Lab Assignment 1

Author: Vendramin Nicolò

Imports

```java
import java.io.IOException;
import java.util.StringTokenizer;

import java.io.IOException;
import java.util.Map;
import java.util.TreeMap;
import java.util.HashMap;

import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
```

Main

```java
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "top ten");
        job.setNumReduceTasks(1);
        job.setJarByClass(TopTen.class);
        job.setMapperClass(TopTenMapper.class);
        job.setCombinerClass(TopTenReducer.class);
        job.setReducerClass(TopTenReducer.class);
        job.setOutputKeyClass(NullWritable.class);
        job.setOutputValueClass(Text.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

Mapper

```java
public static Map<String, String> transformXmlToMap(String xml) {
    Map<String, String> map = new HashMap<String, String>();
    try {
        String[] tokens = xml.trim().substring(5, xml.trim().length() - 3)
                .split("\"");

        for (int i = 0; i < tokens.length - 1; i += 2) {
            String key = tokens[i].trim();
            String val = tokens[i + 1];

            map.put(key.substring(0, key.length() - 1), val);
        }
    } catch (StringIndexOutOfBoundsException e) {
        System.err.println(xml);
    }

    return map;
}

public static class TopTenMapper extends Mapper<Object, Text, NullWritable, Text> {
    // Stores a map of user reputation to the record
    private TreeMap<Integer, Text> repToRecordMap = new TreeMap<Integer, Text>();

    public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
        Map<String, String> parsed = transformXmlToMap(value.toString());
        String userId = parsed.get("Id");
        String reputation = parsed.get("Reputation");

        // check that this row contains user data
        if(userId == null || reputation == null){
            return;
        }

        repToRecordMap.put(Integer.parseInt(reputation), new Text(value));

        // If we have more than ten records, remove the one with the lowest reputation.
        if (repToRecordMap.size() > 10) {
            repToRecordMap.remove(repToRecordMap.firstKey());
        }
    }

    protected void cleanup(Context context) throws IOException, InterruptedException {
        // Output our ten records to the reducers with a null key
        for (Text t : repToRecordMap.values()) {
            context.write(NullWritable.get(), t);
        }
    }
}
```

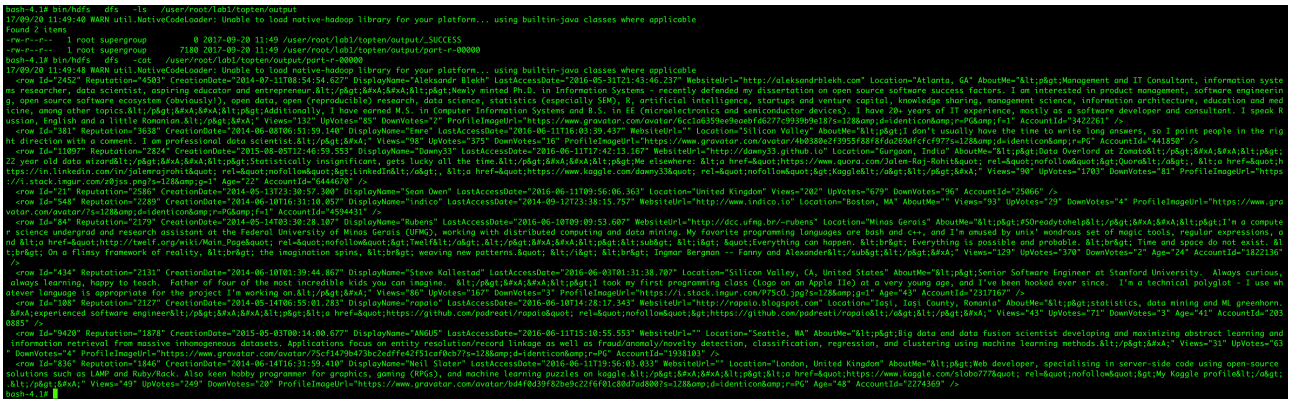
Reducer

```
public static class TopTenReducer extends Reducer<NullWritable, Text, NullWritable, Text> {
    // Stores a map of user reputation to the record
    // Overloads the comparator to order the reputations in descending order
    private TreeMap<Integer, Text> repToRecordMap = new TreeMap<Integer, Text>();

    public void reduce(NullWritable key, Iterable<Text> values, Context context) throws IOException, InterruptedException {
        for (Text value : values) {
            Map<String, String> parsed = transformXmlToMap(value.toString());
            repToRecordMap.put(Integer.parseInt(parsed.get("Reputation")), new Text(value));

            // If we have more than ten records, remove the one with the lowest reputation
            if (repToRecordMap.size() > 10) {
                repToRecordMap.remove(repToRecordMap.firstKey());
            }
        }

        // Sort in descending order
        for (Text t : repToRecordMap.descendingMap().values()) {
            // Output our ten records to the file system with a null key
            context.write(NullWritable.get(), t);
        }
    }
}
```

Output (1)

