Report HW4 – PageRank in Spark

CS6240 Map Reduce – Akash Kadam (Section 2)

Design Discussion

- **RDD: Graph Generation (pageName, outlink List)**
  1) Read the bz2 compressed file
  2) Applying a map so that each line from input file is parsed by the WikiParser which generates the String in the format of "PageName->name~name~name"
  3) Filter is applied to remove all the links which are "invalid" returned by the wikiParser
  4) Apply map which splits the line based on delimiter "->"
  5) Apply map to create a tuple containing (pageName, outlinksList)
  6) Apply keyBy to create K,V pair with pageName as key
  7) Apply mapValues to drop the key component which is present in the value part of the K,V pair

  I persist the input graph, which keeps the RDD in memory to reduce the time to access the RDD

- **RDD: List Generation for all pagesNames in input Graph (pageName, Empty List)**
  1) Get only the values from the above generated input graph
  2) Apply flatmap to flatten all the list of outlinks in the list
  3) Create a K,V pair for all elements in the list
  4) Reduce it by Key
  5) Apply map to generate a line of the form (pageName, empty list)

- **RDD: Final List (containing full graph list with dangling nodes)**
  1) Take a union of the above generated RDD's
  2) Reduce It by Key merging all the subsets for a key match

- **RDD: PageRankList (contains pageName with their initial Page Rank values)**
  1) Apply MapValues on above final list assigning the initial page Rank to all pages

- **RDD: PageRankList computation for 10 iteration**
  1) Compute a page rank list which where we distribute the contribution of a page to all its outlinks if it contains any, or else if it is a dangling node, the dangling factor accumulator is incremented by the page rank value of the dangling node
  2) The PageRankList RDD is updated by computing the new PageRank using the formula

- **Top 100 pages based on Page Rank**
  1) Sort the PageRankList RDD based on the page rank values and take the 100 values

Comparison of the Hadoop MapReduce and Spark implementations of PageRank

1) **Hadoop Job1 vs Spark implementation**
   In the Job1 of Hadoop, which was only a map job generates the input graph, reading the bz2 file. The same thing is achieved in spark with the first three points discussed in design discussion i.e. Graph Generation, List Generation and Final List

2) **Hadoop Job2 vs Spark implementation**
   The map reduce in Job2 computes the page rank for all the pages and it also uses special global counters to maintain the dangling factor count and page Count.
   In Spark, this is same as explained in design discussion point 4 & 5 i.e. PageRankList and PageRankList computation for 10 iterations

3) **Hadoop Job3 vs Spark implementation**
   The map reduce in Job3 gets the100 pages with highest page rank values which uses a HashMap data structure to store all the K,V pair and then sorts the hash map based on values in descending order and outputs the first 100 from the list.
   The same implementation in spark is achieved by sorting the RDD and taking 100 records from the RDD

Advantages and Short Comings of different approaches

1) Spark has flexible APIs and it's easy to write user-defined functions. It also includes spark shell for running commands to get immediate results. Hadoop MapReduce is more difficult to program as compared to Spark
2) Spark's performance is improved if we can cache the data in memory. Hadoop MapReduce is designed for large data sets which cannot be fitted into memory for faster execution.
3) The simple Map Reduce Job written in Hadoop might contain approx. 150 lines of code which the same program written in Scala Spark can will have approx. 20 lines of code.
4) PageRank implementation was very straightforward in Spark Scala. No need to configure and write separate Map Reduce Jobs as in Hadoop one to process the graph structure, second to compute the actual page rank, third to get the top 100 pages by page rank values. No need of using additional data structures, custom objects etc. to implement page rank in spark as compared to Map Reduce

Performance Comparison

Run time comparison of AWS runs for Spark Configuration

M4.large count	Time(min)
6	53
11	139

Run time comparison of AWS runs for Hadoop Configuration

M4.large count	Job1 - Preprocessing (min)	Job2 – 10 Iterations (min)	Job3 - Top-100 (min)	Total Time (min)
6	18	29	3	50
11	10	48	2	60

