Report

CS6240

HW 4

Gautam Vashisht

Design Discussion

PageRank.scala class will read the from .bz2 format file and map it line by line to XMLJavaParser.java class to get the pageName and its links as in RDD of String. Then using flatMap on RDD, it will emit the (String, String) RDD in form of (pageName, LinkofOutlinks separted by "|").

Using mapValues function, each record will be initialized with initial pageRank and then will be joined with previous RDD to get complete pair of (pageName, pageRank|listofEdges) in fullPair variable. Then, rankShare will be calculated using flatMap by emitting each edge and its rankShare and same for Dangling nodes in the format of (String, Double). Then finalRank share will be computed using pageRank formula and at last danglingShare will be added to finalRank.

Joining finalRank(String, Double) and keyValPair(String, String) will provide us the fullPair for a single iteration. Program will iterate through it for 10 times. After that, we will get the RDD of (rank, node) and will sort it by key in descending order and take first 100 records. The top 100 webpages with their pageRank will be stored in output file.

Comparison of Hadoop MapReduce and Spark implementation

First of all, in Hadoop I was using mapReduce program to parse the XML file but Spark program is passing record by record to XMLJavaParser.java and storing the output in RDD of string. Previous program was handling dangling nodes in parser only, while In spark, adjacency list is customized using flatMap and map functions to add dangling nodes in adjacency list returned by parser.

Mappper is emitting the same key, value pair in both the implementation. Spark program is using flatMap to emit (edgeNode, rankShare) and ("dang", rankShare) for nodes in edges and dangling node respectively.

In Reducer phase also, basic logic is same. It is just that In spark, we are reducing by reduceByKey function and computing pageRank using map function in scala.

One difference is that, In spark implementation program is adding danglingShare at the end of map and reduce, so we don't need to take care of adding danglingShare at the end of last iteration.

Spark using scala is less verbose as compared to code in MapReduce. As spark is using RDD i.e, in memory execution it will make it faster. As programmer can customized the memory storage using persist, so it provides more flexibility. Spark uses data in memory storage, so it might cause memory issue, if its functionalities are not customized properly while MapReduce kills the data in memory as process is executed without causing memory issue.

Performance Comparison

Spark Run on 6 m4.large machines

Spark program is completed in 1 hour and 44 minutes for 6 m4.large machines

Spark Run on 11 m4.large machines

Spark program is completed in 52 minutes for 11 m4.large machines

Hadoop Run on 6 m4.large machines

MapReduce program was completed in 1 hour and 16 minutes for 6 m4.large machines

Hadoop Run on 11 m4.large machines

MapReduce program was completed in 37 minutes for 11 m4.large machines

Surprisingly, Hadoop Map Reduce program was completed in lesser time than compared to Spark run. It might be due to the data shuffling involved in spark version. As in scala, program is emitting and merging lot of map operations which might have caused this issue. Though, I will look more into this that how to resolve the shuffling or other related issue.

Top 100 webpages for Spark (Full Data Set) – PageRank, WebPage

(1.1983583994530579E-5,United_States_09d4)

(7.192138213073082E-6,Biography)

(5.992673694925062E-6,2006)

(4.651999407823781E-6,United_Kingdom_5ad7)

(3.8531990301281695E-6,Canada)

(3.841271744926034E-6,England)

(3.817341691616178E-6,Geographic_coordinate_system)

(3.742561622442789E-6,2005)

(2.997951792931907E-6,Australia)

(2.956454665652581E-6,France)

(2.944801280333169E-6,India)

(2.835249960701435E-6,2004)

(2.740643892907408E-6,Germany)

(2.4550488439877603E-6,Record_label)

(2.407380769348794E-6,Internet_Movie_Database_7ea7)

(2.392986141838585E-6,2003)

(2.3558595942069394E-6,Japan)

(2.269771661250159E-6,Music_genre)

(2.0407662469318825E-6,Football_(soccer))

(2.02849249945887E-6,2001)

(1.9117628406971923E-6,2002)

(1.8927376288889488E-6,Population_density)

(1.8463276367702008E-6,Politician)

