CIS 5516: Principles of Data Management - Spring 2017

About Me

Course Home

Schedule

Projects

Assignments

Announcements

Phase 1: Crawl Publication Records From Google Scholar (GS).

Requirements

1. Input: A list of author names in a text file.
2. Your tool needs to read the author names from the file, go to GS with each of them, and retrieve the publication records for each author.
3. Ambiguous Names: Some author names may be ambigous, which means that there can be multiple people with the same name. Your tool needs to recognize this and gather all (if possible) such poeple from GS before it proceeds to collecting the publications records.
4. For each author, you need to extract the following pieces of information: name, homepage, email, areas of interests, position, affiliation, and h-index.
5. For each publication record, you need to extract Title, the list of authors, publication venue, and year.
6. Your crawler needs to be polite.

Deliverables

- Give a table or a chart with the number of different people for the following names: Jie Wu, Wei Zhang, Slobodan Vucetic, Qiang Zhang, and Krishan Kant.
- Give the average h-index for the above names. For example, there appear to be 17 people who share the name Krishna Kant in GS. You need to get the h-index of each of them and compute the average. Repeat the process for the 5 names given above. This needs to be done automatically.

Start early!

Last Updated: 2/1/2017