According to my experiments and observation from above 2 tables the Hadoop configuration time is less by 5 min when run on 6 machines as compared to Spark. When run on 11 machines the time taken is more than double in Spark configuration as compared to Hadoop configuration. The is totally opposite from my expectations from spark architecture. In theory, I expected the run time for executing in spark to be considerably less than the Hadoop Run time. The reason could be my code in spark could have been optimized when getting the top 100-page rank values. Instead of sorting the entire RDD I could have parallelized the process by dividing into splits, sort the splits and get only top 100 pages from the splits. Finally, sort the resultant list and fetch the top 100. It might also be the case that for huge set of data Spark might be more efficient as compared to Hadoop implementation.

According to my understanding, the other important factor in degradation of Spark performance is that it is running on Hadoop Yarn, which is highly resource demanding and the data is too big to fit entirely into the memory for the preprocessing step.

Top 100 result from Spark and Hadoop comparison for local and AWS execution

1) **Output for Hadoop implementation**

Local	6 Machines AWS	11 Machines AWS
0.003569625792468806 United_States_09d4 0.0028382076913557428 Wikimedia_Commons_7b57 0.0023365970691297156 Country 0.0015716213589640395 England	4.8354501316565985E-4 United_Kingdom_5ad7 2.274417957849596E-4 2003 1.6722993239519924E-4 World_War_II_d045 1.5542921882227682E-4 1999	1.6723157697931906E-4 World_War_II_d045 1.2796589321863747E-4 1997 8.447999905636918E-5 Paris 8.288502957005226E-5 1974 6.514657778075556E-5 1963 5.540033726560612E-5 1939 5.4537280450422105E-5 Rome

0.0015609374356412862 Europe 0.0015471260915105342 United_Kingdom_5ad7 0.001540060961541869 Water 0.0015336737129266014 Germany 0.0014985136908930862 France 0.0014697291097649932 Earth 0.0014643519517427734 Animal 0.0013515080807248007 City 0.0012382042222981065 Week 0.001160112838520621 Asia 0.0011360986407891504 Sunday 0.001118206599957752 Monday 0.001107698693003402Wiktionary 0.001107594952344537Wednesday 0.0010926093943232752 Money 0.0010820580032024998 Plant 0.0010807920687108852 Friday 0.0010684361117729536 Saturday 0.001054507091664898 Thursday 0.00104697167014417 Tuesday 0.0010440057437382411Computer 0.0010434909654742392 English_language 0.0010247375976506642 Italy	1.3566296917073874E-4 Wikimedia_Commons_7b57 1.346010848977608E-4 1998 1.0005886660037758E-4 Film 9.373540028379748E-5 Ireland 8.88196228098041E-5 1980 8.751570853754391E-5 Album 8.469816463636737E-5 1979 8.447294911063003E-5 Paris 8.288437603171155E-5 1974 7.821748947892611E-5 1945 7.79856292548301E-5 1969 7.718708863327996E-5 1978 6.721194773264558E-5 Asia 6.303900865743857E-5 People's_Republic_of_China_82bf 6.0405841308137165E-5 1947 5.467590318351779E-5 1953 5.079251348039959E-5 Rock_music 4.881748344313319E-5 Arabic_language 4.816731851330361E-5 Sunday 4.808819441859829E-5 Binomial_nomenclature 4.791884008072315E-5 Republic_of_Ireland_10e7 4.780341500369632E-5 1938 4.713603184750893E-5 School	5.1795066649871706E-5 1920 4.928179355440393E-5 1936 4.8648584607164496E-5 Plant 4.713604327065189E-5 School 4.38130062945977E-5 Building 3.678782922292051E-5 Sydney 3.539859386973007E-5 Italian_language 3.492777071780135E-5 Atlantic_Ocean_570a 3.3005941055370206E-5 September_12 3.29463798614201E-5 List_of_countries 3.272881617289423E-5 Technology 3.223233142936291E-5 Middle_East_c1c7 3.2059780221594884E-5 September_22 3.2038059464636036E-5 November_9 3.18322767612703E-5 April_29 3.182003682031813E-5 October_7 3.165844575104991E-5 July_25 3.154415112010381E-5 History 3.1388334322521576E-5 Chordate 3.125956661627916E-5 August_25 3.0733092441104195E-5 January_21 3.066029581528463E-5 Judaism 3.0625335707417905E-5 May_2 3.0175036163326557E-5 June_18 3.0163782817771704E-5 Bird

