Homework 4 Instructions
Due Saturday, 24th Feb by 11:59 PM

1. Short answer – give a single regular expression pattern that will match to the following conditions. To test, or to try patterns out, you can use code similar to this:

```
import re

a = "Jackson or Johnson or…"
m = re.findall("Jack\w+", a)

if m:
   print("It matched!")
   print(m)
```

   a) Match to Jackson and to Johnson

   b) Match to Jackson and Johnson, but not to Jason

   c) Match to an id of the form NM_123456.2 where the 2 letters at the beginning can be any combination of uppercase letters, and the numbers can be any set of numbers. There is always an underscore between the letters and the numbers. There is always a .2 (period number, and can be any number) at the end.

   d) Match to an id of the form NM_123456.7 where the letters at the beginning can be any combination of two upper or lowercase letters, the numbers can be any combination of numbers, and the .7 (period followed by any number) at the end may or may not be there.

   e) A standard 10-digit phone number with some using – as a separator (e.g. 123-456-7890) and some will have a dot (.) as a separator (e.g. 123.456.7890).

2. Short answer – write a single re.findall command to capture the data asked for. Samples of input data are given. Assume the input data are being read in one line at a time, into a variable called "linein". Pay attention to details, details matter. Your answers should be like:

   m = re.findall("regex pattern", linein)

a) Capture the components of the time separately as hours, minutes, and seconds

Date: Mon, 14 May 2001 19:36:00 (PDT)
Date: Fri, 7 Aug 2000 12:37:00 (PDT)
Date: Wed, 11 Jan 2001 03:16:00 (PDT)

b) Capture the information between the pairs of tags (not including the tags). There are 2 pairs of tags per line. A pair of tags is like this: <a>

hello hello 123 stuff to ignore here 123412bhje
<a>what??? stuff to ignore here asd13asf
who! Hooooo! stuff to ignore here df7887a

c) Capture the information associated with each tag, but don't capture the tag or the equals sign. A tag is like this: /xyz=

/gene=apoE /defin=apolipoprotein E
/gene=BIN1 /defin=bridging integrator 1
/gene=CLU /defin=clusterin

3. Consider the employee database with two relations

employee (employee name, street, city)
works (employee name, company name, salary)

a) Write a function called avg salary that takes a company name as an argument and finds the average salary of employees at that company.

b) Write an SQL statement, using that function, to find companies whose employees earn a higher salary, on average, than the average salary at "First Bank".

4. Write python code to output each instructor's name, the names of the classes they teach, and the names students in each of those classes from the university database. Sort the output by instructor name, class name, student name.

a. Do this using a single big query
b. Do this using multiple queries