CHEM452/CHBE413

Chemical Data Science and Engineering

Homework 1

Due Date: September 4th, 2025

Please complete and submit this homework as a Jupyter notebook. *For every plot your axes must be labeled, and if there are multiple data sets on the same plot, a legend must be provided!* **If you want some extra practice prior to starting the coding assignment, I would suggest doing the assigned readings in PPNM as well as** “InfoTutorialforPythonNumpy.ipynb” provided on the course website.

1. **(5 points)****What brings you to class? Take some time to reflect on what you hope to gain from this course.** It can be as simple as “I want to practice programming,” or as specific as, “I want to learn machine learning so I can apply it to X problem because I’m passionate about Y.” In a few sentences, please detail some of your specific interests and motivations as it pertains to this class. This will help us shape the course to further suit individual interests and needs.
2. **Reading, Writing, and Manipulating Chemical Data with Python. (**dataset:*ChemicalNames.txt)****.*** “*ChemicalNames.txt*” is a tab-delimited file containing the names, molecular weights (g/mol), and chemical formulas of many compounds. Here, we will practice some basic Python scripting necessary for formatting and sorting this chemical data into usable formats.
3. **(3 points)** Write a function called ReadToDict() that takes as input the filename you would like to read and outputs a list containing two dictionaries: one named MWDict (keys: molecule names, values: molecular weights) and FormulaDict (keys: molecule names, values: molecular formulas). Then use this function to Read the content of “ChemicalNames.txt” into two dictionaries. *Hint: Use split(‘\t’) to help parse the tab-delimited strings in the .txt file.*
4. **(2 points)** Write a function called ConvertMW() that will convert a molecular weight value in g/mol to kiloDaltons (kDa). Use this function to convert all the molecular weights in MWDict to kDa and write out a new file with the same format as ChemicalName.txt, but with the units of kDa. Name this file “*ChemicalNames\_kDa.txt*”.
5. **(2 points)** Use a for loop and conditional statements to filter through your FormulaDict and identify only acids containing the element sulfur. Write out a file called “*ChemicalNames\_kDa\_Sulfur\_acids.txt*” that is the same in structure as “*ChemicalNames\_kDa.txt*” but only contains Sulfur containing acids.
6. **(1 point)** Convert your Sulfur containing acid compounds into a numpy compatible array and use np.mean() and np.std() to compute the average molecular weight and associated standard deviation for Sulfur-containing acids.
7. **(2 points)** Using matplotlib, make a histogram of the molecular weights (in kDa) of the full dataset. Title your plot appropriately and provide a legend and axes labels. It will probably be helpful to modify the ‘bins’ attribute of plt.hist() to make the spread of your plot more interpretable.