0.0010120953137671767
India
9.926473551099425E-4
Government
9.66146160566674E-4
Number
9.203367818438623E-4
Spain
9.063302046275335E-4 Day
8.955447110490885E-4
Japan
8.614731954669139E-4
People
8.528386747071949E-4
Canada
8.505841050108146E-4
Human
8.40387400671836E-4 index
8.237933035789281E-4
Wikimedia_Foundation_83d9
8.121810181404896E-4
China
8.116984201313178E-4
Energy
7.942337257722643E-4
Australia
7.882278768942304E-4 Sun
7.84242798389517E-4 Food
7.763994446010676E-4
Science
7.662976122720625E-4
Mathematics
7.196756733141486E-4
Television
7.058744091842173E-4
Russia
6.822048992927936E-4 Year
6.774453887348216E-4 State
6.761575949567515E-4
Music
6.643191921890625E-4
Greece
6.642143824992926E-4
Capital_(city)
6.641367711651933E-4
Language

4.553978250329162E-5
Guitar
4.5426553463947566E-5
1935
4.3932204320676405E-5
March_4
4.363721154049141E-5
1934
4.199987081250976E-5
Pennsylvania
3.94766949485392E-5 1910
3.9314226881802174E-5
Hindu_calendar
3.883077125536793E-5
Pop_music
3.8819698803327604E-5
Colombia
3.7356061907092204E-5
Politics
3.645756213132425E-5
Iceland
3.6408170738210795E-5
South_Korea_aa29
3.387369071772775E-5
Lebanon
3.3455962436719246E-5
Los_Angeles
3.3380401390188986E-5
September_15
3.3250237980243064E-5
October_10
3.320051047114196E-5
January_20
3.289191851491922E-5
April_20
3.280656855885125E-5
November_17
3.25988527836435E-5
March_21
3.25883690376793E-5
December_3
3.2578821937048505E-5
Pacific_Ocean_bdb7
3.250142245083809E-5
February_14

2.983036447091657E-5
August_26
2.9774891347448514E-5
June_7
2.9554444732312736E-5
January_2
2.9396060958881637E-5
January_19
2.9356322010925687E-
Slovenia
2.9018930147838315E-5
Alaska
2.886950580396677E-5
July_30
2.8090530327512994E-5
February_21
2.7691597021983032E-5
December_21
2.6731760310998217E-5 1894
2.6720525738026866E-5
Vienna
2.629074810820181E-5
Computer_science
2.566630665107635E-5 Artist
2.3937415340229245E-5
Theatre
2.3359017735582752E-5
Commonwealth_of_Nations_38fc
2.304085473954951E-5 Jesus
2.278328362004196E-5
Azerbaijan
2.2210423624383035E-5
Architecture
2.213733292833538E-5
Greek_mythology
2.205633718689317E-5
Cornwall
2.1522063943507222E-5
Jordan
2.0858239299995685E-5
Ancient_Greece_e4c9
2.0646794612325702E-5
Edinburgh
2.0529636117609155E-5 1875

