**Final Project Proposal Outline**

**Group members: Yingjun Guan, Jing Du, Qi Qi**

1. Purpose of your project

From PubMed, a database of publication bibliography, our group focuses on the data structure of author’s affiliation information and intend to clean the attribute for discovering its quantitative distribution, and furthermore, the clustering structures/ relationship within them. Namely, what are the affiliations and how they are related to each other. At the same time, our group also compares various software on testing their performance on scalability and other functions.

* Describe what you intend to do for the final project
  + With the large database of PubMed, we firstly test the performance of different data cleaning software, like OpenRefine, Data Wrangler, etc. And report on how large file they can handle, in preparation of the following steps.
  + From the database, the attributes include (PMID, author, author order, journal name, year of publication, author affiliation, Abstract, Mesh Term, etc.) select the attribute of author affiliation (and others if needed for future cleaning)
  + Preliminary cleaning by detecting and deleting those null values, some author affiliation information is missing, or not included in the paper.
  + [Hard] The affiliation information is usually in the format of {department, university, city, state, country}, which looks like “Nursing Department, National University Hospital, Singapore.” Noticing that in this case, the city and state are of missing value.

Parsing the data (with the delimiter of comma), and assign the contents to the correct columns {department, university, city, state, country, and other potential level as needed}

* + Take out one parsed column, for example, “department”. Take all those department values (only for those entries whose department value is not empty), and get prepared for deeper cleaning in the next step.
  + Clean the department information retrieved last step and do a deep cleaning on at least but not constrained to the following points:
    - Word Syntax or abbreviation. Eg: some department are taken “dept” for short, similarly, “univ”, “inst”, etc.
    - Wrong spelling or typos. Eg: If the “St John’s University” is spelled as “St Johns University” or “St John University”, they should get cleaned [make sure though there’s no such St John University beforehand.]
    - Same department but different expression. Eg: department of Nursing and Nursing department. This is tricky, sometimes, they are different, but sometimes, they not. Have them discussed.
    - Find other things worth cleaning for the information.
  + With the cleaned data, take column “department” as an example, use SQLite (count, sort, group by, order by, etc.) to discover the quantitative distribution of the data. [We might also need to join columns of department and others to do some calculation, not sure yet.]
  + [Related. Fancy output expected.] Using Machine Learning method to cluster the department to see their relationship. (This can be done with Python or Weka)
* What is the motivation behind it?
  + The PubMed database is of great value with containing tons of information of publication bibliography. However, the data are very dirty, and needs cleaning.
  + The study behind relationship of subjects and affiliations are sparse so far. The affiliations have strong similarity and dissimilarity from one another.
  + And Yingjun’s engaged in a related research now. Yingjun is right now under advisory of Prof. Torvik.

1. Dataset Description

PubMed is a database for Medline retrieval on the world-wide web. The dataset contains PubMed author names together with the papers annotated by publication count, time-span, affiliations, topics, journals, co-authors, citations as well as imputed data from MapAffil. The following links provide more information on the dataset.

[dataset information: https://www.nlm.nih.gov/pubs/factsheets/pubmed.html

dataset download: http://abel.ischool.illinois.edu/cgi-bin/download/request.pl]

In the database of PubMed, there are more than 19.6 Million records of medical publication. The dataset is way too large. We will try use it to test the scalability of task one, but might cut a subset of the database for the rest of the project. The way of cutting the subset will be discussed in the report.

The dataset has the following information (fields or attributes):

[PMID au\_order lastname firstname year journal affiliation type city state country lat lon fips]

There are more than 19 million publication with more than 39 million of author information (due to the co-authorship phenomenon).

1. Methods

* Data cleaning tools you intend to use (if any)
  + For data cleaning scalability comparison, try use the following software: OpenRefine, OpenWrangler, and other cleaning tool.
  + For data parsing, try use the RegEx.
  + For data querying, try use SQLite, Oracle or others.
  + For final machine learning and visualization, try use Python.
* Other methods or steps (not necessarily data cleaning) that you will be using
  + Machine learning as mentioned above.

1. Project deliverables

In the end of this project, it will be expected to see the cleaned data and cool visualization of the data. All the figure outputs will be covered in the presentation slides, and the written report.

As the same time, we will also provide the cleaned data (file in the format of csv) and the original dataset download link.