**Lab 4 Submission**

**Data Wrangler Script and Screenshot: CMSC**

**Data Wrangler Script and Screenshot: World Cup 1**

**Data Wrangler Script and Screenshot: World Cup 2**

**UNIX Tools Command: CMSC**

**cat cmsc.txt | sed 's/[()]//g' | awk '/^(CMSC[0-9]{3})/ {c=$0;}**

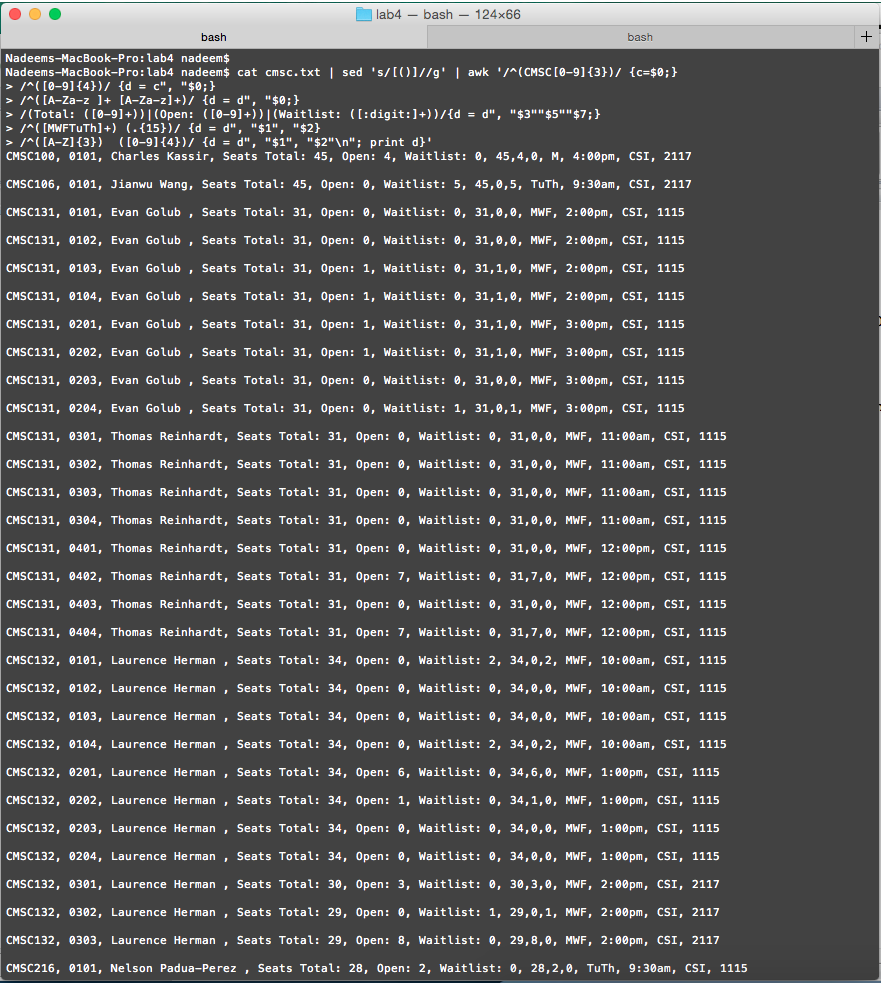
**/^([0-9]{4})/ {d = c", "$0;}**

**/^([A-Za-z ]+ [A-Za-z]+)/ {d = d", "$0;}**

**/(Total: ([0-9]+))|(Open: ([0-9]+))|(Waitlist: ([:digit:]+))/{d = d", "$3""$5""$7;}**

**/^([MWFTuTh]+) (.{15})/ {d = d", "$1", "$2}**

**/^([A-Z]{3}) ([0-9]{4})/ {d = d", "$1", "$2"\n"; print d}'**

****

**UNIX Tools Command: World Cup 1**

**cat worldcup.txt | awk '{gsub(".style=\".+\"", "");print}'|**

**sed 's/|{{fb|\('[A-Z]'\{3\}\)}}/\1/' | awk '{gsub("\\[|\\]", "");print}' | awk '{gsub("FIFA World Cup\\|[0-9]{4}","");print}' | awk '{gsub("#..\\\*","");print}' | awk '{gsub("(\\|)\*[0-9]{1,}(\\|)+[0-9]{1,}(\\|)+[0-9]{1,}","");print}' | awk '{gsub("(\\|)|-","");print}' | awk '{gsub("style=whitespace:nowrapTCH<sup>","");print}' | awk '{gsub("<sup>|#.\*","");print}' | awk '{gsub("\\(|\\)|,|\\}","");print}' |**