6.558004294140977E-4 Scotland 6.485312855226407E-4 Metal 6.427576455595093E-4 Wikipedia 6.368723682070818E-4 Greek_language 6.343081507947149E-4 Planet 6.295749251339534E-4 2004 6.133404238667168E-4 Sound 6.105259179352422E-4 Religion 6.066352013465335E-4 London 6.054776986421647E-4 Africa 5.729885095381664E-4 Poland 5.698185227142411E-4 Geography 5.661639236914086E-4 Liquid 5.638531282780253E-4 20th_century 5.628737793687771E-4 Law 5.548665936502902E-4 World 5.501996847195113E-4 19th_century 5.480131311912433E-4 Scientist 5.468927906682265E-4 Society 5.362820007999706E-4 Atom 5.263711162093771E-4 History 5.247285586337481E-4 Latin 5.225613693067157E-4 Light 5.220497096199476E-4 Sweden 5.143097312344817E-4 War	3.2374375306048845E-5 November_29 3.230409382126497E-5 February_28 3.221593026667028E-5 Venezuela 3.205968552198247E-5 September_22 3.2057850690412036E-5 March_20 3.191583929857617E-5 May_10 3.140673312515485E-5 March_13 3.134134668385888E-5 October_29 3.126998614663054E-5 Taiwan 3.125964479523161E-5 August_25 3.1216435538976406E-5 August_9 3.1107774568294506E-5 Virginia 3.103868875311036E-5 Syria 3.095348641016176E-5 January_15 3.065819626917218E-5 Major_League_Baseball_9421 3.0627933856840005E-5 March_17 3.0565135722355754E-5 Chemistry 3.052882524428556E-5 April_10 3.0487846799259952E-5 Luxembourg 3.0103992345995835E-5 February_23 3.0090376420183674E-5 March_22 3.0060203163544087E-5 February_10	2.0487662712784347E-5 DC_Comics_9a75 2.0221764424659004E-5 Protein 1.989686156188272E-5 Victoria_(Australia)_71e6 1.9891008329349817E-5 Irish_language 1.8859477115905436E-5 Longitude 1.880080368384751E-5 Rock_and_roll 1.8329239201719545E-5 Software 1.720271404887847E-5 Papua_New_Guinea_eb0d 1.708049506085575E-5 Drums 1.7041294787845608E-5 Andorra 1.695768162293038E-5 Geometry 1.6331462621444438E-5 Country 1.6218489262932302E-5 Parliament 1.5669488799074233E-5 South_Carolina_aeeb 1.5643350429704713E-5 New_Mexico_2a18 1.515286436565101E-5 San_Francisco 1.514191581176134E-5 Norfolk 1.4939714010745752E-5 Boat 1.4299924796394643E-5 Carbon 1.421585924053703E-5 Saint 1.3897951687807932E-5 Alexander_the_Great_5bb4 1.3681160447508281E-5 Violin 1.3251443941078908E-5 Danish_language 1.3096962751583653E-5 1821 1.3096773903360864E-5 1830

