MSE 5050/7050 Materials Informatics

Homework Assignment #1

**Due on Canvas on Thursday Jan 27th by 10:45am**

**1.** Clearly, one of the big challenges for materials informatics is the lack of materials property databases. However, as we discussed, there have also been exponential growth in the number of materials publications with no sign of slowing. Therefore, an enormous amount of materials data is *available* for data mining. For this problem, the class will work collectively to assemble a materials property database from scratch! We will be using this database for several follow up homeworks.   
  
Go to the [following link](https://docs.google.com/spreadsheets/d/1Jd77VqQ4hUk3gKVYxv71QjRoFd7JXq8kF_BpooYhjNk/edit?usp=sharing) where you will find a google spreadsheet. The spreadsheet has a number of different columns which you will need to populate by extracting the data from a paper (DOI link provided). You should be able to access the paper from school wifi or it can be accessed offline via [~~sci-hub.se~~](https://sci-hub.se/)… err, I mean via [marriot library off-campus access](https://www.lib.utah.edu/help/off-campus.php). There are columns for the material itself (formula, family, final form, how it was made), as well as material property (electrical conductivity, Seebeck coefficient, and thermal conductivity). When you read the data for materials properties it is critical that you read the data as accurately as possible by using software such as [webplotdigitizer](https://automeris.io/WebPlotDigitizer/). If you haven’t used this software before, I’ve made [a slick YouTube tutorial](https://youtu.be/Mv5nqAPCKA4) you can refer to and I’ll be happy to do demonstrations during office hours.

Finally, it’s critical that we collect this data as accurately as possible since it will serve as the training data for future homeworks. I know that not everyone in this class is a subject matter expert in thermoelectrics, so **please** feel free to ask for help when interpreting the papers you are assigned to data mine.

**2.** Fortunately, some materials databases are starting to take shape. For this problem, I want you to download pymatgen and learn how to use the powerful MPRester API. Step 1: Download [pymatgen](https://pymatgen.org/installation.html). If you are having problems with installation, read the installation guide carefully. I had to download miniconda and pip install through that python installation to get it to work on my machine. Step 2: [obtain an API key](https://materialsproject.org/open) for Materials Project.   
**(a) task 1.** Look up the Materials Project ID for the 5 CIF cards found in the canvas file section (HW1 folder).

**(b) task 2.** Find all silicates (compounds that have at least silicon and oxygen, although other elements are allowed) with a unit cell larger than 1nm3.  
**(c) task 3.** Final all materials project entries for members of the n=2 Ruddlesden-Popper structure. CIF file in HW1 folder entitled RPn2.cif  
**(d) task 4.** Choose a material property and download all compounds that contain said property and have an e\_above\_hull less than 50 meV (i.e. are fairly stable), and also download the target property to your local computer (folder of CIF files and CSV file with target properties, respectively). Convert the CIFs to pymatgen Structure objects and store the pymatgen Structure objects, MPIDs, and target properties in a DataFrame with column titles, "structure", "mpid", and the name of the property, respectively.  
**(e) task 5.** Download all chemical formulas for entries that have been experimentally verified (i.e. Exptl. ICSD), MPID, and e\_above\_hull, and store these in a DataFrame with column names "composition", "mpid", and "e\_above\_hull". Reduce the list to a unique set of chemical formulas while keeping track of the repeat values from the other columns. Hint: use df.groupby or python sets