AYUSH SINGH

# CS 6240: Assignment 4
Pagerank in Spark

## Design Discussion

I followed the 3 step design strategy mentioned in the homework as follows:

**Preprocessing:** For starters I handled the case where the html file had char '&' which is invalid so I added '&amp;' as per world wide web consortium. Apart from that, for a sample graph here's how we handle parsing from Map to Reduce in Spark:

1. Load input file
2. Apply map to parse each line of input using WikiParse
3. Output format of Parser for each line *PageName***\t***AdjacencyList*
4. Split each line and convert into tuple of format: *(PageName, (AdjacentNodes, 0))*
5. Now to handle a special edge case, for all dangling nodes that do not have a line will be passed through above step and would not be taken into account. So, in order to handle above edge case I've mapped all values in adjacent list and then union those with original graph list and filter all duplicates which leaves us with valid list that we can start with.
6. Calculate the total number of links and initialize it as broadcast variables so that it is available in distributed mode.
7. Once we have the count, I updated all values of pageranks to 1/totalNumberLinks.
8. I initialized the graph as a variable and persisted it in memory since it going to be updated multiple iterations which would provide us performance boost.
9. Next I filter out all dangling nodes from normal nodes into two separate RDDs

**PageRank:** I used the solution 2 for computing dangling node pagerank share for *i+1* iteration at the beginning of *i*th iteration phase and making it available to all tasks. I grasped the concept of PageRank from pseudo code which I found logical but since that did not handling dangling nodes, I started programming spark solution for it based off that code and took it from there.

1. **Output format** preprocessing phase (*PageName***,** *(AdjacencyList***,** *initPageRank))*

** All elements in a line are guaranteed to be tab separated from parser*

2. **Iterations: 10**
    a. Calculate the dangling loss, keeping only pageranks for all dangling nodes and performing a reduce on them
    b. Update pageranks of all dangling nodes with dangling loss calculated in (a)
    c. Update all normal nodes by dividing old pageranks by total number of adjacent nodes and then calculating them with new pagerank
    d. Finally update the graph by taking a union between dangling nodes from step b and c

**TopK Sort:** Sorting came out pretty easy because of already implemented sorting algorithms in spark using sortBy and taking the top 100 from them. Finally for pretty printing purposes after taking the top 100 into *PageName**\t**PageRank*

## Compare

**Preprocessing:** In MapReduce I had written a custom Map-Reduce job for parsing as a separate job while in spark I had to do it step by step.
**PageRank:** Implementing pagerank in MapReduce was tricky and inelegant compared to that of Spark, since I had to conditionalize at several steps which seemed tricky too in the beginning with spark but became pretty obvious towards the end. In MapReduce, the pipeline had to be put together starting from the output of preprocessing step as input to pagerank and then consecutive pageranks output acting as input for next generation needed to be written on filesystem whereas in spark this was not the case saved a lot of I/O. Calculating dangling node was not possible in map phase since the computation sum of pageranks of all dangling nodes had to be done in reducer because of which we needed a global counter for reflecting the change in next iteration. One downside or more of a tricky part of spark was the part where we had to segregate dangling and adjacent nodes computation since scala is more of an functional language than imperative.
**TopK:** For sorting too, I had to write another MapReduce job the uses TopK design pattern and use of a TreeMap limiting the size of TreeMap to top 100 pageranks to increase efficiency. While in the case of spark, sorting is an in-built feature where we can simply define the key to be used for sorting and then simply take 100 from the sorted list.
The major tradeoff of Spark is that for efficient computation RDD need heap the size of data structure (size of which cannot be computed on paper) which puts the program at risk

of heap allocation issues while MapReduce is disk based which makes it free of these issues. One advantage of Spark being the ability of to move data around for next computation so it can run reduce without waiting for all MapReduce jobs to complete. The amount of data shuffled is one of the major causes apart from heap allocation issues and the amount of issues one faces when trying to segregate data into two parts and combining them which becomes an issue in case of bigger datasets.

