CS6240 Parallel Programming
Section 2 – Homework 4
Ajay Vardhan

The whole program consists of 3 steps:

1. Parsing
2. PageRank calculation
3. TopK

Parsing:

First we read the bz2 input files using scala and ivoke the map function from spark on this file. This map executes whatever function we specify inside it's parameters on each line of the input file.

```
val tempList = sc.textFile(input)
  .map(line => {…})
```

Inside this map, we call the SAX Parser from the previous assignment, which will parse each line and return the page name with it's adjacency list. We then store that as a key-value pair –> *(pageName, list(adjacencyList))* in the **adjList** RDD. We then iterate through the adjacencyList for each page and store that *(linkName, EmptyAdjacencyList)* in the same RDD. This step will allow us to catch hold of the dangling nodes in the links. We can then reduce this RDD by key and append the lists for each page, which will give us complete list of pages and the dangling nodes along with their adjacency list. We can then store this RDD in our memory using `.persist(StorageLevel.MEMORY_AND_DISK)` as we will be using this RDD for our future processing.

PageRank Calculation:

The output from the parser will be sent to the pagerank calcuator function which would first append the RDD with the initial pageranks for all pages. This can be done with a simple map. Now this RDD's values are sent to a flatMap and the temporary page rank (pagerank of the incoming page/size of adjacency list) is calculated for each link in the adjacency list. We also accumulate the delta value during this map. This is then reduced by key such that each page has the accumulated pagerank. This is stored in the **tempRanks** RDD. The original adjacency list, without any pageranks, is then left joined with this tempRanks so that we the temporary page ranks for all the pages along with it's adjacency list and pagerank. We can then perform a map in this joined RDD to calculate the actual pagerank with the formula. These set of operations are performed for 10 iterations and returned to the main function. The final output will be of the format RDD: *(pageRank, (adjacencyList, pagerank))*.

TopK:

Once we get the pageranks output RDD, we perform a map on this so that we have the pagerank as the key and the page name as the value *(pageRank → PageName)*. This is then sorted by key and the top 100 is taken from this sorted RDD. This can be written to the output file.

Comparison between Scala Spark and MapReduce program:

The configuration steps are similar for both programs. Since there is no job initiation in Spark, there are no equivalent steps for the MR job steps in the Scala Program. Here's the comparison between rest of the program:

- Reading input: Scala reads the input files using *sc.textFile(input)*, while the MR program adds the input path to the job, which will in-turn read the input files while the job is being executed.
- *.map(line => {…})* in Scala reads each line from the input and executes the parser on each line. This is done using the ParserMap in MapReduce. The output from the parser in Scala is then divided into pagename and adjacency list and emitted as *(pageName, Links)*. Each link in the Links for each page is emitted as *links.map(link => (link,l))*. In Mapreduce this is done inside ParserMap in lines *context.write(new Text(pageName),new Text(output));* and *context.write(new Text(link),new Text(""));* There is no difference between the two programs for this statement since both do the same thing. We try to find the dangling nodes by emitting each link with an empty list in both programs and reduce by key.
- *adjList.map(page=>(page._1,(page._2,initialPageRank)))* initializes the adjacency list with the initial pageranks for all the pages. MapReduces does not initialize list, but rather calculates the page rank in the Map phase and appends that to the node object of each page. Scala is better here since we don't create a separate object for each page and we can just work with whatever RDD we have in the memory. MapReduce program uses extra memory for this phase. Scala also stores this data in the memory and we can just access it directly in the future steps but mapreduce stores this in the disk which needs to be extracted each time.
- *delta = sc.doubleAccumulator* is used in Scala to store the global accumulator for Delta value. In Mapreduce, we have a global counter which is accessed each time in Mapper. This counter is also updated each time the mapper comes across the dangling node. This is a lot of I/O operations to the global counter which is highly inefficient. Once we update a value in the global counter, the actual value is updated only after the job gets over. This makes it difficult for us to access the updated value in the same job. But the accumulator in Scala updates the value in real time and can be accessed any where any time.

