CS 838 - Project Stage 3

Jale Dinler, Riccardo Mutschlechner, Steven Lamphear

dinler@wisc.edu, riccardo@cs.wisc.edu, slamphear@wisc.edu

April 2, 2017

- We've matched albums from our pitchfork reviews and discogs data sets. We've created p_table from Pitchfork data and d_table from Discogs data. p_table has 18391 rows and 4 columns. The first row is the header followed by rows containing album information. The columns include information such as title (album name), artist and year (album release), Each row is uniquely identified by the 'reviewid' column.
  d_table has 466936 rows and 4 columns. The first row is the header followed by rows containing album information. The columns include information such as title (album name), artist and year (album release), Each row is uniquely identified by the 'id' column.

- We've applied overlap blocker and attribute equivalence blocker to the column 'title'. There are 7682 tuples remaining after blocking.

- After sampling, we've labeled 3000 tuples.

- We've applied all 6 matchers to set I with CV, the results are shown in the below table:

	DT	SVM	RF	Log Reg	Lin Reg	NB
Precision	99.2%	97.1%	99.4%	99.4%	99.4%	99.4%
Recall	99.2%	99.8%	99.3%	99.5%	99.1%	99.3%
F1	99.2%	98.4%	99.4%	99.6%	99.3%	99.3%

- As we already got good results after the first CV, we went ahead and applied all 6 matchers to set J, results were also good with set J. Since RF, Log Reg, Lin Reg and NB have similar precision, we looked at their recall and picked RF as our matcher.

	DT	SVM	RF	Log Reg	Lin Reg	NB
Precision	98.88%	97.92%	99.38%	99.5%	99.5%	99.5%
Recall	99.13%	99.38%	99.38%	98.88%	99.01%	99.01%
F1	99.01%	98.64%	99.38%	99.19%	99.25%	99.25%

- We've adapted the code from the examples given in user manual for py_entitymatching. Assuming time estimates question is for run time, overlap blocker is 6 minutes, attribute equivalence blocker is 10 seconds, labeling is 2 hours, finding the best matcher (arranging the code, run time) is 20 minutes since we've got desired results without debugging. Getting to blocking step was the painful part.