1. **Clustering molecules**

<http://www.rdkit.org/docs/Cookbook.html#clustering-molecules>

For large sets of molecules (more than 1000-2000), it’s most efficient to use the Butina clustering algorithm.

**def** ClusterFps(fps,cutoff=0.2):

**from** **rdkit** **import** DataStructs

**from** **rdkit.ML.Cluster** **import** Butina

*# first generate the distance matrix:*

dists = []

nfps = len(fps)

**for** i **in** range(1,nfps):

sims = DataStructs.BulkTanimotoSimilarity(fps[i],fps[:i])

dists.extend([1-x **for** x **in** sims])

*# now cluster the data:*

cs = Butina.ClusterData(dists,nfps,cutoff,isDistData=**True**)

**return** cs

The return value is a tuple of clusters, where each cluster is a tuple of ids.

Example usage:

**from** **rdkit** **import** Chem

**from** **rdkit.Chem** **import** AllChem

**import** **gzip**

ms = [x **for** x **in** Chem.ForwardSDMolSupplier(gzip.open('zdd.sdf.gz')) **if** x **is** **not** **None**]

fps = [AllChem.GetMorganFingerprintAsBitVect(x,2,1024) **for** x **in** ms]

clusters=ClusterFps(fps,cutoff=0.4)

1. 分子指纹和分子相似性
2. Topological Fingerprints：哈希拓扑路径
3. MACCS Keys：166个公共MACCS密钥基于SMARTS的实现。
4. Atom Pairs and Topological Torsions
5. Morgan Fingerprints (Circular Fingerprints)
6. 2D Pharmacophore
7. Pattern
8. Extended Reduced Graphs
9. 将分子从smiles转为二维图片[mol\_to\_image.py](rdkit_code/mol_to_image.py)