**awk '{gsub("([0-9])\*\\{\\{sort dash","");print}' |**

**awk 'BEGIN {ctr=0} /^([A-Z]{3})/ {c=$0; d=$0;}**

**/^([0-9]) ([0-9]{4})+/ {ctr = ctr+1;}**

**{ i = 1**

**x = 2**

**l = $1**

**if(l ~ /[A-Z]{3}/) {**

**ctr = 0**

**d = $1**

**}else if(l ~ /align=center/) {**

**ctr++**

**}**

**else {**

**for(i = 1; i <= l; i++) {**

**d = c", "$(i+1)", " ctr;**

**print d;**

**}**

**d = c;**

**}**

**} END {d=""}'**

**Python Script: CMSC**

**import re**

**file\_name = "cmsc.txt"**

**with open(file\_name) as f:**

**lines = f.readlines()**

**num\_lines = len(lines)**

**i = 0**

**while i < num\_lines:**

**l = lines[i]**

**if re.match('^CMSC[0-9]{3}', l):**

**class\_name = l.rstrip()**

**info = class\_name + ", "**

**while(i + 1 < num\_lines and re.match('^[0-9]{4}', lines[i+1].rstrip())):**

**section = lines[i+1].rstrip();**

**prof\_name = lines[i+2].rstrip()**

**seats = lines[i+3].rstrip()**

**days\_time = lines[i+4].rstrip().split(" ")**

**days = days\_time[0]**

**days\_time = days\_time[1] + " - " + days\_time[3]**

**building, room = lines[i+5].rstrip().split(" ")**

**m\_seats = re.match('\d+',seats)**

**t, o, w = (seats[7:-1].split(", "))**

**t, o, w = t.split(": ")[1], o.split(": ")[1], w.split(": ")[1]**

**print(class\_name + ", " + section + ", " + \**

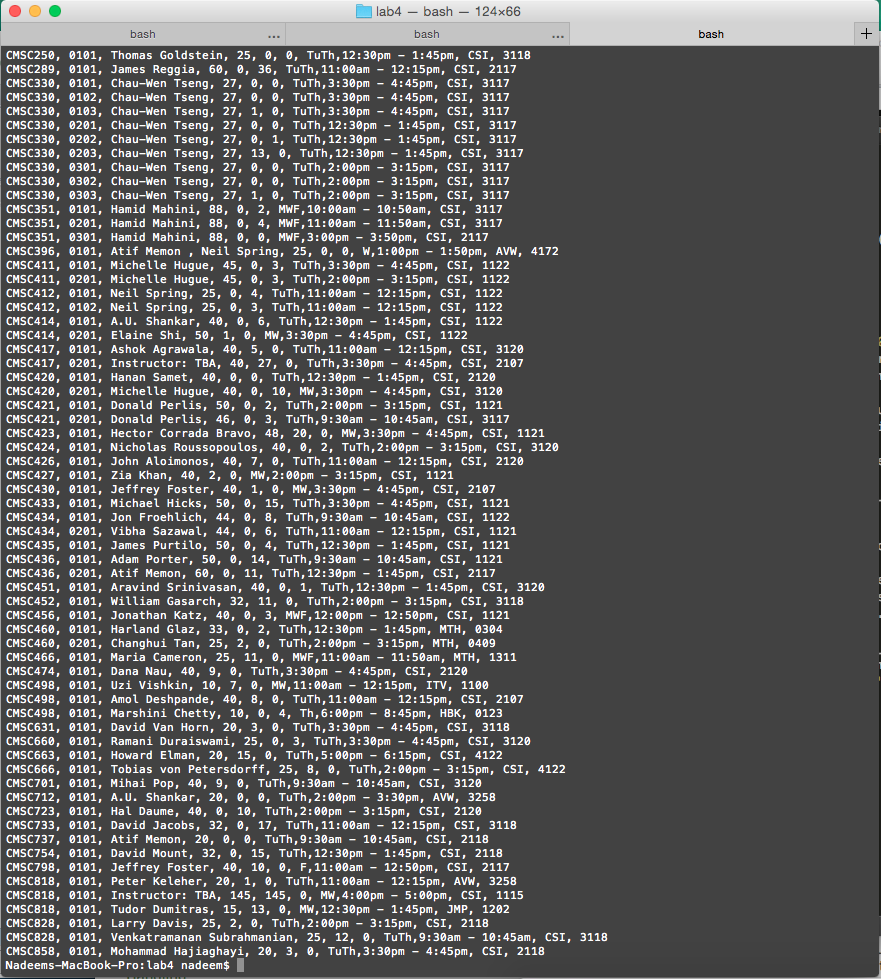
**prof\_name + ", " + t + ", " + o + ", " + w + ", " + \**