## Performance Comparison

| **Framework** | **6 Machines** | **11 Machines** |
|---|---|---|
| **Spark** | 2hr 15 min | 1hr 11 min |
| **MapReduce** | 1hr 4 min | 37 min |

The reason behind spark running slow in my case the amount of computations it takes to handle the edge case (dangling node present in adjacent list but no line in it), the reasoning behind that is on my local machine on looking at the rows the part that I mentioned above took 92% of the time (55 seconds for parsing and making graph but 5 seconds for pagerank computation). I can only imagine the case in case of big dataset.

## Top 100 Pages

Sorted from highest ranking to lowest
The  answers seems reasonable looking at the keywords in the sample dataset while in the mega dataset since the dataset belongs to year 2006 I'm pretty happy with the result in 10 iterations placing 2006 on the top with most pagerank.
The answer don't differ a lot from each other for e.g. in the first two lines I've mentioned the pageranks of MapReduce and spark respectively which do not differ more than ~.005%.

**SIMPLE WIKIPEDIA DATA MAPREDUCE**

United_States_09d4	0.005355777601255395 0.005548825295802055
Country	0.004027824046355395 0.0041218694048619315
Wikimedia_Commons_7b57	0.0036022714985541565
Week	0.003084465720751175
Earth	0.0026752943563853387
Water	0.002546000793679085

| | |
|---|---|
| Europe | 0.002523803647437281 |
| Sunday | 0.002435140370849413 |
| Monday | 0.002384176065598382 |
| Wednesday | 0.0023542561216078825 |
| United_Kingdom_5ad7 | 0.002304059494456286 |
| Friday | 0.0022884877073897055 |
| Saturday | 0.0022622064161520465 |
| Thursday | 0.0022291401408207814 |
| Tuesday | 0.0022154232994574677 |
| Day | 0.002214672571305371 |
| index | 0.0021428657009387743 |
| Asia | 0.002048696137474343 |
| Animal | 0.002019930329063209 |
| France | 0.0018433959710020043 |
| City | 0.0017698940247303772 |
| Money | 0.0017096034585455699 |
| Government | 0.0016776941838347683 |
| Number | 0.0016539807641294132 |
| Energy | 0.0015556498502864366 |
| Sun | 0.0015435842672796775 |
| English_language | 0.0015398752446945734 |
| Plant | 0.0015367815772914688 |
| England | 0.0014954793083519719 |
| India | 0.001474593356160982 |
| Germany | 0.001465669214320054 |
| Italy | 0.0013886001687866954 |
| Wiktionary | 0.0013628464893483602 |
| Wikimedia_Foundation_83d9 | 0.0013460253833042647 |
| Computer | 0.0013324712512722398 |
| People | 0.001305977143147495 |
| Planet | 0.0012922282160995502 |
| Science | 0.0012772149801701903 |
| Canada | 0.0012494104235240542 |

