Assignment3 report

Chengze Li

(1) Design Discussion

preprocessing1:

use the file provided by professor, only change the name of main function, and use this function in mapper
the datatype get from this step is : PageName [linkPagelists]

preprocessing2:

we get the data PageName [LinkPageLists] from first steps, then we extract the nodes which exist in other nodes' linkpagelist but not exist as a independent node

```
map(pagename, linkpagelist) {
    for each page in linkpagelist
        emit(page, [])
    emit(pagename, linkpagelist)
}

reduce(pagename, [[], [], ....[linkpagelist]]){
    list = []
    for each value in values
        if(value == []) {
            continue;
        } else {
            list = value;
        }
    emit(pagename, list)
}
```

pagerank:

here the datatype we handle is pagename weight|[linkpagelist]

```
map(pagename, weight|linkpagelist) {
    split value into weight and linkpagelist
    for (each node in linkpagelist)
        emit(node, weight / linkpagelist.size)
    emit(pagename, linkpagelist)
}
```

```
reduce(pagename,[v1, v2, ...]){
    int newweight = deltaCounter / #pages;
    list = []
    for each v in values{
        if v is a number
            newweight += v;
        if v is a non-empty list
            list = v;
    }
    if list == [] {
        deltaCounter += newweight
    }

    emit(pagename , newweight|[list])
}
```

TopK

```
map(pagename, weight | linkpagelist) {
    split value into weight and linkpagelist

    emit(weight, pagename);
}

// set the comparator to sort weight descendingly
// set the reducer number to 1
Reducer{
    int counter;
    setup{
        counter = 0;
    }
    reduce(weight, [pagename1, pagename2, ....]) {
        for each pagename in values
            if (counter < 100) {
                counter ++;
                emit(pagename)
            } else {
                break;
            }
    }

}
```

Data Transferred of PageRank phrase

```
1:          Map input records=18611
             Map output records=433462
             Map output bytes=18129003
             Map output materialized bytes=19003710
             Input split bytes=153
             Combine input records=0
             Combine output records=0
             Reduce input groups=18611
             Reduce shuffle bytes=19003710
             Reduce input records=433462
             Reduce output records=18611

2:          Map input records=18611
             Map output records=433462
             Map output bytes=18139431
             Map output materialized bytes=19014138
             Input split bytes=153
             Combine input records=0
             Combine output records=0
             Reduce input groups=18611
             Reduce shuffle bytes=19014138
             Reduce input records=433462
             Reduce output records=18611

3:          Map input records=18611
             Map output records=433462
             Map output bytes=18140011
             Map output materialized bytes=19014718
             Input split bytes=153
             Combine input records=0
             Combine output records=0
             Reduce input groups=18611
             Reduce shuffle bytes=19014718
             Reduce input records=433462
             Reduce output records=18611

4:          Map input records=18611
             Map output records=433462
             Map output bytes=18139431
             Map output materialized bytes=19014138
             Input split bytes=153
             Combine input records=0
             Combine output records=0
```

```
            Reduce input groups=18611
            Reduce shuffle bytes=19014138
            Reduce input records=433462
            Reduce output records=18611

5:          Map input records=18611
            Map output records=433462
            Map output bytes=18125057
            Map output materialized bytes=18999764
            Input split bytes=153
            Combine input records=0
            Combine output records=0
            Reduce input groups=18611
            Reduce shuffle bytes=18999764
            Reduce input records=433462
            Reduce output records=18611

6:          Map input records=18611
            Map output records=433462
            Map output bytes=18129003
            Map output materialized bytes=19003710
            Input split bytes=153
            Combine input records=0
            Combine output records=0
            Reduce input groups=18611
            Reduce shuffle bytes=19003710
            Reduce input records=433462
            Reduce output records=18611

7:          Map input records=18611
            Map output records=433462
            Map output bytes=18138932
            Map output materialized bytes=19013639
            Input split bytes=153
            Combine input records=0
            Combine output records=0
            Reduce input groups=18611
            Reduce shuffle bytes=19013639
            Reduce input records=433462
            Reduce output records=18611

8:          Map input records=18611
            Map output records=433462
            Map output bytes=18142295
```

```
        Map output materialized bytes=19017002
        Input split bytes=153
        Combine input records=0
        Combine output records=0
        Reduce input groups=18611
        Reduce shuffle bytes=19017002
        Reduce input records=433462
        Reduce output records=18611

9:      Map input records=18611
        Map output records=433462
        Map output bytes=18138456
        Map output materialized bytes=19013163
        Input split bytes=153
        Combine input records=0
        Combine output records=0
        Reduce input groups=18611
        Reduce shuffle bytes=19013163
        Reduce input records=433462
        Reduce output records=18611

10:     Map input records=18611
        Map output records=433462
        Map output bytes=18142124
        Map output materialized bytes=19016831
        Input split bytes=153
        Combine input records=0
        Combine output records=0
        Reduce input groups=18611
        Reduce shuffle bytes=19016831
        Reduce input records=433462
        Reduce output records=18611
```