```
bash-4.1# bin/hadoop   jar   work/lab1/topten/topten.jar   lab1code.TopTen /user/root/lab1/topten/input   /user/root/lab1/topten/output
17/09/20 11:48:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/09/20 11:48:57 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
17/09/20 11:48:57 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
17/09/20 11:48:57 INFO input.FileInputFormat: Total input paths to process : 1
17/09/20 11:48:58 INFO mapreduce.JobSubmitter: number of splits:1
17/09/20 11:48:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1505902951661_0006
17/09/20 11:48:58 INFO impl.YarnClientImpl: Submitted application application_1505902951661_0006
17/09/20 11:48:58 INFO mapreduce.Job: The url to track the job: http://ca2689501238:8088/proxy/application_1505902951661_0006/
17/09/20 11:48:58 INFO mapreduce.Job: Running job: job_1505902951661_0006
17/09/20 11:49:03 INFO mapreduce.Job: Job job_1505902951661_0006 running in uber mode : false
17/09/20 11:49:03 INFO mapreduce.Job:  map 0% reduce 0%
17/09/20 11:49:08 INFO mapreduce.Job:  map 100% reduce 0%
17/09/20 11:49:12 INFO mapreduce.Job:  map 100% reduce 100%
17/09/20 11:49:13 INFO mapreduce.Job: Job job_1505902951661_0006 completed successfully
17/09/20 11:49:13 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=7244
                FILE: Number of bytes written=245903
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5856430
                HDFS: Number of bytes written=7180
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=2501
                Total time spent by all reduces in occupied slots (ms)=2254
                Total time spent by all map tasks (ms)=2501
                Total time spent by all reduce tasks (ms)=2254
                Total vcore-seconds taken by all map tasks=2501
                Total vcore-seconds taken by all reduce tasks=2254
                Total megabyte-seconds taken by all map tasks=2561024
                Total megabyte-seconds taken by all reduce tasks=2308096
        Map-Reduce Framework
                Map input records=13995
                Map output records=10
                Map output bytes=7199
                Map output materialized bytes=7244
                Input split bytes=127
                Combine input records=10
                Combine output records=10
                Reduce input groups=1
                Reduce shuffle bytes=7244
                Reduce input records=10
                Reduce output records=10
                Spilled Records=20
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=146
                CPU time spent (ms)=1720
                Physical memory (bytes) snapshot=448331776
                Virtual memory (bytes) snapshot=1466036224
                Total committed heap usage (bytes)=288358400
        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=5856303
        File Output Format Counters
                Bytes Written=7180
```

Output (2)


bash-4.1# bin/hdfs dfs -ls /user/root/lab1/topten/output
17/09/20 11:49:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 root supergroup 0 2017-09-20 11:49 /user/root/lab1/topten/output/_SUCCESS
-rw-r--r-- 1 root supergroup 7180 2017-09-20 11:49 /user/root/lab1/topten/output/part-r-00000
bash-4.1# bin/hdfs dfs -cat /user/root/lab1/topten/output/part-r-00000
17/09/20 11:49:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
<row Id="2452" Reputation="4503" CreationDate="2014-07-11T08:54:54.627" DisplayName="Aleksandr Blekh" LastAccessDate="2016-05-31T21:43:46.237" WebsiteUrl="http://aleksandrblekh.com" Location="Atlanta, GA" AboutMe="<p>Management and IT Consultant, information systems researcher, data scientist, aspiring educator and entrepreneur.</p>

<p>Newly minted Ph.D. in Information Systems - recently defended my dissertation on open source software success factors. I am interested in product management, software engineering, open source software ecosystem (obviously), open data, open (reproducible) research, data science, statistics (especially SEM), R, artificial intelligence, startups and venture capital, knowledge sharing, management science, information architecture, education and medicine, among other topics.</p>

<p>Additionally, I have earned M.S. in Computer Information Systems and B.S. in EE (microelectronics and semiconductor devices). I have 20+ years of IT experience, mostly as a software developer and consultant. I speak Russian, English and a little Romanian.</p>
" Views="132" UpVotes="85" DownVotes="2" ProfileImageUrl="https://www.gravatar.com/avatar/6cc1a6359ee9aa6bfd6277c9939b9e187s=128&d=identicon&r=PG&f=1" AccountId="3422261" />