5.107358111666356E-4 Netherlands 5.10282463169602E-4 Culture 4.963818702341552E-4 Turkey 4.951977894040488E-4 God 4.941945411094556E-4 Building 4.908936706726269E-4 Plural 4.864410244922377E-4 Information 4.785643796209303E-4 Chemical_element 4.7532363848730674E-4 Portugal 4.7359746933450775E-4 Centuries 4.723824264938957E-4 Inhabitant 4.6663908272972913E-4 Denmark 4.630235544365656E-4 Austria 4.612487187353598E-4 Cyprus 4.5552459843812253E-4 Ocean 4.5118553026021876E-4 Moon 4.4943723048416784E-4 Species 4.484523107976154E-4 Disease 4.476856771120241E-4 Book 4.475359153453163E-4 Biology 4.4621220880561E-4 University 4.44387415007513E-4 Capital_city	2.9998582920213797E-5 April_11 2.9991057855400983E-5 May_29 2.99479535994E-5 April_19 2.99272760786E-5 June_29 2.9779416073E-5 April_13 2.9772671141660597E-5 September_20 2.9656296450831902E-5 October_22 2.962699035325277E-5 September_4 2.959704818542207E-5 September_27 2.94948357728E-5 June_24 2.94353997654E-5 April_18 2.937858827822184E-5 Saudi_Arabia_53d2 2.9376634745414983E-5 June_9 2.918074502324E-5 July_3 2.8090356319114033E-5 February_21 2.80817817765563E-5 December_18 2.796238407533053E-5 Film_director 2.7917664630282268E-5 Operating_system 2.777895815E-5 Michigan 2.7690575E-5 December_21 2.720797559326E-5 Cyprus 2.59895350504E-5 Latvia 2.5959282075E-5 Microsoft 2.5797263435588943E-5 Cold_War_f6d9 2.535588124E-5 Queensland 2.5341755778416955E-5 World_Conservation_Union_6f54 2.5305055020435483E-5 1887 2.5130923390627448E-5 Jew	1.3067252876198543E-5 Socialism 1.290675148132386E-5 County_Durham_194a 1.28051593900536E-5 Video_game_publisher 1.2671542148268114E-5 Suriname 1.2287996877363151E-5 Prussia 1.2272122685351372E-5 Provinces_of_Italy_be7d 1.2171546151468484E-5 New_York_Yankees_cfbf 1.205375307833746E-5 List_of_unmanned_aerial_vehicles 1.2002575458192506E-5 National_Rail_49a3 1.1996289880329649E-5 Military_of_the_United_States_8982 1.1916977000180672E-5 Latin_alphabet 1.1807885227639601E-5 1799 1.1734741117878795E-5 Batting_average 1.122839762820156E-5 1825 1.1072470289633193E-5 Executive_(government) 1.1051409565678962E-5 2008 1.1000924659651159E-5 9th_century 1.099069986187371E-5 University_of_Michigan_a8fd 1.039365584697802E-5 Timeline_of_aviation

2) Output for Spark implementation

Local	6 Machines AWS	11 Machines AWS
(United_States_09d4,0.004863778015124289) (Wikimedia_Commons_7b57, 0.0037398725686378725) (Country,0.0030895255517587043) (Europe,0.002070419413715612) (United_Kingdom_5ad7, 0.0020408076719647342) (Water,0.0020350549655741676) (England,0.002024481374775782) (France,0.001973515285170049) (Earth,0.0019324868468396665) (Germany,0.0019322803455490934) (Animal,0.001913449246690474) (City,0.0018448619415539747) (Week,0.0016937372358817942) (Sunday,0.001560370239989838) (Monday,0.0015369838869736649) (Asia,0.0015328129517206627) (Wednesday,0.0015218236091938105)	(United_States_09d4,0.0018137332545390373) (2006,0.0016451462081480154) (United_Kingdom_5ad7,8.645398804234425E-4) (2005,7.514286531935765E-4) (England,5.629101100571953E-4) (Canada,5.576282513123333E-4) (Biography,5.312604790263531E-4) (France,5.22826266795084E-4) (2004,5.177772629957998E-4) (Australia,4.6035348903718243E-4) (Germany,4.5842210107615065E-4) (Geographic_coordinate_system,4.2412413636707557E-4) (2003,4.19314438100985E-4) (Japan,3.991451420985376E-4) (India,3.931665902666479E-4) (Italy,3.372951670288572E-4) (Internet_Movie_Database_7ea7,3.320064166152448E-4) (2001,3.2961221445569394E-4) (2002,3.294682470246997E-4) (2000,3.1043899953104247E-4) (Europe,3.083878020710851E-4) (World_War_II_d045,3.037524452892909E-4)	(United_States_09d4,0.0018137332545390384) (2006,0.0016451462081480165) (United_Kingdom_5ad7,8.645398804234424E-4) (2005,7.514286531935763E-4) (England,5.629101100571953E-4) (Canada,5.576282513123334E-4) (Biography,5.31260479026353E-4) (France,5.228262667950835E-4) (2004,5.177772629958E-4) (Australia,4.603534890371829E-4) (Germany,4.584221010761505E-4) (Geographic_coordinate_system,4.2412413636707584E-4) (2003,4.193144381009853E-4) (Japan,3.9914514209853764E-4) (India,3.931665902666479E-4) (Italy,3.3729516702885726E-4) (Internet_Movie_Database_7ea7,3.320064166152448E-4) (2001,3.2961221445569394E-4) (2002,3.294682470246997E-4) (2000,3.1043899953104236E-4) (Europe,3.083878020710854E-4)

