# Homework 1: Math 495, ISU, F20

**Due on 8/28 in canvas**

---

**Instructions: Please do HW1 in Jupyter Notebook using Python. Submit an HTML file and ipynb file following the link in Modules, and then HW1, and finally clicking the submit button to access files on your computer. When working on the HWs, please do not display unnecessary data unless it is asked. Each problem is with 20 points.**

1) The longest known word in the English language is given by pneumonoultramicroscopicsilicovolcanoconiosisis. You can google to find out what it means. Read more about dictionaries and sets in python.
   a. Using python in Jupyter notebook, make a list of all the letters in this word.
   b. Change the list in (a) to a set to find out a unique set of alphabets used in this word.
   c. Considering a set U of 26 English alphabets as a Universal set, find the complement of the set in (b) using python.
   d. Create a dictionary with these unique letters in the set (b) as keys and the number of times these letters appear in the word as values. Your dictionary should look like $D = \{a : 2, c : 6, \cdots\cdots\}$ as a appears 2 times, and c appears 6 times in the original word.
2) Another word, supercalifragilisticexpialidocious, has 34 letters. Find a set B of unique letters in this word. Let's denote a set you found in problem 1(b) by A. Do the following using python in Jupyter notebook.
   a. Find the union and intersection of A and B.
   b. Find $A - B, B - A$ and $A \triangle B$.
   c. Verify and find out if the statement $A - B = A \cap B^c$ or $A - B = A^c \cap B$ is true.
   d. Verify both De Morgan's law for set theory for A and B.

3) Upload the .csv file titanic.csv that came with this HW in the notebook by following instructions here and name it as titanic. Make sure to import panda as pd and numpy as np. Read the details about the titanic data here. Answer the following questions. Learn more about the Boolean mask about filtering data.

   a. What happens If you type isna(titanic.cabin).any() and run it? Explain how this is related to either the disjunction or the conjunction in logic lecture we did.

b. How many passengers who paid the minimum fare survived? Hint- Statement p: Passenger survived. Statement q: Passenger paid the minimum fare.
c. What was the name of the female who was traveling in the first class from Cherbourg to Cooperstown, NY, and was 18 or younger? (full credit will be given to a solution that combines multiple statements using and/or and is a one-line code.)
d. Did anybody who didn't pay any fare, had no cabin and didn't have a destination survive?

4) Continuing with the titanic data set, answer the following questions.

a. Filter the titanic data with the condition that $|fare - mean(fare)| < 0.5$ and get a subset of the titanic data.
b. Excluding the NaN values for the variable ticket, find out how many tickets were used by more than a passenger?
c. Create a new column with a column name "luck" in the titanic data set with the following rules. Let's call a passenger "unlucky" if the passenger was in the first-class, paid more than the average fare, and did not survive. On the contrary, a passenger is "lucky" if the passenger was not in the first class, paid less or equal to the average fare, and survived. Let's call everybody else "averageluck". Show the head of the data with a visible luck column.
d. Find the number of lucky, unlucky, and passengers with average luck using a groupby function.

5) Write two data frames you see in the picture below in Jupyter notebook. You will have to manually type these and make two data frames DF1 and DF2.

| | DF1 | | | | DF2 | | |
|---|---|---|---|---|---|---|---|
| ID | Name | City | Age | ID | Name | City | GPA |
| s01 | Julie | Des Moines | 25 | s01 | Julie | Des Moines | 3.25 |
| s05 | Carl | Iowa city | 34 | s08 | David | Minneapolis | 3.4 |
| s06 | Luke | Chicago | 29 | s03 | Bradley | Tempe | 3.3 |
| s08 | David | Minneapolis | 45 | s04 | Nina | Pittsburg | 4 |

a. Create a new data frame using the merge function with left join on ID. Explain which set operation this join corresponds to. You will see that there are some NaN values in the new data frame. Explain why? Read more about joins and concatenation here.
b. Do the same as in (a) with right join.
c. Do the same as in (a) with the inner join. Why are there no NaNs here?
d. Do the same as in (a) with an outer join.