<row Id="381" Reputation="3638" CreationDate="2014-06-05T06:51:59.140" DisplayName="Emre" LastAccessDate="2016-06-11T16:03:39.437" WebsiteUrl="" Location="Silicon Valley" AboutMe="<p>I don't usually have the time to write long answers, so I point people in the right direction with a comment. I am professional data scientist.</p>
" Views="98" UpVotes="375" DownVotes="16" ProfileImageUrl="https://www.gravatar.com/avatar/4b0380e2f3955f88f8fda269dfcfc977s=128&d=identicon&r=PG" AccountId="441850" />
<row Id="11097" Reputation="2824" CreationDate="2015-08-05T12:46:59.553" DisplayName="Dawny33" LastAccessDate="2016-06-11T17:42:13.167" WebsiteUrl="http://dawny33.github.io" Location="Gurgaon, India" AboutMe="<p>Data Overlord at Zomato</p>

<p>22 year old data wizard</p>

<p>Statistically insignificant, gets lucky all the time.</p>

<p>Me elsewhere: Quora, LinkedIn, Kaggle</p>
" Views="90" UpVotes="1703" DownVotes="81" ProfileImageUrl="https://i.stack.imgur.com/z0jss.png?s=128&g=1" Age="22" AccountId="6444670" />
<row Id="21" Reputation="2586" CreationDate="2014-05-13T23:30:57.300" DisplayName="Sean Owen" LastAccessDate="2016-06-11T09:56:06.363" Location="United Kingdom" Views="202" UpVotes="679" DownVotes="96" AccountId="25066" />
<row Id="548" Reputation="2289" CreationDate="2014-06-10T16:31:10.057" DisplayName="indico" LastAccessDate="2014-09-12T23:38:15.757" WebsiteUrl="http://www.indico.io" Location="Boston, MA" AboutMe="" Views="93" UpVotes="29" DownVotes="4" ProfileImageUrl="https://www.gravatar.com/avatar/5s=128&d=identicon&r=PG&f=1" AccountId="4594431" />
<row Id="84" Reputation="2179" CreationDate="2014-05-14T03:30:28.107" DisplayName="Rubens" LastAccessDate="2016-06-10T09:09:53.607" WebsiteUrl="http://dcc.ufmg.br/~rubens" Location="Minas Gerais" AboutMe="<p>$OreadytohelpI</p>

<p>I'm a computer science undergrad and research assistant at the Federal University of Minas Gerais (UFMG), working with distributed computing and data mining. My favorite programming languages are bash and c++, and I'm amused by unix' wondrous set of magic tools, regular expressions, and Twelf.</p>

<p>_{<i> "Everything can happen.
 Everything is possible and probable.
 Time and space do not exist.
 On a flimsy framework of reality,
 the imagination spins,
 weaving new patterns." </i>
 Ingmar Bergman -- Fanny and Alexander}</p>
" Views="129" UpVotes="370" DownVotes="2" Age="24" AccountId="1822136" />
<row Id="434" Reputation="2131" CreationDate="2014-06-10T01:39:44.867" DisplayName="Steve Kallestad" LastAccessDate="2016-06-03T01:31:38.707" Location="Silicon Valley, CA, United States" AboutMe="<p>Senior Software Engineer at Stanford University. Always curious, always learning, happy to teach. Father of four of the most incredible kids you can imagine. </p>

<p>I took my first programming class (logo on an Apple IIe) at a very young age, and I've been hooked ever since. I'm a technical polyglot - I use whatever language is appropriate for the project I'm working on.</p>
" Views="86" UpVotes="167" DownVotes="3" ProfileImageUrl="https://i.stack.imgur.com/P75cO.jpg?s=128&g=1" Age="43" AccountId="2317167" />
<row Id="108" Reputation="2127" CreationDate="2014-05-14T06:55:01.283" DisplayName="rapaio" LastAccessDate="2016-06-10T14:28:17.343" WebsiteUrl="http://rapaio.blogspot.com" Location="Iasi, Iasi County, Romania" AboutMe="<p>statistics, data mining and ML greenhorn.
experienced software engineer</p>

<p>https://github.com/padreati/rapaio</p>
" Views="43" UpVotes="71" DownVotes="3" Age="41" AccountId="2030885" />
<row Id="9420" Reputation="1878" CreationDate="2015-05-03T00:14:00.677" DisplayName="ANGUS" LastAccessDate="2016-06-11T15:10:55.553" WebsiteUrl="" Location="Seattle, WA" AboutMe="<p>Big data and data fusion scientist developing and maximizing abstract learning and information retrieval from massive inhomogeneous datasets. Applications focus on entity resolution/record linkage as well as fraud/anomaly/novelty detection, classification, regression, and clustering using machine learning methods.</p>
" Views="31" UpVotes="63" DownVotes="4" ProfileImageUrl="https://www.gravatar.com/avatar/75cf1479b473bc2edfe42f51cef0cb77s=128&d=identicon&r=PG" AccountId="1936101" />
<row Id="836" Reputation="1846" CreationDate="2014-06-14T16:31:59.410" DisplayName="Neil Slater" LastAccessDate="2016-06-11T19:56:03.033" WebsiteUrl="" Location="London, United Kingdom" AboutMe="<p>Web developer, specialising in server-side code using open-source solutions such as LAMP and Ruby/Rack. Also keen hobby programmer for graphics, gaming (RPGs), and machine learning puzzles on kaggle.</p>

<p>My Kaggle profile</p>
" Views="49" UpVotes="249" DownVotes="20" ProfileImageUrl="https://www.gravatar.com/avatar/bd4f0d39f82be9c22f6f01c80d7ad800?s=128&d=identicon&r=PG" Age="48" AccountId="2274369" />
bash-4.1#