Report for W11-Practical
CS1003 W09 Practical 170025298 25th ,April 2018

Overview:

This practical uses Hadloop to map all the data and reduce it to count how many times the urls apperaed. In my practical I added one JsonRead class and changes some in main class and mapper class. The changes will be shown in design part.

This practical have several extension in extension folder and the basic program is W11Practical.

Design:

ScanWordsMapper- map – void:

```
// the value is a line from the file.
ReadJson readJson = new ReadJson();

String line = value.toString();
Scanner scanner = new Scanner(line);

List<String> urlList = readJson.read(line);

if (!urlList.isEmpty()) {
    for(int i = 0; i < urlList.size(); i++) {
        String word = urlList.get(i);
        output.write(new Text(word), new LongWritable( value: 1));
    }
}
scanner.close();
```

I found that the previous code does not fit to my expectation so I change the code as above:

Because the list return may empty, so I putted output only available when the url list does exist.

This method uses list the provided by ReadJson class which is a list that contains the elements that needed. For loop is used for get all the elements in this list.

ReadJson – read(String input) – List<String>

This class read Json file that provided by mapper and it uses JsonReader API to read the Json line. Because of the nullable url and expanded_url, So I uses judgement to judge if it is null to prevent null pointer exception:

```
if(object.containsKey( o: "entities")){
    object = object.getJsonObject("entities");
    if(object.containsKey( o: "urls")) {
        JsonArray urls = object.getJsonArray( s: "urls");
        for (int i = 0; i < urls.size(); i++) {
            for (String key : urls.getJsonObject(i).keySet())
                if (key.equals("expanded_url")) {
                    JsonObject linkObject = urls.getJsonObject(i);
                      JsonValue value = linkObject.getJsonString("key");
                    String link;
                      System.out.println(linkObject.isNull(key) + key);
                    if (!linkObject.isNull(key)) {
                        if (linkObject.getString(key) != null) {
                            link = "\"" + linkObject.getString(key) + "\""; //<- issue in 17:50 Tuesday
                            list.add(link);
                        }
                    }
                      System.out.println(linkObject.getJsonString(key).toString());
                }
        }
    }
}
```

Main:

Map reduce area is not changed but I found that the hadloop may return file exist when files are available. So I add file.delete to prevent it:

```
//set output file:
File output = new File(output_path);
if (output.exists()) {
    String[] childFile = output.list();
    if (childFile == null) {
        output.delete();
    }
    else {
        for (String aChildFile : childFile) {
            File file = new File( pathname: output_path + "/" + aChildFile);
            file.delete();
        }
        output.delete();
    }
}
```

Testing:

```
pc2-143-c:~/Documents/cs1003/W11Practical hl74$ stacscheck /cs/studres/CS1003/Practicals/W11/Tests/
Testing CS1003 Week 11 Practical
- Looking for submission in a directory called 'src': found in current directory
* BUILD TEST - basic/build : pass
* COMPARISON TEST - basic/Test01_no_arguments/progRun-expected.out : pass
* TEST - basic/Test02_data-very-small-00/test : pass
* TEST - basic/Test03_data-very-small-01/test : pass
* TEST - basic/Test04_data-very-small-all-files/test : pass
* TEST - basic/Test05_1_minute_of_data/test : pass
* TEST - basic/Test06_10_minutes_of_data/test : pass
* INFO - basic/TestQ_CheckStyle/infoCheckStyle : pass
--- submission output ---
Starting audit...
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./log4j.properties:0: File does not end with a newline. [NewlineAtEndOfFile]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./CountWordsReducer.java:8: Missing a Javadoc comment. [JavadocType]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./CountWordsReducer.java:10:1: File contains tab characters (this is the first instance). [FileTabCharacter]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./CountWordsReducer.java:12:9: Missing a Javadoc comment. [JavadocMethod]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./CountWordsReducer.java:18:20: 'for' is not followed by whitespace. [WhitespaceAround]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ScanWordsMapper.java:11: Missing a Javadoc comment. [JavadocType]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ScanWordsMapper.java:13:1: File contains tab characters (this is the first instance). [FileTabCharacter]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ScanWordsMapper.java:15:9: Missing a Javadoc comment. [JavadocMethod]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ScanWordsMapper.java:27:28: 'for' is not followed by whitespace. [WhitespaceAround]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ReadJson.java:1: Using the '.*' form of import should be avoided - javax.json.*. [AvoidStarImport]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ReadJson.java:2: Using the '.*' form of import should be avoided - java.io.*. [AvoidStarImport]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ReadJson.java:17: Expected an @return tag. [JavadocMethod]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ReadJson.java:17:37: Expected @param tag for 'input'. [JavadocMethod]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ReadJson.java:36:11: 'if' is not followed by whitespace. [WhitespaceAround]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ReadJson.java:36:43: '{' is not preceded with whitespace. [WhitespaceAround]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ReadJson.java:38:15: 'if' is not followed by whitespace. [WhitespaceAround]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./ReadJson.java:41: 'for' construct must use '{}'s. [NeedBraces]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./W11Practical.java:14: Missing a Javadoc comment. [JavadocType]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./W11Practical.java:16:1: File contains tab characters (this is the first instance). [FileTabCharacter]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./W11Practical.java:16:9: Missing a Javadoc comment. [JavadocMethod]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./W11Practical.java:20: Line has trailing spaces or tabs. [RegexpSingleline]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./W11Practical.java:21:19: 'if' is not followed by whitespace. [WhitespaceAround]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./W11Practical.java:49: Line has trailing spaces or tabs. [RegexpSingleline]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./W11Practical.java:52: Line has trailing spaces or tabs. [RegexpSingleline]
[WARN] /cs/home/hl74/Documents/cs1003/W11Practical/src/./W11Practical.java:70: Line has trailing spaces or tabs. [RegexpSingleline]
Audit done.
---

8 out of 8 tests passed
```

