CS6240 Section01 – HW3 Report
Submitted by: Rajkumar Murukeshan

Design Discussion

1) Pre-processing Job:

Pre-processing job takes the input file (Bz2Compressed File) and and creates an output file that contains the Page Name and its adjacency list separated by a delimiter ":" and the adjacency list is delimited by using a delimiter "~".

Note : **Bz2WikiParser** does not handles the self links (links that has reference to its own page). It does not remove the self links present in the adjacency list. Hence there can be fluctuations in the output result

This job make use of only mapper and the output is emitted to the file without a reducer. Another job takes the output of pre processing job and adds a default PageRank (equally distributed across all pages) and emits the output to file in the format given below:

<pagename : pageRank : adjacencyList>

PseudoCode

```
enum Counter{
   NodeCounter,
   Dangling_PR_Sum
};

map(key, value){
   String line= value.toString();
   // perform Bz2WikiParser and return the output as String
   String preProcessedResult= Bz2WikiParser(line);
   context.write(pageName, adjacencyList);
   NodeCounter.increment(1);
}
```

Default PageRank job PseudoCode

```
map(key, value){
   String line = value.toString();
   // get pageName, adjacencyList from line
   defaultPageRank = 1/NodeCounter;
   context.write(pageName, defaultPageRank, adjacencyList)
```

```
    // increment dangling node pagerank sum
    if(adjacencyList is empty){
        // increment Dangling_PR_Sum with the pageRank of dangling Node
     }
}
```

After the job is completed, **RESET** the counter that calculates the page rank sum of all dangling nodes.

2) **PageRank Job:**

PageRank job takes the output of Pre-processing job as an input and performs calculation for obtaining the page rank of all pages.

Formula used to calculate the Page Rank :

$$PR(A) = ((1 - \alpha)/ \text{Total No of Pages}) + \alpha(\delta + \Sigma (z(b)))$$

where, α = Damping Factor (0.85 is used)
δ = Sum of Page Ranks from Dangling Nodes
z (b) = PR of a page B/ Total number of outlines of B

and b are the pages that has an outline to A.

PseudoCode

```
map(key,value){
    String line= value.toString();
    // get pageName, pageRank and adjacency list from line;
    emit(pageName, pageRank, adjacencyList);
    for(each node in adjacency list){
        emit(pageRank/|adjacency list|);
    }
}

reduce(pageName, Iterable<Object> values){
    prSum = 0.00;
    for(each value in values){
        if(contribution){
            prSum += value.getPageRank();
        } else {
            // get adjacencyList
        }
```

```
    }
    constant1 = (1- DampingFactor)/NodeCounter;
    constant2 = (DampingFactor) * Dangling_PR_Sum;
    pageRank = constant1 + constant2 + (DampingFactor) * prSum;
    emit(pageName,pageRank with adjacencyList);
}
```

This job is iterated for 10 times where the output of an iteration is given as an input to the next iteration.

Note: After every iteration, **RESET** the page rank sum of all dangling nodes.

3) **Top-k Job:**

This job takes the output of the PageRank job, after the 10th iteration as its input.

PseudoCode

```
Class Page{
    String pageName;
    Double pageRank;

    // getters, setters and constructors

}

Mapper {

// create a priority Queue data structure with customized comparator

PriorityQueue pq = new PriorityQueue(100, Comparator)

map(key, value){
    String line= value.toString();
    // get pageName and pageRank from line;
    Page pg = new Page(pageName, pageRank);
    pq.add(pg);
}

// mapper clean up
cleanup(){
    context. write(top 100 pages in descending order by pageRank);
}
```

```
}
```

Reducer {

```
   int count = 100;

   reduce(Page, Iterable<> values){
      for(each value in values){
         if(count<100){
            context.write(pageName, pageRank);
            count++;
         } else {
            break;
         }
      }
   }
}
```

Performance Comparison

Time Values:

- 6 m4.large machines (1 master and 5 workers)

  Pre-processing Time: **38 minutes**
  Time to run 10 PageRank Iterations: **22 minutes**
  Time to find top-100 pages: **3 minutes**

- 11 m4.large machines (1 master and 10 workers)

  Pre-processing Time: **23 minutes**
  Time to run 10 PageRank Iterations: **16 minutes**
  Time to find top-100 pages: **2.35 minutes**

SpeedUp:

Pre-processing Time: **1.65**
Time to run 10 PageRank Iterations: **1.375**
Time to find top-100 pages: **1.27**

From the above speed up values, we see that pre-Processing phase has good speedUp than the other two. Also, the top-K job phase shows less speedup when compared to other 2 phases. It is because, all the computations will be made in a single reducer in top-K Job.