(1.8456474140455742E-6,2000)

(1.7187785992553996E-6,Personal_name)

(1.6678308992706756E-6,Europe)

(1.6185603248066548E-6,Scientific_classification)

(1.6085530789062761E-6,Italy)

(1.5916646944425264E-6,1999)

(1.5233845195684103E-6,London)

(1.5206565798284002E-6,Album)

(1.5043933410108925E-6,Actor)

(1.4757653593418732E-6,World_War_II_d045)

(1.4754324497999125E-6,Television)

(1.4618510607921713E-6,Census)

(1.4578673825675448E-6,Spain)

(1.4280236831747564E-6,Record_producer)

(1.4040347563004253E-6,Studio_album)

(1.3941074865946395E-6,1998)

(1.3518979807365079E-6,Sweden)

(1.3391939860735415E-6,Scotland)

(1.316730132027956E-6,Race_(United_States_Census)_a07d)

(1.3149656051098906E-6,Public_domain)

(1.281405739251515E-6,1997)

(1.2620811243954954E-6,California)

(1.2176913992284918E-6,Marriage)

(1.21724475271818E-6,1996)

(1.2134643325488366E-6,Square_mile)

(1.190385555854082E-6,Km²)

(1.183671968585202E-6,Norway)

(1.1721459993364824E-6,Russia)

(1.1689417460301855E-6,New_Zealand_2311)

(1.1643463112860017E-6,Film)

(1.1638321217605E-6,Per_capita_income)

(1.1587275309858643E-6,United_States_Census_Bureau_2c85)

(1.15503236196397E-6,Poverty_line)

(1.1543662062863345E-6,Wiktionary)

(1.1118542771854385E-6,Animal)

(1.10991831926306E-6,1995)

(1.0867513664498322E-6,Writer)

(1.084919549240522E-6,China)

(1.083489908287087E-6,Poland)

(1.0802516077284572E-6,New_York_3da4)

(1.077765490948881E-6,Brazil)

(1.0646900468560723E-6,1994)

(1.0646271562305105E-6,White_(U.S._Census)_c45a)

(1.059982309782747E-6,School)

(1.0595855210958827E-6,New_York_City_1428)

(1.033413534881644E-6,English_language)

(1.0267314432147992E-6,1993)

(1.001861650986148E-6,Poet)

(9.92525684259602E-7,Population)

(9.923099599168666E-7,Ireland)

(9.769296047766601E-7,1992)

(9.715657573672622E-7,Corporation)

(9.698946495312912E-7,1990)

(9.509058215830798E-7,1991)

(9.02219995091783E-7,Netherlands)

(8.883866156245144E-7,Company_(law))

(8.86932815671388E-7,1989)

(8.791286276637619E-7,Hispanic_(U.S._Census)_1387)

(8.788721455949224E-7,Latino_(U.S._Census)_5f0e)

(8.782375878673946E-7,USA_f75d)

(8.714000963421453E-7,Binomial_nomenclature)

(8.709756459735537E-7,Mexico)

(8.631150079731736E-7,1980)

(8.480158878597537E-7,1982)

(8.400986959482091E-7,South_Africa_1287)

(8.393914579358669E-7,1983)

(8.365788766815005E-7,1981)

(8.335051498047496E-7,1986)

(8.299264839234578E-7,1985)

(8.259474022915401E-7,Building)

(8.232685830273817E-7,1984)

(8.193732932333522E-7,1987)

(8.147682246335296E-7,1988)

(8.12136417926044E-7,Texas)

(8.088859114377144E-7,1979)

(7.935143382121487E-7,British_Columbia_90ec)

(7.919570775229079E-7,Band_(music))

Top 100 webpages for Spark (Small Data Set) – PageRank, WebPage

(2.7966269143827812E-11,United_States_09d4)

(2.075844358981264E-11,Wikimedia_Commons_7b57)

(1.922036564657725E-11,England)

(1.7785359847793844E-11,Germany)

(1.2692655557923408E-11,France)

(1.1114003180526154E-11,Inhabitant)