(Friday,0.0014841639888779083) (Saturday,0.00146758600460587) (Thursday,0.0014484580724867268) (Tuesday,0.001438133475596252) (Money,0.001437475426771646) (Wiktionary,0.0014050804826912667) (Plant,0.001381209753481258) (Government,0.0013485522169754574) (English_language, 0.001343337340269417) (Italy,0.001340452954634702) (Computer,0.0013323489546598567) (India,0.0013206161664924669) (Number,0.001279236124797564) (Day,0.001235585686002158) (Spain,0.0012140620725472685) (Canada,0.001158768701818937) (People,0.0011320975658140337) (Japan,0.0011257699964812744) (Human,0.0011120792950610452) (Wikimedia_Foundation_83d9, 0.0010891383936349106) (Australia,0.0010689464578663218)	(London,2.96201089391174E-4) (Population_density,2.797895121537483E-4) (Record_label,2.788642665267945E-4) (English_language,2.783481909878285E-4) (1999,2.752262167043083E-4) (Race_(United_States_Census)_a07d,2.6624655707639867E-4) (Russia,2.596234177988341E-4) (Spain,2.553675674495535E-4) (Wiktionary,2.4563082484133265E-4) (Wikimedia_Commons_7b57,2.448878031422648E-4) (1998,2.396925498050341E-4) (Music_genre,2.330623649024811E-4) (1997,2.3047027064267766E-4) (New_York_City_1428,2.2932628941638327E-4) (Scotland,2.2683973605127584E-4) (1996,2.1653571400972714E-4) (Television,2.0983425313684086E-4) (Square_mile,2.070246400829858E-4) (Census,2.0565806720834265E-4) (1995,2.0415495266860878E-4) (California,2.015274808538534E-4) (China,1.9793250620294198E-4) (Netherlands,1.9607119070399364E-4)	(World_War_II_d045,3.03752445289291E-4) (London,2.96201089391174E-4) (Population_density,2.797895121537482E-4) (Record_label,2.7886426652679454E-4) (English_language,2.7834819098782846E-4) (1999,2.7522621670430825E-4) (Race_(United_States_Census)_a07d,2.662465570763988E-4) (Russia,2.596234177988341E-4) (Spain,2.553675674495532E-4) (Wiktionary,2.4563082484133276E-4) (Wikimedia_Commons_7b57,2.4488780314226493E-4) (1998,2.3969254980503416E-4) (Music_genre,2.330623649024811E-4) (1997,2.304702706426777E-4) (New_York_City_1428,2.293262894163834E-4) (Scotland,2.268397360512759E-4) (1996,2.1653571400972712E-4) (Television,2.0983425313684089E-4) (Square_mile,2.070246400829859E-4) (Census,2.0565806720834273E-4) (1995,2.0415495266860883E-4) (California,2.015274808538534E-4) (China,1.979325062029421E-4)

