Types of clustering to do:

Label Propagation

MCODE

IPCA

Graclus

Louvain algorithm

Infomap

Measures to use:

Conductance (implemented by Professor Slota)

Modularity (implemented by Professor Slota)

Pairwise similarity within papers in the same cluster based on title and abstract.

Use cosine similarity

TF-IDF

Compute average internal similarity

Use sklearn and binary, bag of words, count vector, tf-idf and try and see which ones are better.

Steps:

1. python GraphingParser.py allPapers.txt keywords.txt allNodes.txt allQuery.txt

2. python GraphProcess.py allNodes.txt allQuery.txt 1 Hashed.txt

3. First for Louvain/Label propagation which are in the same code file,

python PGAEncoder.py Hashed.txt PGAEncoded.txt 1

4. ./PGAClustering.exe PGAEncoded.txt LouvainClusters.txt LPAClusters.txt

This will get the modularity and conductance for the label propagation and louvain and also will give the assignment of the clusters that each node belongs to for the Louvain and LPA clusters.

5. For Graclus

python GraclusEncoder.py Hashed.txt GraclusEncoded.txt 1

./graclus Hashed.txt 1000

6. Once the results from Graclus are obtained:

python GraclusClustering.py OutputFile GraclusResults.txt

7. ./ClusterEvaluation.exe GraclusResults.txt

This will give modularity and conductance for Graclus clustering

8. For Infomap

python InfomapEncoder.py Hashed.txt InfomapEncoded.txt 0

Then run Infomap

run InfomapClustering.py on it

Then run ClusterEvaluation with the results

Then do same for MCODE and IPCA.