partitions as well as 20 partitions over 10 and 20 iterations respectively. Additionally, the same observations were made without partitions. To conclude, we can say that more important websites are more likely to gain links from other sites.

7

7. CONCLUSION

We implemented the PageRank Algorithm in this project using PySpark and calculated the ranks of different nodes using the Power Iteration and Naive Implementation techniques.

Although the Power Iteration technique provides the ranks of all nodes, it is vulnerable to spider traps and dead ends, which are bound to come across on the Internet. In fact, some of the ranks became zero after 250 iterations due to spider traps. As a result, we used Naive Implementation with a damping factor 'd' of 0.85 to avoid the impact of spider traps and dead ends. We were also able to observe that the node with the most incoming nodes has a higher rank.

Furthermore, we can clearly notice that having more number of partitions takes less time to complete the execution than having none. The time required to calculate the page ranks of the nodes decreases as the number of partitions increases which highlights the importance and efficiency of parallel computation in a real time scenario.

8

8. REFERENCES

https://techblogmu.blogspot.com/2018/03/explain-how-dead-ends-are-handled-in_22.html#:~:text=A%20dead%20end%20is%20a,nodes%20with%20no%20arcs%20out.

https://snap.stanford.edu/data/web-Google.html

https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf

https://www.link-assistant.com/news/google-pagerank-algorithm.html