Harsh Shukla
CS6240-Section-01 Assignment-4
PageRank Using Scala

a) Describe briefly how each step of your program is transforming the data. Be precise, e.g., by showing the structure of the input and output as a table.

1. The step parses each line using pa parser and creates an RDD
   With effect to data we filter the invalid nodes from the overall data.
2. Combine all the nodes in an adjacency list of all nodes that are present in the graph. Create a new RDD and make a union with the completeRDD.Next we merge all the elements in the adjacency list with the same keys.
3. Now we put the mapValues over the complete RDD initializing the pageRank to 1/totalnodes on the graph. Create a new RDD for combined graph
4. We need to keep the dangling node offset and therefore we keep a counter for multiple iterations.
5. This step maps over the combinedGraph extract page rank and adjacency list from each node Update the dangling counter if you find one. If no dangling counter is found distribute page rank across all the nodes in the adjacency list (outlinks) and creates a RDD contribution graph.
6. Join the combinedGraph with contributionGraph, now extract sum of contributions from inlinks to a page from contributionGraph RDD to calculate a new page rank value and now create a new RDD and read next iterations
7. Create a new RDD top100 with pageName and Rank from combinedGraph
   Take top 100 from every partition and put them out globally.

Steps	Input Format	Output Format
1	RDD[(String, (Double, List[String]))]	`RDD[(pageName, (pageRank, adjacencyList))]`
2	`RDD[(String, (Double, List[String]))]`	`RDD[(pageName, (pageRank, adjacencyList))]`
3	`RDD[(String, (Double, List[String]))]`	`RDD[(pageName, (pageRank, adjacencyList))]`
4	N/a	N/A
5	`RDD[(String, Double)]`	`RDD[(outlink, contribution)]`
6	`RDD[(String, (Double, List[String]))]`	`RDD[(pageName, (pageRank, adjacencyList))]`
7	N/A	N/a

b) For each step, state if the dependency is narrow (no shuffling) or wide (shuffling). How many stages does your Spark have?

Shuffling is a process in which we move data between nodes.

Steps(from a)	Dependecy
1	Narrow
2	Wide
3	Wide
4	Narrow
5	Wide
6	Wide
7	Narrow

There are total **142 stages** for this mapreduce program.

c) Report for both configurations the Spark execution time. For comparison, also include the total execution time (from pre-processing to top-k) of the corresponding Hadoop executions from Assignment 3.

Approach(6 machines)	**Timings**
Spark	**27 minutes**
Mapreduce	**26 minutes**
Approach(11 machines)	**Timings**
Spark	**12 minutes**
Mapreduce	**17 minutes**

d) Discuss which system is faster and briefly explain what could be the main reason for this performance difference.

As seen from statistics in step 3 spark program is faster with 11 machines but is almost equal in 6 machines. This might be due to:

i. Spark has no spawining while Mapreduce needs spawing all the time.
ii. Spark has a caches memory for data and does not depend on middle stage like mapreduce . Mapreduce stores intermediate data in HDFS.
iii. The slower performance for smalled configuration could be because of container allocation strategy in yarn, where we stop for existing execution to complete before executing new containers.

Top 100 pages Local:

(0.004234526888762993,United_States_09d4)
(0.0032433377379234036,Wikimedia_Commons_7b57)
(0.002627999167791734,Country)
(0.001806477634398407,England)
(0.0017731764318439285,United_Kingdom_5ad7)
(0.0017722308538755678,Europe)
(0.001748914319614091,Water)
(0.0017100798171341646,France)
(0.0016764013330700413,Animal)
(0.0016497038227778864,Earth)
(0.001633723120474448,Germany)
(0.0015546859267226711,City)
(0.0014184266012062614,Week)
(0.0013269914987875209,Sunday)
(0.0013118004739144065,Asia)
(0.001307810476570479,Monday)
(0.001295326851439692,Wednesday)
(0.0012632780065331543,Friday)
(0.0012495625194945366,Saturday)
(0.0012495344861278002,Money)
(0.0012369578263260782,Wiktionary)
(0.0012334033796575557,Thursday)
(0.0012246343503808967,index)
(0.001224464416010219,Tuesday)
(0.00120132438682016,Plant)
(0.0011962420178951779,Computer)
(0.0011734933034579203,English_language)
(0.00116101609097121,Italy)
(0.0011581521532315988,Government)
(0.0011455967532039365,India)
(0.0010970742669945991,Number)
(0.001047500255070609,Spain)
(0.0010358250045802786,Day)
(0.0010038631917882678,Canada)
(9.884172815555095E-4,Japan)
(9.79386346471023E-4,People)
(9.680417793975043E-4,Human)
(9.400919000831374E-4,Australia)
(9.328162797563749E-4,Wikimedia_Foundation_83d9)
(9.215565736588138E-4,China)
(9.060487433447444E-4,Energy)
(8.903793391327451E-4,Food)
(8.884154440252286E-4,Sun)
(8.823713905787439E-4,Science)
(8.603778205554602E-4,Mathematics)
(8.180511293943909E-4,Television)
(8.113099334476417E-4,Capital_(city))
(8.095719913595919E-4,Russia)
(8.038188135014618E-4,Year)
(7.866447622101328E-4,Music)

(7.720494485992087E-4,State)
(7.598741956147506E-4,Language)
(7.404784079843066E-4,Metal)
(7.292280303606577E-4,Wikipedia)
(7.214976296810177E-4,Greek_language)
(7.140640254724753E-4,Religion)
(7.129487636296815E-4,2004)
(7.036263721483927E-4,London)
(7.025957879541771E-4,Sound)
(7.01150839025582E-4,Planet)
(7.00930368300168E-4,Scotland)
(6.762121160416637E-4,Greece)
(6.740268798293555E-4,Africa)
(6.658632145849031E-4,20th_century)
(6.517308198823369E-4,19th_century)
(6.490682309701924E-4,Law)
(6.383721617869327E-4,Geography)
(6.302658497718048E-4,World)
(6.298943421202023E-4,Liquid)
(6.201067410081968E-4,Scientist)
(6.198171535268883E-4,Society)
(5.987698975269798E-4,Atom)
(5.945455219782833E-4,Latin)
(5.932679667621986E-4,War)
(5.899551558753892E-4,History)
(5.879439510742463E-4,Light)
(5.793055741833649E-4,Building)
(5.769711460292272E-4,Netherlands)
(5.766584093090907E-4,Culture)
(5.72615980686873E-4,God)
(5.718239897463178E-4,Centuries)
(5.583506186934667E-4,Turkey)
(5.582737269616431E-4,Plural)
(5.574747315438643E-4,Information)
(5.56827516489253E-4,Chemical_element)
(5.521750445977727E-4,Poland)
(5.450980706858818E-4,Sweden)
(5.407257402742484E-4,Portugal)
(5.253272645980527E-4,Disease)
(5.230571660641617E-4,Denmark)
(5.229938046869019E-4,Austria)
(5.207276774846071E-4,Species)
(5.194029987569419E-4,North_America_e7c4)
(5.169569580550111E-4,Capital_city)
(5.128788198653849E-4,Cyprus)
(5.120847930680905E-4,Ocean)
(5.064606764336833E-4,List_of_decades)
(5.06253878081463E-4,Biology)
(5.036503585256746E-4,Book)
(5.009888830571796E-4,Moon)