Summary:

as we can see from the reports above, input records of reducer and mapper remain same during the whole phrase, only difference the output bytes of mapper and input bytes of reducer

Performance Comparison
for 11 machines test:

(1) preprocessing:

GC time elapsed (ms)=174987
CPU time spent (ms)=17016360

(2) pagerank (10 iterations)

GC time elapsed (ms)=31971
CPU time spent (ms)=979900

GC time elapsed (ms)=33387
CPU time spent (ms)=981740

GC time elapsed (ms)=31344
CPU time spent (ms)=978160

GC time elapsed (ms)=31605
CPU time spent (ms)=971010

GC time elapsed (ms)=31693
CPU time spent (ms)=979200

GC time elapsed (ms)=30939
CPU time spent (ms)=987780

GC time elapsed (ms)=32061
CPU time spent (ms)=972490

GC time elapsed (ms)=33050
CPU time spent (ms)=975310

GC time elapsed (ms)=31335
CPU time spent (ms)=992060

GC time elapsed (ms)=31971
CPU time spent (ms)=979900

(3) top-k

GC time elapsed (ms)=19455
CPU time spent (ms)=166600

6-machines test

(1) preprocess:

```
GC time elapsed (ms)=180622
CPU time spent (ms)=17770690
```

(2) pagerank:

```
GC time elapsed (ms)=48575
CPU time spent (ms)=1094370

GC time elapsed (ms)=48250
CPU time spent (ms)=1104930

GC time elapsed (ms)=49497
CPU time spent (ms)=1105890

GC time elapsed (ms)=46347
CPU time spent (ms)=1102920

GC time elapsed (ms)=47865
CPU time spent (ms)=1104040

GC time elapsed (ms)=45966
CPU time spent (ms)=1095500

GC time elapsed (ms)=47407
CPU time spent (ms)=1101210

GC time elapsed (ms)=46583
CPU time spent (ms)=1096120

GC time elapsed (ms)=47808
CPU time spent (ms)=1096080
```

top-k

```
GC time elapsed (ms)=34286
CPU time spent (ms)=202310
```

summary: from the data above, we can see the time spent of 11-machines for preprocess and page rank phrase is less than that of 6 machines, but the top-k phrase, both spend almost same times, I think the reason is I set the number of reducer to one for this phrase, so the number of workers cannot affect the final result.

Top-100 pages for simple dataset:

United_States_09d4	0.00779508470850387 2
Country	0.00555696355703316 5
Wikimedia_Commons_7b57	0.0054565858066 50162
Week	0.0038648293747876126
Earth	0.0036433900656126987
Water	0.0035660358651619416
Europe	0.003540515903866633
United_Kingdom_5ad7	0.003325854294505277
Sunday	0.003160578208253685
Monday	0.0030990432765067074
Wednesday	0.0030616375267338534
Animal	0.0029795039495123933
Friday	0.0029788738973597494
Saturday	0.0029446612279560636
Thursday	0.0029030539274661647
Tuesday	0.002884472971560805
France	0.0028125854658935397
Asia	0.002798708024421315
index	0.002794136706286621
Day	0.002788166953902265
City	0.002691970788244633
England	0.0025179681169618527
Germany	0.0024328194104712066
Money	0.0024248486126411646
Government	0.0023455953278073067
Number	0.002294529268189954
Plant	0.0022438081003952532
English_language	0.00222115678678706
India	0.0021375734711041564
Energy	0.00208397461743284
Wiktionary	0.0020788011523132435
Sun	0.002064131682206717
Italy	0.0020504855648718883
Computer	0.0020049342590691705
Wikimedia_Foundation_83d9	0.00189576660090349
People	0.0018723762202493702
Canada	0.0018331045051205384
Science	0.0017812590294018335
Human	0.0017689478225635647
Spain	0.0017177941221262173
Planet	0.0017138072231851708

China	0.0016751090219201948
Japan	0.001651089201198659
State	0.0016054393243545232
Year	0.0015843089458683487
Australia	0.0015793151017126658
Food	0.0015737829814565726
Mathematics	0.001567487919357772
Russia	0.0014930550238222765
Wikipedia	0.0014901570175595183
Capital_(city)	0.001488433993697315
Greek_language	0.0014499467232963416
Geography	0.0014066631834356847
Language	0.0013713010369973799
Atom	0.0013450111577835225
Metal	0.0013334540754134446
Society	0.0013232837918391335
Liquid	0.0013160553590647994
Africa	0.0013095856135534514
Greece	0.0013027292467736253
Sound	0.0012959123603408864
World	0.001267486256568155
Scotland	0.0012583857276836485
Law	0.0012376568023006666
Religion	0.0012326366909645831
Television	0.0012324623240550562
Moon	0.0012230712023575568
Light	0.0012218903901886926
Scientist	0.0012146637480146731
Culture	0.001209829601424602
History	0.0012093489183293924
2004	0.001206837758318248
Cyprus	0.0011851686949997233
Turkey	0.001174922141926617
Plural	0.0011735546225080803
20th_century	0.0011440031594343132
Latin	0.001130608973722957
Music	0.001121613817468201
Poland	0.001117249800996265
19th_century	0.0010934176772870296
Sweden	0.0010927986426162695
Gas	0.0010853395268795147
War	0.0010820043151925638
Information	0.0010807464122587515
Circle	0.0010798941828956542

