1. **Project goal**

Analysis Wikipedia dump data by Hadoop and get which title the users most interested in. Use Word Cloud to see the result at a time. So we can find out what is people’s interests in that month.

1. **used DataSet**



We use the meta-history data which contains Wikipedia edit records. The data has title, edit time, and editor ID. We use title, and timestamp so we are going to find out what is the people’s interests per month.

1. **What I did**

**(1) summary the metadata**.

Each part of wiki dump files were about 50~70GB, and we don’t need the <text></test>area which occupying most space of the data, so we take the information which we need (title, timestamp). We use shell command for this work.

First we take the line we need..

**cat test | grep “<title>\|<timestamp>” > out**

Erase “\n” then make the file a line.

**cat test | tr “\n” “\t” > out**

Change <title> to \n<title> and delete empty spaces.

**cat test | sed ‘s/<title>/\n<title>/g’ | sed ‘s/\t//g’ | sed ‘s/ //g’ | sed ‘s/ //g’ > out**

Delete <title> </title> </timestamp> and change <timestamp> to “\t”.

**cat test | sed ‘s/<title>//g’ | sed ‘s/<\/title>//g” | sed ‘s/<timestamp>/\t/g’ | sed ‘s/<\/timestamp>//g’ > out**

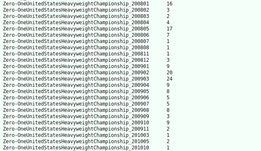
|  |  |
| --- | --- |
|  |  |
| Original dada | Modified data |

We changed 2TB data to 10GB.

**(2) run Hadoop**

We make mapreduce program like wordcount which key is “title + “\t” + year + month”

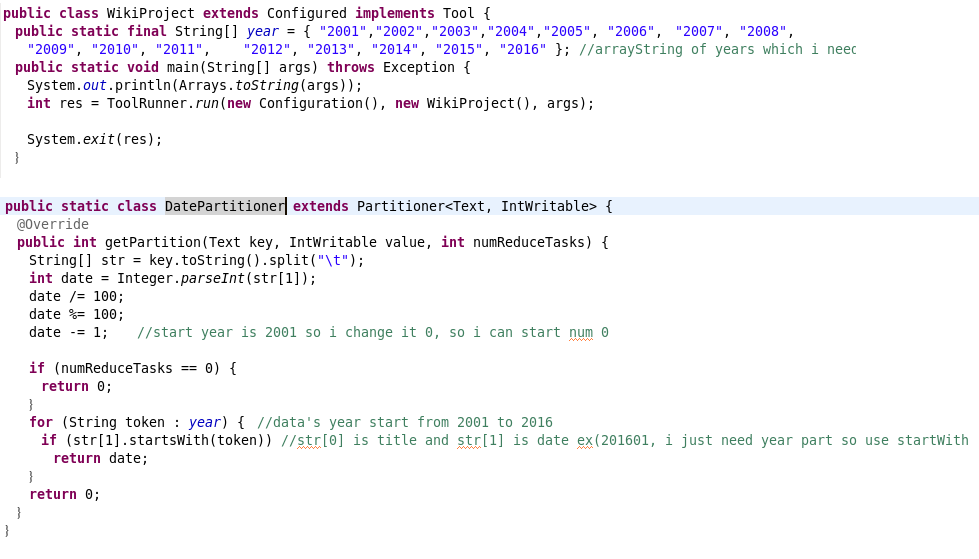
ex) title 20160424

****And value is the count of each keys.



We got problems with GC overhead limit exceeded error and java heap error.

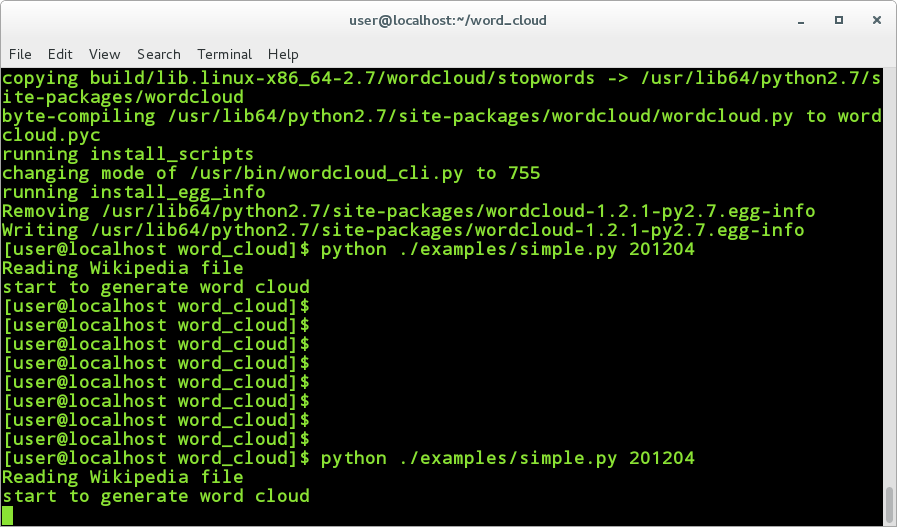
It was solved by setting the overall memory size in the mapred-site.xml file.



Wiki data were from 2001 to 2016, so I made 16 reduce tasks and used a pertitioner, so each task processed the data from 2001 to 2016.

**(3) use Word Cloud**

We use Word Cloud to see the result which the counts of edit histories per year at a time. Existing Word Cloud is just use for show the frequency of the document so we modified it to use our output file.



A user inputs python.file and what he wants know year and month on the terminal, then the program takes the year which he inputted from the XML parsing data. It explores all data and find the data which user wants. We use the dictionary data structure to manage the data. There are useless data which contain like “User:”, “Template:”, “Talk:” among the titles. We use stopword to ignore these titles.



This is the result of July in 2012. We can see the movie “The Dark Knight Rises” and “Amazing spider man” which released in July in 2012 are most edited title.



We can see Fukushima and Libyan CivilWar in March in 2011.

1. **REFERENCES**

* Wikipedia Database

<https://en.wikipedia.org/wiki/Wikipedia:Database_download>

* Open large text files

<http://www.readfileonline.com/>

* Hadoop on Wikipedia

<https://tpmoyer-gallery.appspot.com/hadoopWikipedia>