Stage5 Data Analysis

Zhiwei Fan zfan29@wisc.edu
Lingfeng Huang lhuang58@wisc.edu
Fang Wang fwang64@wisc.edu

December 15, 2016

Term explanation

tableA: book data obtained from Barnes and Noble.
tableB: book data obtained from Good reads.
scheme of both tables: id, title, authors, ISBN13, pages, publisher, publishedYear, publishedMonth, publishedDay
tableC: the table after blocking stage.
tableE: the table after merging stage. It has the same schema as tableA and tableB, but its id is not obtained from tableA nor tableB. Its id is new and in increasing order start from 0.

TODO

random forest scores?

Description of table merging

During the project stage 4, we chose *random forest* as our final selected model because of its high F1 score. Thus, in this final stage, we first obtained the prediction result from running *random forest* to the blocking tableC from stage4 (matching tuple pairs with information from tableA and tableB). By looking at the prediction boolean array generated from random forest, we generate a intermediate table containing tuple pairs that are predicted to be matched. We call this filtered-table. Our goal of merging is to ensure we find all unique tuples from tableA and tableB so we need to use this filtered-table in the next stage.

how did we merge stages?:

To generate tableE, we first keep all the data from tableA (always select the values from the tuple from table A) because tableA has well formatted data and there are lots of missing data in tableB. Thus, when there is a match, tableA's data has higher priority than data in tableB. After this step, we still have't added the tuples that are not present in tableA but present in tableB. Then, using the filtered-table from the last step, we are able to know which tuples in tableB have already presented in the tableA. We don't need these data so we take tableB and find all tuples which IDs are not present in the filtered-table. These are the tuples that are not present in tableA but present in tableB. By adding these tuples to tableE as well, we have a complete tableE.

***Statistics on Table E*:**

Labeling

Feature Construction

Development and Evaluation Sets

Initial Cross Validation Measurement

Final Cross Validation Measurement and Measurement on Evaluation Set

Final Cross Validation Measurement

Measurement on Evaluation Set

Of course, we have *evaluated* our *final selected models random forest* along with *other models* on the *evaluation set* J and the corresponding results are shown as the

List of final features in final feature set

Approximate Time Estimation

Discussion