Ocean	0.0010726202539562955
Building	0.0010630519873470246
Denmark	0.0010362238453787882
Portugal	0.0010354696721030753
Solid	0.001033953931195375
Chemical_element	0.001022402050720792
London	0.001018395546158795
Nation	0.0010153040513736973
Trade	0.0010036348041383128
Electricity	0.0010018546990661068
Austria	9.853712173284843E-4
Continent	9.837196086953863E-4
God	9.736670945175397E-4
Image	9.665078709562856E-4
Netherlands	9.636417222753058E-4

Top-100 pages for full dataset

2006	0.003129761676841186
United_States_09d4	0.003084106663879893
United_Kingdom_5ad7	0.0015401122628263702
2005	0.0013856450467681302
France	0.0010563691251414904
2004	9.474824207767575E-4
England	9.384905759089626E-4
Canada	9.28036220356211E-4
Germany	8.669872683384925E-4
Australia	7.703862240353153E-4
2003	7.595945648308346E-4
Japan	7.101613102397975E-4
Biography	6.851405224082526E-4
India	6.718200778902676E-4
Italy	6.638499915386085E-4
Geographic_coordinate_system	6.422485099561578E-4
2002	5.945674475806181E-4
Europe	5.939109877798996E-4
2001	5.895860684285092E-4
World_War_II_d045	5.8085174827925E-4
English_language	5.767555117079104E-4
2000	5.550584587763412E-4
London	5.405438148986172E-4
Spain	5.256771626654204E-4
Wikimedia_Commons_7b57	5.237377318946748E-4

Russia	5.167191320670436E-4
Wiktionary	5.043957116022048E-4
1999	4.998511894409066E-4
Internet_Movie_Database_7ea7	4.973146316091873E-4
Race_(United_States_Census)_a07d	4.6946063707877563E-4
Population_density	4.4989607329409397E-4
1998	4.289606034809359E-4
New_York_City_1428	4.249407405561955E-4
1997	4.150817885848992E-4
Scotland	3.9993012852321936E-4
1996	3.8758132603765056E-4
Netherlands	3.771388905441717E-4
China	3.7593859141933617E-4
1995	3.6829611147202285E-4
Sweden	3.6643969381296854E-4
Record_label	3.5557233856122757E-4
1994	3.5314357700366484E-4
January_1	3.5233395698329624E-4
Latin	3.475792101418799E-4
1991	3.475077845871846E-4
Square_mile	3.40197757704676E-4
California	3.393795541937075E-4
1990	3.3837383847054207E-4
New_Zealand_2311	3.370430413272965E-4
Television	3.348804543000695E-4
1993	3.324007589777933E-4
French_language	3.310361596729521E-4
1992	3.2143086092010234E-4
New_York_3da4	3.144697108828665E-4
Sexagenary_cycle	3.141265357260364E-4
Public_domain	3.084475631874964E-4
index	3.0809246152302313E-4
Census	3.077634565293695E-4
1989	3.065352279896954E-4
1980	3.051892098962106E-4
Ireland	3.0120389943691215E-4
Soviet_Union_ad1f	3.005059596870991E-4
Football_(soccer)	2.9962166261472353E-4
Poland	2.9789207648203643E-4
1986	2.947589168904814E-4
Music_genre	2.9299369136669796E-4
1974	2.925401088165239E-4
1979	2.9235977078838067E-4
1945	2.9053588260886254E-4

```
1970	2.887905464423691E-4
1981	2.8695500683367943E-4
Mexico2.8586155082312685E-4
Norway		2.8474060844772223E-4
1982	2.844600489911612E-4
United_States_Census_Bureau_2c852.8439434831168254E-4
1985	2.841784722386043E-4
Population	2.839566299873262E-4
Switzerland	2.828685469107125E-4
Egypt	2.822409021722773E-4
1976	2.818618406814446E-4
1969	2.794306009317297E-4
1975	2.790033481596211E-4
1984	2.765674963846599E-4
Gregorian_calendar	2.760010753403061E-4
1983	2.757596061667399E-4
Film	2.7508728483789034E-4
Greece2.745933837414386E-4
1987	2.738121955350754E-4
1972	2.737788806629311E-4
Paris	2.730324671309625E-4
South_Africa_1287	2.730159781659288E-4
Brazil	2.7226188865147287E-4
Greek_language		2.707223268651165E-4
Portugal		2.695489894380744E-4
1988	2.675588012559257E-4
Austria2.6720459494902055E-4
1977	2.6663575669959744E-4
1973	2.665354072849195E-4
1971	2.6489234729184494E-4
Denmark		2.633544557633728E-4
```

Summary: I think the result make sense, because the full data set was created at 2006, so 2006 is the most important page in the dataset, also, because the data is for US, so the US country page is also a very important page;