| | |
|---|---|
| Human | 0.0012150250625922526 |
| State | 0.0011455735684420825 |
| China | 0.0011447345133376349 |
| Year | 0.0011414494449842518 |
| Spain | 0.001112032104007919 |
| Wikipedia | 0.0010729174818036337 |
| Japan | 0.001068713408744579 |
| Mathematics | 0.0010629869193989033 |
| Food | 0.001059291804649356 |
| Australia | 0.0010474679016676523 |
| Geography | 0.0010385557642966254 |
| Russia | 0.001035641036663431 |
| Greek_language | 0.00103330695704128 |
| Capital_(city) | 0.001022697843835778 |
| Atom | 9.90285448638182E-4 |
| Society | 9.627595314902565E-4 |
| Liquid | 9.468816031824316E-4 |
| Language | 9.435983218665094E-4 |
| Moon | 9.263060285517775E-4 |
| Africa | 9.220092243932372E-4 |
| Metal | 9.11448578054992E-4 |
| World | 9.024211689218325E-4 |
| Sound | 8.925511641359888E-4 |
| Cyprus | 8.884938768969182E-4 |
| Light | 8.812600713629492E-4 |
| Culture | 8.795620015871022E-4 |
| Greece | 8.762906211813577E-4 |
| History | 8.696008691701632E-4 |
| Law | 8.670235864565729E-4 |
| Turkey | 8.55910398497589E-4 |
| Scientist | 8.523872220858339E-4 |
| Plural | 8.483304168102193E-4 |
| Religion | 8.34546008939885E-4 |

```
Scotland    8.321642740565298E-4
Circle    8.105646859580774E-4
Gas    8.017461004974286E-4
2004    7.950566438796031E-4
Ocean    7.741850211831742E-4
20th_century    7.732364121600164E-4
Poland    7.643205099705985E-4
Solid    7.636188505689122E-4
Information    7.627704616486011E-4
Sweden    7.593342065596659E-4
Television    7.582907444708231E-4
Nation    7.496152107354E-4
War    7.429907486244271E-4
Trade    7.40708869926796E-4
Denmark    7.354747520959662E-4
Building    7.32203871729194E-4
19th_century    7.320051566582628E-4
Continent    7.316903496399993E-4
Portugal    7.29182994494692E-4
Electricity    7.136784745130115E-4
Chemical_element    7.08076888666907E-4
Austria    6.882539890516241E-4
Image    6.815065310045475E-4
Republic_of_Ireland_10e7    6.780880414452558E-4
Music    6.748331871234114E-4
Belgium    6.669926820299989E-4
Time    6.638254426235764E-4
God    6.552185960502633E-4
```

### SIMPLE WIKIPEDIA DATA SPARK

```
United_States_09d4    0.005548825295802055
Wikimedia_Commons_7b57    0.0041218694048619315
England    0.003817371277739225
```

| | |
|---|---|
| Germany | 0.0035332786671783993 |
| France | 0.002525059984194524 |
| Inhabitant | 0.0022125292087488285 |
| City | 0.0019428475458643211 |
| Wiktionary | 0.001778641618089768 |
| Computer | 0.0016454297465091105 |
| Japan | 0.0015939145952123552 |
| Animal | 0.0015873020080940702 |
| United_Kingdom_5ad7 | 0.0015120371716174274 |
| Country | 0.0014822397792028155 |
| India | 0.0014622273465472201 |
| Europe | 0.0014091767849129496 |
| Australia | 0.001350428013761719 |
| Italy | 0.0013487930818108173 |
| Water | 0.0013259229829391584 |
| Canada | 0.0013022397106097928 |
| Television | 0.0012564213873361462 |
| English_language | 0.0012440309959703174 |
| Spain | 0.0012439654924628993 |
| Plant | 0.001229689883995675 |
| Earth | 0.0011519893331854796 |
| London | 0.0011032814951953671 |
| Football_(soccer) | 0.0010776824681638303 |
| Scotland | 0.00107199200206158 |
| Greece | 0.0010707386174618092 |
| China | 0.0010595623348008474 |
| Money | 0.0010485541377480148 |
| Music | 0.0010194743435322999 |
| Metal | 9.815116345407182E-4 |
| Food | 9.785253418010495E-4 |
| Capital_(city) | 9.689441566104012E-4 |
| 2005 | 9.668547497985122E-4 |
| Brazil | 9.604593284520008E-4 |