(9.75178981883102E-12,City)

(8.922354452364002E-12,Wiktionary)

(8.249475957539157E-12,Computer)

(7.989263132256183E-12,Japan)

(7.955861697277059E-12,Animal)

(7.575684695927904E-12,United_Kingdom_5ad7)

(7.425172400591612E-12,Country)

(7.324085797698805E-12,India)

(7.056117322828553E-12,Europe)

(6.759366108120827E-12,Australia)

(6.751107755939377E-12,Italy)

(6.635586537577547E-12,Water)

(6.515957825448375E-12,Canada)

(6.284520761814727E-12,Television)

(6.221934538878271E-12,English_language)

(6.2216036682062055E-12,Spain)

(6.149494855199112E-12,Plant)

(5.757014597577556E-12,Earth)

(5.510982045643055E-12,London)

(5.3816764922147224E-12,Football_(soccer))

(5.3529328658374346E-12,Scotland)

(5.3466017818813E-12,Greece)

(5.2901482528871465E-12,China)

(5.234543756262678E-12,Money)

(5.0876561858869485E-12,Music)

(4.8958993238098125E-12,Metal)

(4.880814991310529E-12,Food)

(4.832418602961475E-12,Capital_(city))

(4.821864611834572E-12,2005)

(4.78956012246475E-12,Brazil)

(4.729309726727944E-12,Netherlands)

(4.721896345517592E-12,Movie)

(4.703101284715376E-12,Human)

(4.695006406566403E-12,U.S._state_5a68)

(4.676985787652663E-12,Capital_city)

(4.655338567318036E-12,2006)

(4.490092784256378E-12,Greek_mythology)

(4.395852383956565E-12,Poland)

(4.38960306042723E-12,Russia)

(4.383708936544532E-12,Book)

(4.3211314634900385E-12,Number)

(4.311314035329429E-12,Mathematics)

(4.1281374845111735E-12,2004)

(4.126594703476583E-12,People)

(4.0428914736794845E-12,Actor)

(4.028232915127809E-12,Language)

(4.012739270588598E-12,Asia)

(3.997026794293364E-12,Government)

(3.985344066239758E-12,California)

(3.957200216647E-12,God)

(3.92410365571784E-12,Year)

(3.881334288987124E-12,Sweden)

(3.770798616512766E-12,Religion)

(3.744793181139335E-12,University)

(3.662934080053536E-12,Fruit)

(3.5975064661192052E-12,Africa)

(3.570208838763067E-12,Science)

(3.474838190492063E-12,Chemical_element)

(3.4430657917576725E-12,Film)

(3.362860102338834E-12,Car)

(3.3073789007753624E-12,Internet)

(3.305947536368759E-12,Disease)

(3.2648924868247937E-12,Company)

(3.2356104444504154E-12,World_War_II_d045)

(3.1931705560393625E-12,River)

(3.1883828964596026E-12,Species)

(3.1740869816257605E-12,19th_century)

(3.146116153350259E-12,Internet_Movie_Database_7ea7)

(3.1418978366884394E-12,Fish)

(3.1310197620120555E-12,Prefecture)

(3.128821100042407E-12,Video_game)

(3.061068835677002E-12,North_America_e7c4)

(3.0295756114327847E-12,Liquid)

(3.0163967844658693E-12,Latin)

(2.9989804853122423E-12,Singer)

(2.9929112571752424E-12,1970s)

(2.971898361113876E-12,Chad)

(2.968909394350211E-12,Island)

(2.867710853608454E-12,Sport)

(2.8503153920772218E-12,War)

(2.8274823245752885E-12,County)

(2.8128865888530896E-12,2001)

(2.810789027396672E-12,Tool)

(2.7478681279090664E-12,1960s)

(2.7297485322461806E-12,German_language)

(2.718558422987326E-12,Band)

(2.7125010288238445E-12,Greek_language)

(2.7078662287641642E-12,Dinosaur)

(2.7060358018213362E-12,Bird)

(2.698031778333844E-12,New_York_City_1428)

(2.694801905181535E-12,Km²)