- *pageRanks.values.flatMap(pages =>{..})* is used in Scala to iterate through the values of the updated adjacency list RDD with their pageranks to calculate the pagerank for each link for each page. MapReduce does this by creating a node for each page, and for each link in the page and then emitting it to the reducer, which will in turn calculate the final pagerank. *context.write(new Text(page[0]),node);* and *context.write(new Text(s),new WritableComparableObject(new Text(df.format(p))));* for each page in MapReduce. In Scala, we can then immediately ReduceByKey this RDD to find the intermediate pagerank values for all the links in the pages using *reduceByKey(_+_)*. In MapReduce, we get each value for all the pages and links in the reducer, accumulate the pageranks for the links and calculate the pagerank for each page in the reducer. Scala joins the intermediate pageranks RDD with the initial adjacencyList and maps the RDD to find the final pageranks for all the pages. In MapReduce, the output for each iteration of the pagerank calculation program is stored in files in the disk, which is then retrieved for the next iteration, whereas in Scala Spark the output is stored in memory which can be accessed easily without any extra processing.
- To find the Top 100 pages, we had to process the whole list once again and find the local top 100 in each mapper using a shared Treemap for each mapper, send these mappers to the reducer and find the top 100 from all these local winners using another shared treemap. This is a lot of processing and memory waste. Scala finished it with *pageRanks.map(page => (page._2._2,page._1)).sortByKey(false).take(100)*. Here the page ranks are just sorted by their pages and the top 100 is emitted. There is no need for any extra RDDs or IO operations.

Running times:

6 Machines:

	6 Machines	**11 Machines**
MapReduce	1:07:23	35:32
Scala Spark	1:17:44	39:43

I expected the Spark program to outperform the MapReduce program but Spark turned out to be a little slower. This could be because Spark evaluates everything lazily and the data is mostly stored in the memory instead of the Disk. If thethe input data or the intermediate RDDs exceed the memory, it will store the extra data in the Disc. The Data is also re-calculated every time. There could be a little extra buffer time to switch between Memory and Disc loading. I used string to store my adjacency list for MapReduce program, but I use a list in Spark. This could also cause some delay in processing.

Top 100 pages:

I got the same results for both the programs. The display was different since I used a double formatter in my MapReduce program to display just the first 16 digits of the page rank. But the pages were all the same. Since I use the same formula and the concept for both the programs, there weren't any differences in the outputs.

MapReduce:

Full dataset:

United_States_09d4 : 0.0004963795733764
Biography : 0.0003012077285819
United_Kingdom_5ad7 : 0.0001985218456853
2006 : 0.0001967050596331
Geographic_coordinate_system : 0.0001810134002697
England : 0.0001657455332593
Canada : 0.000156532209805
2005 : 0.000148691532904
Record_label : 0.0001255303950844
Australia : 0.0001206709517121
Music_genre : 0.0001206665271824
2004 : 0.0001189692747664
France : 0.0001181584262688
India : 0.0001158430694202
Internet_Movie_Database_7ea7 : 0.0001137600727995
Germany : 0.0001101072930085
2003 : 0.0001002594762911
Japan : 0.0000969705936161
Population_density : 0.00009217051081752001 : 0.0000833100969702
Politician : 0.0000800081273703
Europe : 0.0000797775273589
2002 : 0.0000797273248454
Football_(soccer) : 0.0000763385241577
2000 : 0.0000760270201774
Scientific_classification : 0.0000749549376807
Record_producer : 0.0000742734630802
Studio_album : 0.0000738294728309
Census : 0.0000713308463664
Personal_name : 0.0000694697831563
Album : 0.0000679204432178
World_War_II_d045 : 0.0000677468137214
London : 0.0000675260101411
1999 : 0.0000658372671755
Television : 0.0000649603054051