Top- 100 Wikipedia pages:

Simple Data Set:

United_States_09d4:0.0041382193465633 78
Wikimedia_Commons_7b57:0.003120430006557257
Country:0.0026337699819028626
Europe:0.001746615156662618
Water:0.0017262589601026216
United_Kingdom_5ad7:0.0017253611827319085
England:0.001707351947314057
Germany:0.0016603688703253291
France:0.0016463639418695902
Earth:0.0016458976898919334
Animal:0.0016069441242109398
City:0.0015658646014085151
Week:0.001426905675719541
Asia:0.0013007599843611428
Sunday:0.0012934554205942443
Monday:0.0012734573290324118
Wednesday:0.001260332889137803
Friday:0.0012292317994213116
Money:0.0012202541061225195
Saturday:0.0012152425694691338
Wiktionary:0.0012069055261110888
Thursday:0.0011993052763976022
Tuesday:0.0011908709195764988
Plant:0.001187828477203498
English_language:0.0011557503074608303
Government:0.0011441664463965708
Computer:0.0011421091022395429
Italy:0.0011278895351462528
India:0.0011264368800129364
Number:0.0010690713119104085
Day:0.0010421772325826437
Spain:0.0010212409003303 59
Canada:9.891310090211548E-4

Japan:9.83219518763292E-4
People:9.592015815482228E-4
index:9.382091779333301E-4
Human:9.347049639350062E-4
Wikimedia_Foundation_83d9:9.20994150746533E-4
Energy:9.078918843899653E-4
China:8.996840896487081E-4
Australia:8.929245630344868E-4
Sun:8.865215472368123E-4
Science:8.620970350969396E-4
Food:8.596930199905561E-4
Mathematics:8.380877872459494E-4
Capital_(city):7.928094571502788E-4
Television:7.840411079879893E-4
State:7.838561948489433E-4
Russia:7.827895216076777E-4
Year:7.614377910970252E-4
Music:7.365965204679895E-4
Language:7.325583443979326E-4
Greece:7.289792340898346E-4
Scotland:7.204011991232603E-4
Wikipedia:7.183702754863952E-4
Planet:7.135205983132542E-4
Metal:7.111791918986797E-4
Greek_language:7.100846293047824E-4
2004:6.953718310840275E-4
Sound:6.770228195090504E-4
Africa:6.731037079221131E-4
Religion:6.715865190499122E-4
London:6.528898740350391E-4
Geography:6.421343707442346E-4
Poland:6.313506587140175E-4
Law:6.297470214624958E-4
20th_century:6.285784232441156E-4
Liquid:6.276432421912774E-4
World:6.196542867041796E-4
Society:6.139110357916714E-4
19th_century:6.131253660807593E-4
Scientist:6.058117866302839E-4
Atom:5.959765934826966E-4
History:5.889604416798877E-4
Latin:5.834088002738668E-4
Light:5.794126174719832E-4

Sweden:5.7705796702386E-4
War:5.744438300470318E-4
Culture:5.729674691893549E-4
Netherlands:5.60345653200073E-4
Turkey:5.552940858820192E-4
Building:5.548063662836819E-4
Plural:5.494137953546792E-4
Information:5.427266340343299E-4
God:5.421324788813206E-4
Portugal:5.29122625747465E-4
Chemical_element:5.242305539312608E-4
Centuries:5.232086715986285E-4
Denmark:5.214314453499498E-4
Cyprus:5.20240637023904E-4
Austria:5.135365587953691E-4
Capital_city:5.099247076562314E-4
Ocean:5.087583914078682E-4
Moon:5.070359047305244E-4
North_America_e7c4:4.981478324470158E-4
Inhabitant:4.917658603366286E-4
Biology:4.907919801981042E-4
Electricity:4.899078263926507E-4
Disease:4.880011812337834E-4
University:4.8744413040557253E-4

