Instruction for installation:

Using environment: Python 3.7 (recommend using Anaconda package)

Dependent packages:

* Numpy (seems any recent version will do, already in Anaconda package)
* Sklearn (seems any recent version will do, already in Anaconda package)
* Matplotlib (seems any recent version will do, already in Anaconda package)
* Heapdict (v1.0.0 installation guide at <https://anaconda.org/anaconda/heapdict>)
* PyQtGraph (v0.10.0 installation guide at <http://www.pyqtgraph.org/>, optional if don’t want immediate visualization)
* PyOpengl (v3 installation guide at <https://anaconda.org/anaconda/pyopengl>, optional if don’t want immediate visualization)

Instruction for user:

* Put pos file and range file (RRNG format) in the same folder as these scripts
* The main file ‘Cluster\_Analysis\_Main.py’ controls how the scripts run:
  + First section: **data\_filename**: filenames for input pos or txt file contains coordinate and m/z of each data point; **rng\_filename:** range file in the format of RRNG. An example dataset name ‘test\_simple\_syn.pos’ and range file ‘test\_simple\_syn.rrng’ are provided.
  + Second section for clustering analysis:
    - **Ion selection:** defines what ion types will be used for cluster analysis. Options are explained in the script.
    - The **eps** is the maximum distance for searching minpts-th neighbor. It can be arbitrarily large for small system (< 5000 points) but becomes very inefficiency for large system. It should be slightly larger than largest value of minpts-th nearest neighbor distances in the dataset.
    - The **minpts** defines minimum number of neighbors for a point to be considered as core. It also smooths the reachability distance (RD) plot. An optimal minpts must be used to ensure a good quality of RD plot and sensitivity to small clusters. Experimentation with minpts is needed for unknown systems. For irradiated steels, a value between 10-50 is usually satisfactory.
    - The **method**: a string, determine which method for clustering, either ‘DBSCAN’ or ‘OPTICS-APT’.
    - **min\_cluster\_size**: min size of final clusters, below which will not be labeled noise.
  + Third section: Materials property section, currently not used. Saved for future automatic background estimation.
    - **Rho**: atomic density (#/nm3) of the matrix material
    - **Det\_eff**: detector efficiency.
    - **Con**: solute concentration.
  + Forth section: Expert options for hierarchical clustering algorithm. There is no need to change those unless you understand what’s behind.
    - **k**: determine local maximum calculation by comparing with k neighbors to the left and right.
    - **est\_bg**: estimated background RD. Currently automatic selection not available. Will be implemented in future. It still can be used by manual input to filter data.
    - **min\_node\_size**: min size of node allowed. Default to minpts.
    - **significant\_seperation**: a value to determine when hierarchical node needs to be splited. Default 0.75 works for most cases based on tests in original paper.
    - **cluster\_member\_prob**: a value determines the threshold for cluster membership in a RD hist using GMM. default 0.5.
    - **node\_similarity**: determine whether the child should replace current node. Does not affect leaves too much, but will reduce hierarchies for simplicity. Default 0.9.
  + The fifth section: controls output and visualization.
    - **output\_filename**: filename base for output, no extension.
    - **save\_file**: save output files, like cluster\_statistics, indexed pos, etc or not.
    - **show\_visualization**: show visualization or not
    - **show\_background**: show noise/matrix ions or not.

Instruction for output:

Several files will be given as output after clustering:

* \*\_ol\_RD\_CD.txt: a file contains the ordered list, reachability distance, and core distance from the OPTICS algorithm.
* \*\_cluster\_stats.txt: statistics of clusters, their index, center of mass, radius of gyration, equivalent spherical radius, contained ions.
* \*\_clusters.pos or .txt: pos file for visualization clusters in IVAS. Replaced the m/z with cluster\_id. Cluster\_id is move up since IVAS does not allow -1 in range file, which represent noise in this algorithm.
* \*\_clusters.rrng: range file for the corresponding pos to IVAS.
* \*\_log.txt: saved parameters used in the analysis.
* \*.cl: dumped snapshot of clustering class binary object. Can be used for debug purpose. Normally turned off and not output.