| | |
|---|---|
| Netherlands | 9.485313681422204E-4 |
| Human | 9.433428004608947E-4 |
| U.S._state_5a68 | 9.417402319782708E-4 |
| 2006 | 9.338870651012233E-4 |
| Greek_mythology | 9.011728377901181E-4 |
| Poland | 8.825157693906941E-4 |
| Russia | 8.812785711597241E-4 |
| Book | 8.801116929197717E-4 |
| Number | 8.677230337955904E-4 |
| Mathematics | 8.657794466703634E-4 |
| 2004 | 8.295154088806876E-4 |
| People | 8.292099796677795E-4 |
| Actor | 8.12638988048379E-4 |
| Language | 8.097369871160237E-4 |
| Asia | 8.066696615846714E-4 |
| Government | 8.035590132424333E-4 |
| California | 8.01246146831695E-4 |
| God | 7.956744204255607E-4 |
| Year | 7.89122190181667E-4 |
| Sweden | 7.806550041557545E-4 |
| Religion | 7.587719094811818E-4 |
| University | 7.536235316625896E-4 |
| Fruit | 7.374176280032844E-4 |
| Africa | 7.244647175264564E-4 |
| Science | 7.190605203975692E-4 |
| Chemical_element | 7.001796932530667E-4 |
| Film | 6.938896115628842E-4 |
| Car | 6.780110390106711E-4 |
| Internet | 6.670272510687821E-4 |
| Disease | 6.667438793560819E-4 |
| Company | 6.586160820945392E-4 |
| World_War_II_d045 | 6.528190240842213E-4 |
| River | 6.444170659173428E-4 |

| | |
|---|---|
| Species | 6.434692378948752E-4 |
| 19th_century | 6.406390306797231E-4 |
| Internet_Movie_Database_7ea7 | 6.351015578478554E-4 |
| Fish | 6.342664444332303E-4 |
| Prefecture | 6.321128777819253E-4 |
| Video_game | 6.316776017577949E-4 |
| North_America_e7c4 | 6.1826447292494E-4 |
| Liquid | 6.120296602864808E-4 |
| Singer | 6.059726469577596E-4 |
| 1970s | 6.047711027779445E-4 |
| Chad | 6.006111136666962E-4 |
| Island | 6.000193785172712E-4 |
| Sport | 5.799847851737136E-4 |
| War | 5.765409509519231E-4 |
| County | 5.720206167768031E-4 |
| 2001 | 5.691310530775546E-4 |
| Tool | 5.687157922399536E-4 |
| 1960s | 5.562591439048566E-4 |
| German_language | 5.526719505718819E-4 |
| Band | 5.504566094530488E-4 |
| Greek_language | 5.492574080821844E-4 |
| Dinosaur | 5.483398421395952E-4 |
| Bird | 5.479774667616909E-4 |
| New_York_City_1428 | 5.463928850621049E-4 |
| Mammal | 5.407918844187377E-4 |
| Tree | 5.403774580958704E-4 |
| Christianity | 5.390665318617126E-4 |
| Austria | 5.331423176387067E-4 |
| Paper | 5.322931379120414E-4 |
| Word | 5.319969109511678E-4 |
| 2003 | 5.272881338406235E-4 |

**BIG WIKIPEDIA DATA SPARK**

```
2006    0.0014920760969739189
United_States_09d4    0.0013299391361721584
United_Kingdom_5ad7    6.909197635236242E-4
2005    6.441475771352479E-4
France    5.094348783496997E-4
2004    4.3436094669958863E-4
Germany    3.997035998660669E-4
England    3.98531940832174E-4
Canada    3.832928601048221E-4
2003    3.477077304703758E-4
Italy    3.2840550275288396E-4
Australia    3.226610267707091E-4
Japan    3.1510698593140103E-4
English_language    2.982189290731356E-4
India    2.811552832867081E-4
World_War_II_d045    2.792335870587863E-4
Europe    2.7640188613208667E-4
Wikimedia_Commons_7b57    2.73873970236108E-4
2002    2.679878469473865E-4
2001    2.604812158757935E-4
Russia    2.5638029261881206E-4
Spain    2.5207580717932315E-4
London    2.5120344963285246E-4
Wiktionary    2.495211489511024E-4
2000    2.464193712173099E-4
index    2.3659248123397047E-4
1999    2.2631894692557494E-4
Geographic_coordinate_system    2.1765131137219774E-4
Race_(United_States_Census)_a07d    2.1420382619338234E-4
```

```
New_York_City_1428    2.0066172646246853E-4
1998    1.9285813940081624E-4
Sexagenary_cycle    1.8957189935115052E-4
1997    1.893422167179812E-4
January_1    1.8889021088400158E-4
Latin    1.8623506229802309E-4
Netherlands    1.8227741182140446E-4
Internet_Movie_Database_7ea7    1.822301891127428E-4
China    1.804642512035373E-4
Scotland    1.7897972157917102E-4
Population_density    1.7728840466749174E-4
1996    1.7597487057506258E-4
French_language    1.7583910426884224E-4
1995    1.685701495101728E-4
Gregorian_calendar    1.684480767094944E-4
1991    1.646282046741122E-4
1994    1.6242519872599763E-4
Soviet_Union_ad1f    1.5949092793559583E-4
Sweden    1.5946745704803497E-4
1990    1.586201866380676E-4
Biography    1.5700541482397144E-4
1993    1.5234095439014728E-4
Egypt    1.5084070486255213E-4
1945    1.5019808221113237E-4
1992    1.4878331699088595E-4
1980    1.4671397913761116E-4
New_Zealand_2311    1.4609833964421455E-4
Greek_language    1.45429366519352E-4
International_Phonetic_Alphabet_96f8    1.4455191773350167E-4
1989    1.4424253851408266E-4
California    1.4384412326758444E-4
1974    1.4349615844128658E-4
1970    1.4344085345985232E-4
```