(index,0.0010673813267546002) (Energy,0.0010644906663659502) (China,0.001055962796909464) (Sun,0.0010436569099234873) (Science,0.001020628436251588) (Food,0.0010141163644755377) (Mathematics,9.88421272726507E-4) (Capital_(city),9.510608342634838E-4) (Russia,9.344084969497696E-4) (Year,9.275771634721162E-4) (Television,9.16023330447108E-4) (State,9.078010573415974E-4) (Music,8.86806852567735E-4) (Language,8.717825468331334E-4) (Wikipedia,8.465708678633594E-4) (Metal,8.463280377435374E-4) (Greek_language,8.369789424596006E-4) (2004,8.351683583288543E-4) (Planet,8.302194280201893E-4) (Religion,8.151676717281831E-4) (Sound,8.06290174408001E-4) (Scotland,7.99514121524288E-4)	(New_Zealand_2311,1.9517424682156395E-4) (1994,1.9484895491921755E-4) (Football_(soccer),1.9411920238158995E-4) (Sweden,1.9019295356028794E-4) (1991,1.8574742766450317E-4) (1993,1.835847855638909E-4) (New_York_3da4,1.8176904141244016E-4) (1990,1.8147351452335795E-4) (United_States_Census_Bureau_2c85,1.7745463970557474E-4) (1992,1.7615365996106344E-4) (Public_domain,1.7323649982188416E-4) (Film,1.72951640643255E-4) (Scientific_classification,1.7198224957859384E-4) (Actor,1.697179002463602E-4) (Ireland,1.6741256488728983E-4) (1989,1.652917879738159E-4) (Population,1.6305615237851504E-4) (January_1,1.623981824049497E-4) (1980,1.615878542518649E-4) (Marriage,1.6121469041599534E-4) (Latin,1.6095252458897026E-4) (1986,1.5732094923300437E-4) (1979,1.5339692400709816E-4) (1985,1.5324567288705433E-4) (1982,1.52751921764527E-4)	(Netherlands,1.9607119070399359E-4) (New_Zealand_2311,1.9517424682156392E-4) (1994,1.9484895491921755E-4) (Football_(soccer),1.9411920238158992E-4) (Sweden,1.9019295356028788E-4) (1991,1.8574742766450323E-4) (1993,1.8358478556389083E-4) (New_York_3da4,1.817690414124401E-4) (1990,1.8147351452335803E-4) (United_States_Census_Bureau_2c85,1.7745463970557474E-4) (1992,1.7615365996106344E-4) (Public_domain,1.7323649982188425E-4) (Film,1.7295164064325512E-4) (Scientific_classification,1.719822495785938E-4) (Actor,1.6971790024636026E-4) (Ireland,1.6741256488729E-4) (1989,1.6529178797381594E-4) (Population,1.6305615237851507E-4) (January_1,1.6239818240494975E-4) (1980,1.6158785425186506E-4) (Marriage,1.6121469041599534E-4) (Latin,1.6095252458897029E-4) (1986,1.573209492330044E-4)