(2.669740059002058E-12,Mammal)

(2.6676467128382526E-12,Tree)

(2.66102497514476E-12,Christianity)

Top 100 Web Pages for MapReduce Execution(Small Data Set)

United_States_09d4 0.0072598048169058085

Wikimedia_Commons_7b57 0.005384582467441564

Country0.005107633679144425

England0.0032972300817533046

City 0.00305954280405834

Europe 0.003036908599990767

Washington 0.003001720894464254

Earth 0.0029865291336911713

United_Kingdom_5ad7 0.002948292860661309

Water 0.002885789851578552

Animal 0.0028277451707970436

Germany 0.002763193170018334

France 0.0027123888894417135

English_language 0.0025345904308509684

Week 0.0024045400251852117

Sunday 0.002251740867636402

People 0.0022313153921155143

Monday 0.0022162114269899387

Wednesday 0.0021978986005838754

Friday 0.0021454611318054285

Capital_(city) 0.002134669596661304

India 0.0021300278826030705

Saturday 0.0021227829096531876

Wiktionary 0.0021084083874134355

Thursday 0.0020947277216928953

Tuesday 0.002081161695160513

Asia 0.0020702966527493976

Plant 0.0020388013320501385

Computer 0.0020090537547256125

Money 0.0019826922463512314

Government 0.001962969141795501

Number 0.0018784492773523548

Italy 0.0018077581769979291

Day 0.0017417068456007788

Spain 0.0016951852038328693

Japan 0.0016780932445157107

China 0.0016778048563428816

Human 0.0016679962542650187

State 0.0016400629736448085

Wikimedia_Foundation_83d9 0.00154813437621841

Canada 0.0015391202712675782

Energy 0.001526877437998015

Food 0.0015061675886222936

Sun 0.0015025231374946322

Mathematics 0.0014966836408537231

Australia 0.0014934936215895485

index 0.0014922237222228629

Science 0.001480925213121559

Scotland 0.0014469098459344968

London 0.0014188318847682035

World 0.001398935041250769

Geography 0.0013685223976877992

Planet 0.001359971573718269

Year 0.0013326452436722947

Language 0.00132987890209434

Television 0.0013247816771727193

Music 0.0013008207658821633

Society 0.0012820997892285808

Latin 0.0012632035160302908

Wikipedia 0.0012362195291525377

Russia 0.0012342124805092716

Metal 0.001227332537297126

Information 0.001215464995023806

2004 0.0012121815290869764

20th_century 0.0012084030634918503

Greek_language 0.0012079405989542347

Greece 0.0012072192419653302

Plural 0.0011752238424275411

Sound 0.0011749260518373447

Religion 0.0011597420034551895

Nation 0.0011205298472141743

Africa 0.0011126364580391773

19th_century 0.0010979122943288063

Law 0.0010867617694394164

Liquid 0.0010770732574058239

Centuries 0.001028208721590564

Scientist 0.001026053808729831

North_America_e7c4 0.0010117330673649488

Victoria 9.973300437512266E-4

Building 9.956495993492748E-4

Atom 9.94801340025492E-4

Light 9.87105051640115E-4

Poland 9.820900113063189E-4

War 9.818466823034947E-4

Population 9.772976222230652E-4

Capital_city 9.738178728355547E-4

National_anthem 9.692936582842959E-4

Continent 9.614758463573718E-4

God 9.42475808711365E-4

History 9.424224058243688E-4

List_of_decades 9.228903326220452E-4

Netherlands 9.209094906787538E-4

Inhabitant 9.154824440460432E-4

Sweden 9.116632723280301E-4

Chemical_element 8.987691782199881E-4

Culture 8.974296132266942E-4

Square_kilometre 8.914510702172977E-4

Great_Britain_b139 8.823644953530025E-4

Ocean 8.704363439350916E-4

Species 8.685973501108987E-4

Top 100 webpages for Big Data Set

United_States_09d4 0.0032094111085680713

2006 0.0026092684119483098

Washington 0.002161516065237816

United_Kingdom_5ad7 0.0014619742668534788

