Microsoft - DAT278x

From Graph to Knowledge Graph - Algorithms, Theory and Applications

LAB 2 – Knowledge Graph

Heterogeneous Knowledge Graph Exploration – Instructions

In this lab, you will construct a heterogeneous academic graph from a subset of Microsoft Academic Graph (MAG). You will also explore this heterogenous graph to examine its properties.

The input to this Lab exercise contains two files. The first paper-author file is in the same format with the first lab's input file. It covers a set of academic publications, each of which is associated its authors. It is formatted as a tab-separated value file with two columns. The first column represents the paper Id, and the second column represents one of its associated authors' Ids. The second paper-venue file is also a two-column tab-separated-value file. Its first column represents the paper Id and the second one represents its venue name. Note that each paper may have multiple authors, but it is associated with only one venue. Below is a sample input of the paper-venue file:

PaperId	VenueName
2125526935	IJCAI
107169386	IJCAI
2742657268	KDD

The goal is to build a heterogeneous author-paper-venue graph. An triple is linked among one author, one paper, and one venue if and only if the author publishes this paper in this venue. The output could be a list of triples of the constructed heterogeneous graph. For example, one sample output is listed below:

AuthorId	PaperId	VenueName
112778	2125526935	IJCAI
1152630	107169386	IJCAI
757571	2742657268	KDD

In this Lab assignment, please follow the Azure Data Lake (ADL) setup guide to upload the input academic publication data (Paper_venue.tsv) into the ADL folder (course-kg/input/). The full file path is "course-kg/input/Paper_venue.tsv". After this uploading, the "course-kg/input/" folder should contain both the "Paper_venue.tsv" file and the "Paper_authors.tsv" file.

The associated U-SQL script is “DAT278x_Lab2_KnowledgeGraph.usql”. In this Lab, you will fill in the missing code in this script to answer the questions. The missing code is surrounded with “//Q1: START CODE HERE” and “//Q1: END CODE HERE” if taking Lab Question 1 as an example.

After filling the code lines for all questions and then running the script, the heterogeneous graph will be constructed and the answers to the questions will be output.