**days + "," + days\_time + ", " + building + ", " + room)**

**i = i + 5**

**else:**

**i += 1**

****

**Python Script: World Cup 1**

**import re**

**import fileinput**

**file\_name = "worldcup.txt"**

**import fileinput**

**lines = []**

**for line in fileinput.input(file\_name, inplace=False):**

**line = re.sub('(\\|)+(\d)+(\\|)+.+', "", line.rstrip())**

**line = re.sub('\\||{|}|fb|style=\".\*\"', "", line.rstrip())**

**line = re.sub('\\[|\\]', "", line.rstrip())**

**line = re.sub('FIFA World Cup(\d){4}(#\d\\\*)\*', "", line.rstrip())**

**line = re.sub('#\d\\^', "", line.rstrip())**

**line = re.sub('(^style=.+)|-|<sup>...</sup>|\\(|\\)', "", line.rstrip())**

**line = re.sub(' , ', " ", line.rstrip())**

**lines.append(line)**

**num\_lines = len(lines)**

**i = 0**

**while(not re.match('[A-Z]{3}', lines[i])):**

**i += 1**

**while(i < num\_lines and re.match('[A-Z]{3}', lines[i])):**

**country = lines[i].rstrip()**

**x = [1,2,3,4]**

**for placing in x:**

**l2 = lines[i+placing]**

**if(not re.match('align=centersort dash', l2)):**

**for year in l2.split(" "):**

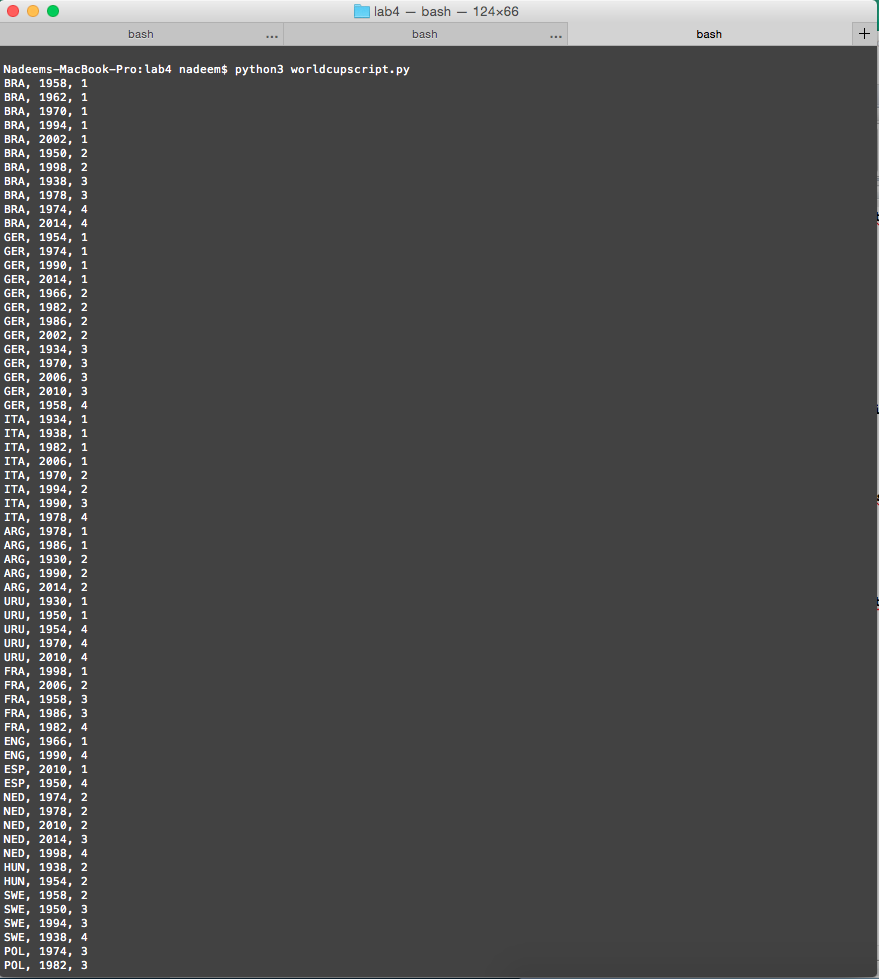
**if len(year) > 1:**

**print (country + ", " + year + ", " + str(placing))**

**i += 5**

**while(i < num\_lines and not re.match('[A-Z]{3}', lines[i])):**

**i += 1**

****

**Python Script: World Cup 2**