England0.0011494348923319672

Canada 0.0010173597196972326

Biography 8.40613765380595E-4

2005 8.342124221936725E-4

English_language 8.3014325093948E-4

France 7.812853418234293E-4

Australia 7.783425191239974E-4

January_1 7.252620722164927E-4

Germany 6.974014894266281E-4

India 6.51255630273762E-4

Wiktionary 6.351355452820647E-4

List_of_state_mottos 6.224674606120439E-4

2004 6.116655546841104E-4

January 6.084320503170525E-4

Japan 5.806138057974453E-4

Portland 5.724208293569503E-4

National_anthem 5.72303109249278E-4

February 5.521947352057445E-4

London 5.37895356223331E-4

Italy 5.23694526937374E-4

December 5.040343012602317E-4

Record_label 4.944230013972825E-4

Richmond 4.900904166094798E-4

Latin 4.886251445442015E-4

Europe 4.8657977771784147E-4

2003 4.630382448043832E-4

January_20 4.6070189808637224E-4

January_4 4.5431702438708917E-4

Population 4.531246022124034E-4

January_3 4.516074224079595E-4

Scotland 4.466948483253124E-4

January_8 4.3899309619426115E-4

January_2 4.3841932340010715E-4

Geographic_coordinate_system 4.375283852295239E-4

January_14 4.3688307054822725E-4

January_10 4.3513746135705054E-4

January_26 4.3500906026344016E-4

January_11 4.3429296660900824E-4

January_9 4.334341502606799E-4

January_7 4.280736695342911E-4

January_5 4.2775172579391935E-4

January_17 4.261145837099414E-4

January_22 4.254999938216917E-4

January_15 4.242944941830425E-4

January_30 4.2269018068877355E-4

January_19 4.2218407926628986E-4

January_13 4.2213173164836024E-4

January_6 4.2127750990908784E-4

January_12 4.211136064564662E-4

Music_genre 4.207226538391014E-4

Spain 4.1911435172529803E-4

January_25 4.177173072025361E-4

January_21 4.176995462679074E-4

January_16 4.1583156909661323E-4

January_31 4.151394589602612E-4

January_29 4.135245664864106E-4

January_24 4.130869718624833E-4

January_27 4.128300614507434E-4

January_18 4.127988043288667E-4

January_28 4.1232078491494173E-4

January_23 4.1174002323074216E-4

2001 4.0155756127585384E-4

Cambridge 3.9514126954705455E-4

2000 3.924773716040936E-4

Birmingham 3.9227095863581397E-4

Sexagenary_cycle 3.8481480260204703E-4

2002 3.8458904460029526E-4

Population_density 3.733877285007588E-4

Indiana 3.704040209156437E-4

World_War_II_d045 3.696366242425039E-4

March 3.652550805171617E-4

Russia 3.5850984678124795E-4

April 3.480037318907536E-4

Football_(soccer) 3.4765476064420434E-4

Gregorian_calendar 3.415532389378652E-4

New_York_City_1428 3.3977874814949017E-4

19th_century 3.3930735706776805E-4

Square_kilometre 3.384923102746544E-4

Internet_Movie_Database_7ea7 3.3831589606889265E-4

Austin 3.345758470401728E-4

Hamilton 3.3129821473515045E-4

Los_Angeles 3.2821010210413114E-4

French_language 3.23999183122228E-4

Scientific_classification 3.211261994694282E-4

October 3.1975717097033294E-4

1999 3.1897635392555857E-4

20th_century 3.1762941276797317E-4

New_Zealand_2311 3.144023920995408E-4

Area 3.142987414815749E-4

Philadelphia 3.1342269832802913E-4

September 3.098686733148951E-4

Census-designated_place 3.0982861973429877E-4

China 3.0850898462725243E-4

May 3.0826540698480995E-4

November 3.063973957333035E-4

Bastrop 3.0500779479025595E-4

Results are nearly same but not exactly. Maximum of top 100 web pages are same with few exceptions. This difference is due to the precision difference in spark and scala as pageRank values are very small so it creates a subtle difference with the difference in precision accuracy.