CSE587 Data Intensive Computing

Project 2: Part 2

Question 1

Taking 12 hours in a day and 60 hours in a week give the percentage of hours every room is not in use in a week for Spring 2016.

To solve this problem we have a chain of two MapReduce. First MR gives the output as <Room Day Time, a number>. The number represents the repetition of room at a particular time in the data to show the enrolment of graduate and undergraduate students separately. In the second MR we find the total number of hours the room is in use and subtracting this sum from 60 to get the number of hours room is not in use. Then by the formula ((not in use hours * 100) / 60) we get the percentage of hours a room is not in use. This is our final output.

Sample output from first MR:

4240RL ARR	A	Before 8:00AM	11
4240RL ARR	F	10:00AM - 10:59AM	0
4240RL ARR	F	11:00AM - 11:59AM	0
4240RL ARR	F	12:00PM - 12:59PM	0
4240RL ARR	F	1:00PM - 1:59PM	0
4240RL ARR	F	2:00PM - 2:59PM	0
4240RL ARR	F	3:00PM - 3:59PM	0

Sample output from second MR:

Abbott 165 95
Abbott B15 98
Ach_A 02 98
Alfier 102 50
Alfier 103 45
Alfier 104 50
Alfier 205 98

Question 2

Which all courses don't have an exam each semester?

For this problem we are using three MRs. First MR runs on data bina_classschedule.csv and find out all the courses each semester. Second MR runs on data bina_examschedule.tsv and find out all the courses which have exams. The output of both the MRs is of the form <Sem Course, 1>. The third MR combines the result of the two and finds the difference outputting the courses which don't have an exam.

Sample output from first MR:

Fall 2011_(MPH)Integrative Project 1
Fall 2011_19c Europe 1
Fall 2011_1st Yr-1st Sem Am Sig Lan 1
Fall 2011_1st Yr-1st Sem Arabic 1
Fall 2011_1st Yr-1st Sem Chinese 1
Fall 2011_1st Yr-1st Sem Greek 1
Fall 2011_1st Yr-1st Sem Hindi 1
Fall 2011_1st Yr-1st Sem Japanese 1
Fall 2011_1st Yr-1st Sem Korean 1
Fall 2011_1st Yr-1st Sem Ukrainian 1
Fall 2011_1st Yr-2nd Sem Chinese 1

Sample output from second MR:

Fall 2011_1st Yr-1st Sem Japanese 1
Fall 2011_1st Yr-2nd Sem Chinese 1
Fall 2011_20 Century Philosophy 1
Fall 2011_20c Italian Literature 1
Fall 2011_2nd Yr-1st Sem Chinese 1
Fall 2011_2nd Yr-1st Sem Japanese 1
Fall 2011_3rd Yr, 1st Sem Japanese 1
Fall 2011_3rd Yr-1st Sem Chinese 1

Sample output from third MR:

Fall 2011_(MPH)Integrative Project 1
Fall 2011_19c Europe 1
Fall 2011_1st Yr-1st Sem Am Sig Lan 1
Fall 2011_1st Yr-1st Sem Arabic 1
Fall 2011_1st Yr-1st Sem Chinese 1
Fall 2011_1st Yr-1st Sem Greek 1
Fall 2011_1st Yr-1st Sem Hindi 1

Question 3

Courses ordered in descending order of their enrolment average, i.e. courses sorted in order of popularity with most popular on top?

This problem is solved by using three MRs on the data bina_classschedule.csv. The first MR gives the output <sem_course, total enrolment>. This output goes as input in second MR where we find the average enrolment over the years giving output as <course, average>. At the third MR this average is sorted in the descending order by making the enrolment average as key and overriding the comparator. Output of the third MR is <enrolment average, course> with enrolment average in the descending order.

Sample output from first MR:

Fall 1931_Economics Topic 1
Fall 1931_Elem French 1st Semester 1
Fall 1931_Hygiene 1
Fall 1931_Intro to Philosophy 1
Fall 1931_Introductory Psychology 1
Fall 1931_Writing 1 1
Fall 1932_Designer Genes 1
Fall 1932_Psychology Topic 1

Sample output from second MR:

(MPH)Integrative Project 3
100yrs Pol His&Art&Soc'ty 10
14th Amendment 0
16c & 17c Spanish Literat 9
16thc Drama: Pre-Lopistas 2
17c Literat & Philosophy 1
17c Literature Studies 4

Sample output from third MR:

2224 General Chemistry Lab-Rec
1852 World Civilization 1
1726 World Civilization 2
1047 Organic Chemistry LAB-REC
958 Gen Chem for Engineers LAB-REC
950 General Chemistry
873 Intro to Mgmt 1
809 Evolutionary Biology

Question 4

Predict the enrolment of next five years of every department taking the enrolment trend of ten years.

To predict the enrolment of next five years we first got the total enrolment of each department of last ten years using one MR. The output of this MR is of the form <year dept, total enrol> with value of year varying ten years. In the second MR by averaging and finding the difference of every point from average we found the slope and intercept of the line and thus applied linear regression. The output of this MR is of the form <year dept, enrol> with value of next five years.

Sample output from first MR:

2006 CE 1371
2007 CE 1386
2008 CE 1605
2009 CE 1886
2010 CE 2027
2011 CE 2230
2012 CE 2572
2013 CE 2863
2014 CE 3017
2015 CE 3084

Sample output from second MR:

CE 2016 3279
CE 2017 3475
CE 2018 3670
CE 2019 3866
CE 2020 4061

Question 5

Trend of number of courses over the years.

In this problem we are showing the trend of number of courses over the years with all courses and courses having enrolment above 30 separately. In the first MR we separated the count of all courses and courses above 30 enrolment. After getting the count we summed them up in the second MR. For every key i.e. semester we get two values which are outputted separately in the third MR. Observing these values we can see the gradual increase in the number of courses every year.

Sample output from first MR:

Fall 1995#17c Literature Studies#counter# 1
Fall 1995#17c Literature Studies#enrolabove30# 0
Fall 1995#19c American Literature#counter# 2
Fall 1995#19c American Literature#enrolabove30# 0
Fall 1995#19c Art#counter# 1
Fall 1995#19c Art#enrolabove30# 0
Fall 1995#19th Century Drama#counter# 2
Fall 1995#19th Century Drama#enrolabove30# 0
Fall 1995#19th-C British Poetry#counter# 1

Sample output from second MR:

Fall 1995#counter# 2231
Fall 1995#enrolabove30# 615
Fall 1996#counter# 2320
Fall 1996#enrolabove30# 607
Fall 1997#counter# 2322
Fall 1997#enrolabove30# 588

Sample output from third MR:

Fall 1995 2231 615
Fall 1996 2320 607
Fall 1997 2322 588
Fall 1998 2350 576
Fall 1999 2350 606
Fall 2000 2280 649
Fall 2001 2830 651
Fall 2002 2888 703

Division of work

- Come up with questions and logic: Done by both Priyanka Singh and Sahil Dureja
- Coding: 3 problems by Sahil Dureja and 2 problems by Priyanka Singh
- Report: Done by Priyanka Singh