(Africa,7.936827605803759E-4) (London,7.902629522044468E-4) (Greece,7.820622439074139E-4) (20th_century,7.69575343016115E-4) (Geography,7.530899963585173E-4) (19th_century,7.49802787897151E-4) (Law,7.462624154134248E-4) (Liquid,7.34346448605637E-4) (World,7.322900881767974E-4) (Society,7.224702294737911E-4) (Poland,7.181348671117164E-4) (Scientist,7.145500885633549E-4) (Atom,7.046425433574212E-4) (History,6.889546307349842E-4) (Latin,6.869242816483627E-4) (War,6.843867625219117E-4) (Light,6.810439555861572E-4) (Culture,6.737966779119843E-4) (Building,6.647206844762677E-4) (Netherlands,6.584936209977671E-4) (God,6.582052374048738E-4) (Turkey,6.560687031332954E-4) (Centuries,6.495753486753825E-4)	(1981,1.526645972761913E-4) (Per_capita_income,1.5198856643792422E-4) (1974,1.5186313242678375E-4) (Norway,1.507232210923623E-4) (French_language,1.505244211142954E-4) (1984,1.4983104000734461E-4) (1987,1.497258482499401E-4) (1983,1.4952712937297462E-4) (South_Africa_1287,1.4887665696552396E-4) (1970,1.4788847780420933E-4) (Mexico,1.476878834977421E-4) (Record_producer,1.4704127846154995E-4) (Album,1.4670630603453582E-4) (1988,1.460682567620388E-4) (1976,1.458627798299412E-4) (Poland,1.4536473773044806E-4) (Switzerland,1.4437074415529823E-4) (1975,1.4422416374454292E-4) (Km²,1.4358328777829463E-4) (1969,1.4299892748262266E-4) (1972,1.4170740106536216E-4) (1945,1.4144218790212518E-4) (Soviet_Union_ad1f,1.4036257209028025E-4) (1977,1.4021495502361352E-4) (Politician,1.4004082979921122E-4)	(1979,1.5339692400709814E-4) (1985,1.5324567288705417E-4) (1982,1.52751921764527E-4) (1981,1.5266459727619126E-4) (Per_capita_income,1.5198856643792422E-4) (1974,1.518631324267837E-4) (Norway,1.5072322109236234E-4) (French_language,1.505244211142954E-4) (1984,1.4983104000734472E-4) (1987,1.4972584824994014E-4) (1983,1.4952712937297473E-4) (South_Africa_1287,1.48876656965524E-4) (1970,1.478884778042092E-4) (Mexico,1.4768788349774222E-4) (Record_producer,1.4704127846154992E-4) (Album,1.4670630603453588E-4) (1988,1.460682567620388E-4) (1976,1.4586277982994118E-4) (Poland,1.453647377304479E-4) (Switzerland,1.4437074415529826E-4) (1975,1.4422416374454292E-4) (Km²,1.4358328777829457E-4) (1969,1.429989274826229E-4) (1972,1.4170740106536216E-4) (1945,1.4144218790212526E-4)

(Plural,6.495654392414617E-4) (Sweden,6.445400874890911E-4) (Information,6.440407200481515E-4) (Chemical_element,6.404856288564219E-4) (Portugal,6.292390370067036E-4) (Denmark,6.139789767762575E-4) (Capital_city,6.12237323930583E-4) (Austria,6.095087330504669E-4) (Cyprus,6.06902788102297E-4) (Ocean,5.991994168519884E-4) (North_America_e7c4,5.967125139316956E-4) (Disease,5.924699277488833E-4) (Moon,5.916739246871882E-4) (Species,5.870775181428077E-4) (Biology,5.80228801217618E-4) (List_of_decades,5.777725747408758E-4) (University,5.756053815495595E-4)	(Greece,1.3952628234711764E-4) (1978,1.391321889735811E-4) (Brazil,1.38853852777451E-4) (Poverty_line,1.3867610871676702E-4) (1973,1.375814888412942E-4)	(Soviet_Union_ad1f,1.4036257209028034E-4) (1977,1.4021495502361344E-4) (Politician,1.400408297992112E-4) (Greece,1.395262823471177E-4) (1978,1.3913218897358093E-4) (Brazil,1.3885385277745085E-4) (Poverty_line,1.3867610871676713E-4) (1973,1.375814888412941E-4)

Output comparison of Hadoop and Spark

According to the results above there is a significant difference between the two configurations. This is because I had some bugs in my Hadoop implementation because of which I was getting slightly different page rank values. There was a preprocessing bug in my Hadoop implementation where I was using ‘,’ as my delimiter for outlink list. This would have affected my implementation for page rank in Hadoop implementation as there are pageNames having ‘,’ in them which led to resultant records to split incorrectly.

The output of the 6 and 11 machine configuration is nearly same. The page rank values of few pages are slightly different. I suspect this because the processing happens in parallel and the data split might not be the same in both the execution thus affecting the page rank computation in very small precision.