8/8 tests passed.

The hardness happened when I read Json which returns null pointer exception, My classmates asks me to use try and catch to prevent Null in ReadJson, but I found that the exception happens as a result of null in url arrays

I have done several in-program tests such as I test whether the object is null in ReadJson class, which alert me to add if(object.isnull) to prevent the object is null condition.

Examples:

```
18/04/27 16:37:03 DEBUG lib.MutableMetricsFactory: field org.apache.hadoop.metrics2.lib.MutableRate org.apache.hadoop.security.UserGroupInformation$UgiMetrics
  .loginSuccess with annotation @org.apache.hadoop.metrics2.annotation.Metric(sampleName=Ops, about=, always=false, type=DEFAULT, valueName=Time, value=[Rate of
  successful kerberos logins and latency (milliseconds)])
18/04/27 16:37:03 DEBUG lib.MutableMetricsFactory: field org.apache.hadoop.metrics2.lib.MutableRate org.apache.hadoop.security.UserGroupInformation$UgiMetrics
  .loginFailure with annotation @org.apache.hadoop.metrics2.annotation.Metric(sampleName=Ops, about=, always=false, type=DEFAULT, valueName=Time, value=[Rate of
  failed kerberos logins and latency (milliseconds)])
18/04/27 16:37:03 DEBUG lib.MutableMetricsFactory: field org.apache.hadoop.metrics2.lib.MutableRate org.apache.hadoop.security.UserGroupInformation$UgiMetrics
  .getGroups with annotation @org.apache.hadoop.metrics2.annotation.Metric(sampleName=Ops, about=, always=false, type=DEFAULT, valueName=Time, value=[GetGroups])
18/04/27 16:37:03 DEBUG impl.MetricsSystemImpl: UgiMetrics, User and group related metrics
18/04/27 16:37:03 DEBUG util.Shell: Failed to detect a valid hadoop home directory
java.io.IOException: HADOOP_HOME or hadoop.home.dir are not set.
    at org.apache.hadoop.util.Shell.checkHadoopHome(Shell.java:326)
    at org.apache.hadoop.util.Shell.<clinit>(Shell.java:351)
    at org.apache.hadoop.util.StringUtils.<clinit>(StringUtils.java:80)
    at org.apache.hadoop.security.SecurityUtil.getAuthenticationMethod(SecurityUtil.java:610)
    at org.apache.hadoop.security.UserGroupInformation.initialize(UserGroupInformation.java:273)
    at org.apache.hadoop.security.UserGroupInformation.ensureInitialized(UserGroupInformation.java:261)
    at org.apache.hadoop.security.UserGroupInformation.loginUserFromSubject(UserGroupInformation.java:791)
    at org.apache.hadoop.security.UserGroupInformation.getLoginUser(UserGroupInformation.java:761)
    at org.apache.hadoop.security.UserGroupInformation.getCurrentUser(UserGroupInformation.java:634)
    at org.apache.hadoop.mapreduce.task.JobContextImpl.<init>(JobContextImpl.java:72)
    at org.apache.hadoop.mapreduce.Job.<init>(Job.java:142)
    at org.apache.hadoop.mapreduce.Job.getInstance(Job.java:185)
    at org.apache.hadoop.mapreduce.Job.getInstance(Job.java:204)
    at W11Practical.main(W11Practical.java:53)
18/04/27 16:37:04 DEBUG util.Shell: setsid exited with exit code 0
18/04/27 16:37:04 DEBUG security.Groups:  Creating new Groups object
18/04/27 16:37:04 DEBUG util.NativeCodeLoader: Trying to load the custom-built native-hadoop library...
18/04/27 16:37:04 DEBUG util.NativeCodeLoader: Failed to load native-hadoop with error: java.lang.UnsatisfiedLinkError: no hadoop in java.library.path
18/04/27 16:37:04 DEBUG util.NativeCodeLoader: java.library.path=/usr/local/idea/bin:/cs/home/hl74/usr/lib:/usr/java/packages/lib/amd64:/usr/lib64:/lib64:/lib
  :/usr/lib
```