Italy : 0.0000632999590828
1998 : 0.0000592929796894
Actor : 0.0000585816246104
Marriage : 0.0000580508978683
Public_domain : 0.0000579728465948
Square_mile : 0.0000575710550211
Km2 : 0.000055854339381
Per_capita_income : 0.0000556790255762
1997 : 0.0000553872588404
United_States_Census_Bureau_2c85 : 0.0000551081834618
Poverty_line : 0.0000549797143183
Spain : 0.0000544812425546
Scotland : 0.0000534982454683
California : 0.0000534542376655
1996 : 0.0000528715241651
English_language : 0.0000526639750111
Wiktionary : 0.0000523641380725
Film : 0.0000521763279065
Animal : 0.0000516117092817
Population : 0.0000512440517563
White_(U.S._Census)_c45a : 0.0000494439668716
Sweden : 0.0000492691052152
New_York_City_1428 : 0.0000486016428744
1995 : 0.0000484713954455
School : 0.0000484318111996
Writer : 0.0000471650510085
Russia : 0.0000471089555599
New_York_3da4 : 0.0000469866923321
1994 : 0.0000463754422042
China : 0.0000463458068601
New_Zealand_2311 : 0.0000461922528967Norway : 0.0000454794676142
1993 : 0.0000446881341361
1992 : 0.0000424952748481
1990 : 0.0000420617489019
Poet : 0.0000420105128676
1991 : 0.0000418775003426
Brazil : 0.0000416188783121
Corporation : 0.0000415694286005
Latino_(U.S._Census)_5f0e : 0.0000410181664038
Hispanic_(U.S._Census)_1387 : 0.0000410077938645
USA_f75d : 0.0000408666192044
Ireland : 0.0000408092119751
Website : 0.0000400008334025
Poland : 0.0000397801463561
Binomial_nomenclature : 0.0000391930967244
Netherlands : 0.000038916591021
1989 : 0.0000386666743122

Race_(United_States_Census)_a07d : 0.0000382288887407
1980 : 0.0000377687530177
Company_(law) : 0.0000375334683054
Building : 0.0000369472587079
Band_(music) : 0.0000368474153688
1982 : 0.0000365102120758
All_Music_Guide_0e49 : 0.0000363928255755
1986 : 0.0000363921898214
1983 : 0.0000361912995949
Native_American_(U.S._Census)_1a7a : 0.000036034877043
1981 : 0.0000360060560732
1985 : 0.0000359958806123
1984 : 0.0000359827000107
1987 : 0.0000358487948299
1988 : 0.0000355106395725
Rock_music : 0.0000351555081881
1979 : 0.0000351214038387

Simple Dataset:

United_States_09d4 : 0.0010003884949175
Wikimedia_Commons_7b57 : 0.0008400444266429
England : 0.0007040868171239
Germany : 0.0006426416935204
France : 0.0004798383492342
City : 0.0004138498523064
Inhabitant : 0.0004076637959362
Wiktionary : 0.0003646176245935
Country : 0.0003543602532404Animal : 0.0003497121922316
Japan : 0.0003356708472961
United_Kingdom_5ad7 : 0.0003351307225881
Computer : 0.000333962186426
Water : 0.0003093583570966
Europe : 0.0003046168373541
India : 0.0003040704711936
Spain : 0.0002862751020975
Australia : 0.0002858460526661
English_language : 0.0002821516939026
Italy : 0.0002820428932741
Canada : 0.0002760210802776
Television : 0.0002733396362203
Plant : 0.0002647377781386
Earth : 0.000264361326668
London : 0.000241804561945
Money : 0.0002415552986099
China : 0.0002398613512482
Greece : 0.0002358295243245