| | |
|---|---|
| European_Union_e368 | 1.4334171287636775E-4 |
| 1979 | 1.4202366295932683E-4 |
| Square_mile | 1.4088137097768069E-4 |
| 1986 | 1.404619609612325E-4 |
| 1969 | 1.3917077048323565E-4 |
| 1976 | 1.385833488942994E-4 |
| New_York_3da4 | 1.3807215052200365E-4 |
| 1981 | 1.3759483030844076E-4 |
| 1975 | 1.373522585895231E-4 |
| Public_domain | 1.3625766174258822E-4 |
| 19th_century | 1.3585219832868968E-4 |
| 1982 | 1.3507222791698785E-4 |
| 1972 | 1.3486427170254686E-4 |
| Ireland | 1.346846113607906E-4 |
| Greece | 1.3432708987580412E-4 |
| 1985 | 1.3391270939924898E-4 |
| Poland | 1.3359057167464184E-4 |
| Switzerland | 1.3343608993229795E-4 |
| People's_Republic_of_China_82bf | 1.333429603744648E-4 |
| Portugal | 1.3314666664759677E-4 |
| Paris | 1.3305379337486407E-4 |
| German_language | 1.3240932419006092E-4 |
| 1971 | 1.3200428781096066E-4 |
| Austria | 1.3198210050315224E-4 |
| 1973 | 1.315162910676586E-4 |
| Television | 1.3059789980267192E-4 |
| 1984 | 1.2996518561059735E-4 |
| 1983 | 1.295438106675563E-4 |
| Mexico | 1.2951896536378003E-4 |
| 1968 | 1.2949750640369766E-4 |
| World_War_I_9429 | 1.293959133139155E-4 |
| 1977 | 1.2937730398434192E-4 |
| 1967 | 1.2901472998762647E-4 |

```
United_Nations_3208    1.288419414047113E-4
1987    1.2723301598069266E-4
Denmark    1.2711960612723044E-4
Israel    1.2703458560928694E-4
South_Africa_1287    1.2665028789059934E-4
```

**BIG WIKIPEDIA DATA MAPREDUCE**

```
2006    0.0017561678956932896
United_States_09d4    0.0015629256171461271
United_Kingdom_5ad7    8.128548074567297E-4
2005    7.564963469890166E-4
France    5.925301936896899E-4
2004    5.099211551020347E-4
England    4.6864921101578874E-4
Germany    4.673017525794225E-4
Canada    4.5387147485567546E-4
2003    4.0821507598863153E-4
Australia    3.823167537352121E-4
Italy    3.7942937213960887E-4
Japan    3.712414683098072E-4
English_language    3.441478626653587E-4
India    3.306205198168228E-4
World_War_II_d045    3.257223788921249E-4
Europe    3.2413335109059176E-4
Wikimedia_Commons_7b57    3.190315197028015E-4
2002    3.152210480279782E-4
2001    3.0677810669101 24E-4
Russia    2.976467741991308E-4
London    2.939697118177294E-4
Spain    2.9388671993097197E-4
Wiktionary    2.9087192154469244E-4
2000    2.900563417746145E-4
1999    2.6604268563140676E-4
```