I don’t know to to show the example of this because all of these are hadloop running process. But as you can see, there is an exception occurred that I don’t know where it comes from.

```
18/04/27 16:37:08 INFO mapreduce.Job: Counters: 30
    File System Counters
        FILE: Number of bytes read=150618910
        FILE: Number of bytes written=1487977
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
    Map-Reduce Framework
        Map input records=14111
        Map output records=4029
        Map output bytes=217530
        Map output materialized bytes=225734
        Input split bytes=250
        Combine input records=0
        Combine output records=0
        Reduce input groups=914
        Reduce shuffle bytes=225734
        Reduce input records=4029
        Reduce output records=914
        Spilled Records=8058
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=186
        Total committed heap usage (bytes)=1540882432
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=58303670
    File Output Format Counters
        Bytes Written=46737
```

This might be an running example with the argument:

/cs/studres/CS1003/Practicals/W11/data/1_minute/00.json output00
and it returns the result like above.

Evaluation:

Here is one exception occurs in my program but doesn't affect running:

```
java.io.IOException: HADOOP_HOME or hadoop.home.dir are not set.
    at org.apache.hadoop.util.Shell.checkHadoopHome(Shell.java:326)
    at org.apache.hadoop.util.Shell.<clinit>(Shell.java:351)
    at org.apache.hadoop.util.StringUtils.<clinit>(StringUtils.java:80)
    at org.apache.hadoop.security.SecurityUtil.getAuthenticationMethod(SecurityUtil.java:610)
    at org.apache.hadoop.security.UserGroupInformation.initialize(UserGroupInformation.java:273)
    at org.apache.hadoop.security.UserGroupInformation.ensureInitialized(UserGroupInformation.java:261)
    at org.apache.hadoop.security.UserGroupInformation.loginUserFromSubject(UserGroupInformation.java:791)
    at org.apache.hadoop.security.UserGroupInformation.getLoginUser(UserGroupInformation.java:761)
    at org.apache.hadoop.security.UserGroupInformation.getCurrentUser(UserGroupInformation.java:634)
    at org.apache.hadoop.mapreduce.task.JobContextImpl.<init>(JobContextImpl.java:72)
    at org.apache.hadoop.mapreduce.Job.<init>(Job.java:142)
    at org.apache.hadoop.mapreduce.Job.getInstance(Job.java:185)
    at org.apache.hadoop.mapreduce.Job.getInstance(Job.java:204)
    at W11Practical.main(W11Practical.java:53)
```

I haven't get rid of it and have no idea to solve it. This is the only thing that I failed and others I did should be all fine.

Conclusion:

This program teaches me about how to use hadloop to map the data and reduce them. In this program, I studied about how to map the data and reduce it, I have also reviewed about how to read Json file. I have also learned from readJson file that in order to red the file, Read from top to bottom is neccessary or it may return exception that get in the way of running program.

Extension 1: Most popular tweet:

Design: Initially I want to use another map reduce to do it, however after I read the file I found a good way to do it which is like the following:

```
public class GetMaximumNumber {
    public void read(String inputFile) {
        Map<String, Integer> map = new HashMap<~>();
        try {
            BufferedReader reader = new BufferedReader(new FileReader(inputFile));
            String line;
            while ((line = reader.readLine()) != null) {
                String[] element = line.split( regex: "\t");
//                System.out.println(line);
                map.put(element[0], Integer.parseInt(element[1]));
            }
            int maxValue = Collections.max(map.values());
            for(Map.Entry<String, Integer> entry : map.entrySet()) {
                if(entry.getValue() == maxValue) {
                    System.out.println("\n\n\n\n\t The most popular tweet is:  " + entry.getKey() + "\t" + entry.getValu
                }
            }
        }
        catch (FileNotFoundException e) {
            System.out.println(e.getMessage());
        }
        catch (IOException e) {
            System.out.println(e.getMessage());
        }
```

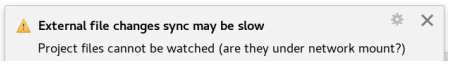
Because that there are only two elements available in the output file, I can use line.split to split the data and then use map to get the maximum data of it.