Music : 0.0002350365942618
Scotland : 0.000234830931415
Food : 0.0002321834442976
Football_(soccer) : 0.0002308334468301
Capital_(city) : 0.0002248335958143
Human : 0.0002224587795103
Metal : 0.0002218690514029
Capital_city : 0.0002161496867612
Mathematics : 0.0002134476779881
Movie : 0.0002127455974942
Netherlands : 0.0002116566237034
Government : 0.0002080389750704
Russia : 0.0002048664428054
Brazil : 0.0002044031752767
U.S._state_5a68 : 0.0002043228133021
Number : 0.0002037510022386
Greek_mythology : 0.0002035912073174
Book : 0.0002035837607729
People : 0.0002034718625752
2005 : 0.0002008819915561
Poland : 0.0001985161664598
2004 : 0.0001976206914645
Language : 0.0001975864646322
2006 : 0.0001945903408034
Religion : 0.0001899476004632
Year : 0.0001890388262255
Actor : 0.000187064591437
God : 0.0001840618030375Asia : 0.0001840504435217
California : 0.0001825962727973
Sweden : 0.0001823023522473
Science : 0.0001807172881675
University : 0.0001799836234908
19th_century : 0.0001790154809193
Fruit : 0.0001749876335488
Car : 0.0001711875989002
Chemical_element : 0.0001684642359408
Africa : 0.0001684361911754
Disease : 0.0001660813138709
Film : 0.0001659795923117
Internet : 0.0001653987455037
World_War_II_d045 : 0.0001651838915599
Species : 0.0001646381721874
Latin : 0.0001638989953097
Company : 0.000162517003906
River : 0.0001596566912962
North_America_e7c4 : 0.0001590850537459
Fish : 0.0001585263020904

20th_century : 0.0001571584250426
Liquid : 0.0001549390952479
1970s : 0.000154833245238
Island : 0.0001544823696856
Centuries : 0.0001540831578346
Greek_language : 0.0001539532754094
Internet_Movie_Database_7ea7 : 0.0001528517618542
Video_game : 0.0001519498045559
Sport : 0.0001515700981988
War : 0.0001504317264891
1960s : 0.0001475172482981
Mammal : 0.0001471607626914
Christianity : 0.0001471008082695
German_language : 0.0001467105285373
Law : 0.000146658334246
Prefecture : 0.000145747812797
Sun : 0.0001451004709616
County : 0.0001445944509396
Singer : 0.0001442421610144
State : 0.000143081130508
Tree : 0.00014278679846
Austria : 0.0001427544656804
Chad : 0.0001421239311374
Child : 0.0001414240428774

Scala Spark:

Full Dataset:

(4.73289576636963E-4,United_States_09d4)
(2.8720139627831784E-4,Biography)
(1.892873759278145E-4,United_Kingdom_5ad7)
(1.8755573686475598E-4,2006)
(1.7265597433856295E-4,Geographic_coordinate_system)
(1.5803550842014144E-4,England)
(1.4925070021961438E-4,Canada)
(1.417748921549489E-4,2005)
(1.196908140573331E-4,Record_label)
(1.150576869620796E-4,Australia)
(1.1505609886911766E-4,Music_genre)
(1.1343534910015323E-4,2004)
(1.1267070865259796E-4,France)
(1.1045437009586346E-4,India)
(1.084696630209505E-4,Internet_Movie_Database_7ea7)
(1.0498618601334561E-4,Germany)
(9.559550046750984E-5,2003)