| | |
|---|---|
| Geographic_coordinate_system | 2.605359080494864E-4 |
| Race_(United_States_Census)_a07d | 2.516181276495843E-4 |
| New_York_City_1428 | 2.349706752953641E-4 |
| 1998 | 2.2683950071502245E-4 |
| index | 2.2537299431849066E-4 |
| 1997 | 2.2229350413132716E-4 |
| Internet_Movie_Database_7ea7 | 2.1809382069533858E-4 |
| January_1 | 2.1776223196021512E-4 |
| Latin | 2.1405555021091627E-4 |
| Sexagenary_cycle | 2.1395666889813969E-4 |
| Netherlands | 2.1235611032947364E-4 |
| Population_density | 2.1084636247127543E-4 |
| Scotland | 2.098679197724891E-4 |
| China | 2.0940714138364917E-4 |
| 1996 | 2.067576559341469E-4 |
| French_language | 2.0358887219561834E-4 |
| 1995 | 1.9784243755527093E-4 |
| 1991 | 1.9256456201430992E-4 |
| 1994 | 1.9057852019116069E-4 |
| Gregorian_calendar | 1.9018501335305366E-4 |
| Biography | 1.8987844110374437E-4 |
| Sweden | 1.877519843059108E-4 |
| 1990 | 1.857354937946 95E-4 |
| Soviet_Union_ad1f | 1.8442914940787583E-4 |
| 1993 | 1.7870174784252857E-4 |
| 1992 | 1.743009157291231E-4 |
| 1945 | 1.7355067815388685E-4 |
| Egypt | 1.7352763673565312E-4 |
| New_Zealand_2311 | 1.7283377472129526E-4 |
| 1980 | 1.7093646700823602E-4 |
| California | 1.6978941783386627E-4 |
| 1989 | 1.686558190404602E-4 |
| 1974 | 1.669965256442851E-4 |

```
Greek_language    1.669815145504117E-4
Square_mile    1.6696225923309852E-4
1970    1.6661433343946735E-4
European_Union_e368    1.663306992990674E-4
1979    1.6525851782910593E-4
1986    1.640052552408662E-4
International_Phonetic_Alphabet_96f8    1.6234059557987404E-4
New_York_3da4    1.6218406912468114E-4
1969    1.6165358996418416E-4
1976    1.6109815081919275E-4
1981    1.6020688572732572E-4
1975    1.5964401164388577E-4
Ireland    1.5804692014397817E-4
1982    1.575246188062329E-4
19th_century    1.573149434213807E-4
Public_domain    1.569875990428344E-4
Switzerland    1.5688485519409865E-4
1972    1.5680244312973025E-4
1985    1.5649492178643963E-4
Poland    1.5635162879973753E-4
Television    1.56043197092036E-4
Greece    1.5599289619959325E-4
Portugal    1.554131856898917E-4
People's_Republic_of_China_82bf    1.546906842926289E-4
Paris    1.5466448380756806E-4
Austria    1.5370765775585613E-4
1971    1.5323070188567075E-4
German_language    1.5320009068820422E-4
1973    1.5269889882460117E-4
Mexico    1.524303669140567E-4
1984    1.5184881565541674E-4
1983    1.513047618813404E-4
1977    1.5059998664674107E-4
```

```
1968    1.5049201699578754E-4
World_War_I_9429    1.5045643188356077E-4
1967    1.4990820328140901E-4
United_Nations_3208    1.4954232897493303E-4
1987    1.4901742052346803E-4
Denmark    1.4872554335581447E-4
South_Africa_1287    1.4799637778879806E-4
Israel    1.46878199854074E-4
```