Full Data Set:

United_States_09d4:0.0012945077389545732
2006:0.0011827306222905703
United_Kingdom_5ad7:6.204844648471361E-4
2005:5.419208598425181E-4
France:4.053060379109539E-4
Biography:4.0107830388496984E-4
Canada:3.9971460333026836E-4
England:3.9882442407127803E-4
2004:3.7626694377342095E-4
Germany:3.447300732634116E-4
Australia:3.276576093057589E-4
Geographic_coordinate_system:3.1267872271285223E-4
2003:3.0287770881646316E-4
Japan:2.8937473694603855E-4
India:2.889790557647477E-4
Italy:2.488779904765754E-4

2001:2.4149980212014597E-4
2002:2.3958094600607665E-4
Europe:2.326710432515422E-4
Internet_Movie_Database_7ea7:2.2992205235734594E-4
2000:2.2622951700374143E-4
World_War_II_d045:2.2194618570963397E-4
London:2.1248667379942897E-4
English_language:2.0615359134857145E-4
Spain:2.0199834958044562E-4
1999:2.0072666878382192E-4
Population_density:1.9898547829908685E-4
Russia:1.922507468317514E-4
Record_label:1.891488813422223E-4
Wiktionary:1.8754567713942302E-4
Race_(United_States_Census)_a07d:1.8680104726982632E-4
Wikimedia_Commons_7b57:1.816751976283209E-4
1998:1.732561106543811E-4
1997:1.6575247557027155E-4
New_York_City_1428:1.6416805554283096E-4
Scotland:1.6253717210332435E-4
Music_genre:1.590886687169112E-4
1996:1.5533989277785427E-4
Sweden:1.520974056074079E-4
Football_(soccer):1.512553533150498E-4
Television:1.490025593268889E-4
1995:1.465769808036504E-4
Square_mile:1.451108457579396E-4
China:1.4508490848468984E-4
California:1.4328965411325106E-4
Netherlands:1.4321663577832286E-4
Census:1.4243695843988783E-4
1994:1.4007297772554274E-4
New_Zealand_2311:1.3942028909596593E-4
1991:1.3453245231834495E-4
1993:1.3234157688510976E-4
1990:1.3220514351731938E-4
Public_domain:1.2978250541175016E-4
New_York_3da4:1.2960012632448763E-4
1992:1.2710434214036333E-4
United_States_Census_Bureau_2c85:1.2348815230790288E-4
Ireland:1.223617708284203E-4
Film:1.2233136942618066E-4
Norway:1.216679455501561E-4

Poland:1.2158489325981783E-4
January_1:1.2112286039049244E-4
Population:1.2000768523499299E-4
1989:1.1961878333767445E-4
Actor:1.1961196789605466E-4
Latin:1.1942606864736057E-4
Scientific_classification:1.1807556676160568E-4
1980:1.1739592977051753E-4
Mexico:1.1436936033945622E-4
French_language:1.1426911034113178E-4
Brazil:1.1400767380581339E-4
1986:1.1396779051564798E-4
1979:1.1154728583379149E-4
Marriage:1.1151630702012191E-4
1985:1.1089964230625492E-4
1981:1.1081354199146538E-4
1982:1.1066178591994702E-4
1974:1.103865778557386E-4
Switzerland:1.087198686325177E-4
1984:1.0826879928752988E-4
1983:1.0811768010863369E-4
South_Africa_1287:1.0793261498479042E-4
1987:1.0791791317143889E-4
1970:1.0783749314201945E-4
Politician:1.068612247643959E-4
1976:1.0630933207888349E-4
1988:1.0548225142987685E-4
1975:1.0511469515638026E-4
Per_capita_income:1.047333022662913E-4
1945:1.0456081547022587E-4
Soviet_Union_ad1f:1.0418509975589408E-4
1969:1.0397364329454593E-4
Paris:1.0361284915343424E-4
Greece:1.0349942197292667E-4
1972:1.0300227918817318E-4
1977:1.0177391993978314E-4
1978:1.010493616733822E-4
Album:1.0058786439330313E-4
1973:1.0043781570346117E-4
Portugal:1.0009919948654873E-4
Record_producer:9.994144340114062E-5