(9.245981568529812E-5,Japan)
(8.78829208452897E-5,Population_density)
(7.943456306893284E-5,2001)
(7.628632914580003E-5,Politician)
(7.606667435068499E-5,Europe)
(7.601844146371114E-5,2002)
(7.278744709163133E-5,Football_(soccer))
(7.249042967605399E-5,2000)
(7.146827406170103E-5,Scientific_classification)
(7.081829087576523E-5,Record_producer)
(7.039676926949181E-5,Studio_album)
(6.801269456872103E-5,Census)
(6.625439250392709E-5,Personal_name)
(6.476088913990664E-5,Album)
(6.459541059319169E-5,World_War_II_d045)
(6.438599618414397E-5,London)
(6.277462067688473E-5,1999)
(6.193851625367969E-5,Television)
(6.035630707368231E-5,Italy)
(5.653498523787382E-5,1998)
(5.585654988866244E-5,Actor)
(5.535050529625594E-5,Marriage)
(5.5276108153333874E-5,Public_domain)
(5.4892997717306676E-5,Square_mile)
(5.3088964662143444E-5,Per_capita_income)
(5.281076506476307E-5,1997)
(5.254467800632953E-5,United_States_Census_Bureau_2c85)
(5.242218363216612E-5,Poverty_line)
(5.195171955853407E-5,Spain)
(5.100966280136976E-5,Scotland)
(5.0967692149440046E-5,California)
(5.041205448093159E-5,1996)
(5.021436107812264E-5,English_language)
(4.995188152830858E-5,Wiktionary)
(4.9749222589622435E-5,Film)
(4.921087415061826E-5,Animal)
(4.8859988546577625E-5,Population)
(4.7143944870991804E-5,White_(U.S._Census)_c45a)
(4.698937686355335E-5,Sweden)
(4.634083866671303E-5,New_York_City_1428)
(4.6216660386396744E-5,1995)
(4.6178874969304005E-5,School)
(4.497104879889741E-5,Writer)
(4.4917588681812244E-5,Russia)
(4.480094863296931E-5,New_York_3da4)
(4.4218167470877886E-5,1994)
(4.4190013566196996E-5,China)

(4.404352128189344E-5,New_Zealand_2311)
(4.336710284213525E-5,Norway)
(4.260930380196898E-5,1993)
(4.0518480852069846E-5,1992)
(4.010532068013221E-5,1990)
(4.0056251737291125E-5,Poet)
(3.9929436988994235E-5,1991)
(3.968549767480156E-5,Brazil)
(3.9635792678004076E-5,Corporation)
(3.911009391493232E-5,Latino_(U.S._Census)_5f0e)
(3.910020375356856E-5,Hispanic_(U.S._Census)_1387)
(3.89656154447531E-5,USA_f75d)
(3.891123766371637E-5,Ireland)
(3.814014870676803E-5,Website)
(3.792967880186043E-5,Poland)
(3.736992253895565E-5,Binomial_nomenclature)
(3.710630393353686E-5,Netherlands)
(3.686797566241694E-5,1989)
(3.645056928427473E-5,Race_(United_States_Census)_a07d)
(3.6011843230538025E-5,1980)
(3.578883844342842E-5,Company_(law))
(3.523317689892701E-5,Building)
(3.514726898472093E-5,Band_(music))
(3.481198780373071E-5,1982)
(3.4699835886968835E-5,All_Music_Guide_0e49)
(3.4699333024053607E-5,1986)
(3.450776612733449E-5,1983)
(3.4358616092806375E-5,Native_American_(U.S._Census)_1a7a)
(3.433114186309076E-5,1981)
(3.43214245344432E-5,1985)
(3.43088693167764E-5,1984)
(3.418120126446158E-5,1987)
(3.385891317657007E-5,1988)
(3.352013654704332E-5,Rock_music)
(3.3487630092246675E-5,1979)
(3.2835669402383317E-5,Wikimedia_Commons_7b57)

Simple Dataset:

(9.67012914628196E-4,United_States_09d4)
(8.120183592863061E-4,Wikimedia_Commons_7b57)
(6.805966191912094E-4,England)
(6.212014633154946E-4,Germany)
(4.638296774936001E-4,France)
(4.000427327521526E-4,City)
(3.940630362136738E-4,Inhabitant)
(3.5245301913094663E-4,Wiktionary)