The drawback of this method is that I can't get the associated users to this tweet.
The result of this shows as follows:

```
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=58303670
    File Output Format Counters
        Bytes Written=46737
18/04/27 17:05:33 DEBUG security.UserGroupInformation: PrivilegedAction as:hl74 (auth:SIMPLE) from:org.apache.hadoop.mapreduce.Job.updateStatus(Job.java:320)


     The most popular tweet is:  "http://du3a.org"  107

Process finished with exit code 0
```



Extension 2 : reorder

This extension uses similar method as the method before. The extension uses buffered reader to read from output file and export it by using print writer and output as the txt file which shows like this:


```
"http://du3a.org"       107
"http://d3waapp.org/"   43
"https://youtu.be/nwmSo7D9S_Y"  39
"http://ghared.com"     29
"http://www.totobet1.com"       28
"http://7asnat.com/"    24
"http://bbc.co.uk/inauguration" 22
"http://prt.nu/0/SEXjapan"      22
"http://ska42.space/a/yE7Sx"    21
"http://dealamoo.com/01/20/huge-anjou-body-scrub-8oz-sale-with-64-off/" 19
"http://playbeach.net/movies/code=ERMTc2MDY15"  19
"http://zad-muslim.com" 19
"https://twitter.com/i/web/status/822519095362920448"   19
"http://prt.nu/0/sexy99"        18
"http://bit.ly/xjQYj7"  17
"http://ncode.syosetu.com/n3699cl/"     16
"http://bit.ly/2cma7La" 15
"http://twipi.org/084015116ca1" 15
"http://Quran.to"       14
"http://bit.ly/1oeDWBG" 14
"http://bit.ly/2i3lgpx?var=1775767525"  14
"http://fb.me/6z5nJ5Yyu"        14
"http://ift.tt/1OrmTa0" 14
"http://ift.tt/2jVTIRd" 14
"http://ind.pn/2iTGP8x" 14
"http://radionightitaly.us"     14
"http://ramen-pasta.com/jp/?p=2088"     14
"http://ver.as/ho_d"    14
"http://www.footballtop.ru/news/dzherrard-o-vozvrashchenii-v-liverpul-takoe-chuvstvo-chto-krug-zamknulsya"      14
"http://www.sexting.hotslut.info/?5isP" 14
"http://www.verseo.com/heated-neck-and-shoulder-scarf.html"     14
"http://youtu.be/QprIRL3RwNI?a" 14
"https://goo.gl/yZdnsU" 14
"https://twitter.com/CloydRivers/status/822234739448905728"     14
"https://twitter.com/ConcaiUrur1990/status/821006099218395136"  14
"https://twitter.com/DiogoRobinson97/status/819251415885479938" 14
"https://twitter.com/Patriota1966/status/822517586252492800"    14
"http://5s.rinode.pw"   13
"http://bit.ly/2azJPIj" 13
"http://buff.ly/2j310nS"        13
"http://dlvr.it/N92pWs" 13
"http://dratef.net/photos/busy-few-days-at-my-sisters-with-the-little-angels-and-sometimes-referred-to-as/"     13
"http://fb.me/8d2ybMwIM"        13
"http://fb.me/Kr0UNi41" 13
"http://ow.ly/1LoI508MJuz"      13
"http://twcm.me/vbmm0"  13
"http://www.2daycellphones.com/US/landing-static/twitter21/?id=http://rover.ebay.com/rover/1/711-53200-19255-0/1?ff3=2&toolid=10039&campid=5337889852&item=152404261989&vectorid=229466&lgeo=1
"http://www.peakfm.co.uk/news/local/chesterfield-man-reported-missing-from-his-home/"   13
"http://y1.treslo.pw"   13
"http://youtu.be/L5NsgRrK6hc?a" 13
"https://cards.twitter.com/cards/18ce54ch3f3/2a0i3"     13
"https://goo.gl/fb/mcNF2y"      13
"https://twitter.com/i/web/status/822519074399916037"   13
"https://twitter.com/i/web/status/822519267363094528"   13
"https://youtu.be/kjqzI8zNkec"  13
```

The program uses ArrayList and uses collections.sort, which compare is overrided in Tweets class implement compareable package. I got this idea from previous practical and uses it this time.

These are all extensions.