# COMP 9102 Data Management and Information Retrieval
# Assignment 3
# Using Clustering for Community Search
# Due Date: Dec 1st, 2017 5:00pm

## Question specification.

In this assignment, you are required to implement and benchmark community search algorithm using clustering. The algorithms should first compute user similarity by vertex similarity and personal page-rank and should perform a K-means clustering.

**Similarity Measure：**

1) See lecture notes for details of vertex similarity and personal page-rank. For PPR, use the iterative method with uniform initial distribution (for $p_i$, e[j]=1/degree(i) if i and j are connected, 0 otherwise). Set $\alpha$=0.1 and a convergence threshold of 10^-5.
2) Make sure your similarity algorithms are correct by testing on a few points (for example, 4 or 5 points).
3) Compute the metrics for full dataset and dump the similarity metrics into disk files if necessary.

**Clustering:**

1) Apply K-means clustering that takes the similarity metrics as input. For simplicity, use Euclidean distance.

**Evaluating Clustering Performance:**

1) You should evaluate the quality of clustering using the label file by following criteria: Purity, Entropy and Normalized mutual information (NMI). See https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html for reference.
3) You can ignore users without tag for evaluation, but do include these users in previous steps.

**Dataset.**

Navigate to moodle page.
1) Download Assignment3.zip to your local folder.
2) Read description.txt for detailed description of data. We treat retweet data as an undirected graph.

## Requirements.

1）Implement and test the requested indexes and algorithms.
2) Run a benchmark experiments where you test the performance of your algorithms. Your benchmark scheme should evaluate the performance of both metrics by clustering quality on different choice of k (for example 5, 10 ,15) .
3) Write a short report about your implementation and the experiments and submit the report together with your code at the course website. Your code should be compilable without problems and you should include basic instructions on how to compile and use it. Your program can be written in your preferred programming language (e.g., C, C++, Java, Python, etc.). Your program must run within reasonable time.

**Submission.**

Please submit your assignment (one **ZIP** file) to moodle on or before 5:00pm, Dec 1st, 2017. Make sure all contents are readable.

Please feel free to post your questions on **Moodle forum**, or contact TA Daniel (dhding2@cs.hku.hk) if you encounter any difficulty in this assignment. We would be happy to help.