(3.4253789445447985E-4,Country)
(3.380448824706787E-4,Animal)
(3.244719774768859E-4,Japan)
(3.239498889937736E-4,United_Kingdom_5ad7)
(3.228203271689595E-4,Computer)
(2.990373623149513E-4,Water)
(2.9445402257435974E-4,Europe)
(2.939258803868799E-4,India)
(2.7672420296175765E-4,Spain)
(2.7630946916611907E-4,Australia)
(2.727383724861821E-4,English_language)
(2.726331970252819E-4,Italy)
(2.668122880481674E-4,Canada)
(2.6422030081973524E-4,Television)
(2.5590542925236814E-4,Plant)
(2.555415512785434E-4,Earth)
(2.337373165307142E-4,London)
(2.3349638209489644E-4,Money)
(2.318589432614719E-4,China)
(2.2796162135984145E-4,Greece)
(2.2719514698814222E-4,Music)
(2.269963468976905E-4,Scotland)
(2.2443719187362253E-4,Food)
(2.2313221984803828E-4,Football_(soccer))
(2.1733255144934922E-4,Capital_(city))
(2.1503696860873726E-4,Human)
(2.1446691081302696E-4,Metal)
(2.0893835308639415E-4,Capital_city)
(2.0632649566711722E-4,Mathematics)
(2.056478306455597E-4,Movie)
(2.0459519294325812E-4,Netherlands)
(2.0109825408483258E-4,Government)
(1.9803155455921093E-4,Russia)
(1.975837336838836E-4,Brazil)
(1.97506053912134E-4,U.S._state_5a68)
(1.96953330296884E-4,Number)
(1.9679885441564027E-4,Greek_mythology)
(1.9679166142598482E-4,Book)
(1.9668350424921727E-4,People)
(1.9418002520183754E-4,2005)
(1.9189313491960587E-4,Poland)
(1.9102754149207728E-4,2004)
(1.9099445695245005E-4,Language)
(1.8809828427336455E-4,2006)
(1.8361044390684418E-4,Religion)
(1.8273199029667251E-4,Year)
(1.808236091529336E-4,Actor)

(1.7792100746226702E-4,God)
(1.779100360176405E-4,Asia)
(1.765043679320174E-4,California)
(1.7622025740204514E-4,Sweden)
(1.7468808252629742E-4,Science)
(1.73978885362401E-4,University)
(1.7304304870286965E-4,19th_century)
(1.6914957251879664E-4,Fruit)
(1.6547632141132636E-4,Car)
(1.6284381620683224E-4,Chemical_element)
(1.6281671014219728E-4,Africa)
(1.6054039464706126E-4,Disease)
(1.6044206316176765E-4,Film)
(1.5988059855488767E-4,Internet)
(1.5967291246748316E-4,World_War_II_d045)
(1.5914540095981412E-4,Species)
(1.5843088896596663E-4,Latin)
(1.5709499834553307E-4,Company)
(1.543301147057579E-4,River)
(1.537775495176935E-4,North_America_e7c4)
(1.5323743917771722E-4,Fish)
(1.5191520230189065E-4,20th_century)
(1.4976991542618598E-4,Liquid)
(1.4966758881797656E-4,1970s)
(1.4932842218727424E-4,Island)
(1.489425328474606E-4,Centuries)
(1.4881698721304955E-4,Greek_language)
(1.4775220871417552E-4,Internet_Movie_Database_7ea7)
(1.468803443761081E-4,Video_game)
(1.465133109923956E-4,Sport)
(1.4541292151299813E-4,War)
(1.4259567267739182E-4,1960s)
(1.4225108456372465E-4,Mammal)
(1.421931283278361E-4,Christianity)
(1.4181586704043193E-4,German_language)
(1.4176542238614502E-4,Law)
(1.4088526591877522E-4,Prefecture)
(1.4025953863579414E-4,Sun)
(1.3977038608185106E-4,County)
(1.394298484641255E-4,Singer)
(1.383075676550904E-4,State)
(1.380230456192148E-4,Tree)
(1.3799179218243208E-4,Austria)
(1.3738228796944058E-4,Chad)
(1.3670575550066614E-4,Child)