Stage 3

Dan Howe and Keith Funkhouser
CS838: Data Science
April 2, 2017

- Describe the type of entity you want to match, briefly describe the two tables (e.g., where did you obtain these tables), list the number of tuples per table.
    - We used the provided Silicon Valley dataset. We chose to match the 'songs' table (downsampled to 3038 tuples) and the 'tracks' table (downsampled to 3500 tuples).
- Describe the blocker that you use and list the number of tuple pairs in the candidate set obtained after the blocking step.
    - We used the overlap blocker on both song name and artist name, with a minimum overlap of one word. That successfully reduced the number of candidates by approximately two orders of magnitude, after which we used a rule based blocker based on the Jaccard similarity of the song title. This reduced the number of tuple pairs to 543.
- List the number of tuple pairs in the sample G that you have labeled.
    - We eliminated 10 tuple pairs due to missing data. That left us with 533 tuple pairs, all of which we labeled.
- For each of the six learning methods provided in Magellan (Decision Tree, Random Forest, SVM, Naive Bayes, Logistic Regression, Linear Regression), report the precision, recall, and F-1 that you obtain when you perform cross validation **for the first time** for these methods on I.
    -

Method	**Precision**	**Recall**	**F1**
Decision Tree	0.886904	0.885568	0.884977
Random Forest	0.913180	0.942320	0.926725
SVM	0.789615	0.984525	0.875998
Linear Regression	0.920335	0.937755	0.928635
Logistic Regression	0.926929	0.948289	0.937370
Naive Bayes	0.883408	0.653312	0.748647

- Report which learning based matcher you selected after that cross validation.
    - We selected logistic regression, based on it having the highest average precision score.

- Report all debugging iterations and cross validation iterations that you performed. For each debugging iteration, report (a) what is the matcher that you are trying to debug, and its precision/recall/F-1, (b) what kind of problems you found, and what you did to fix them, (c) the final precision/recall/F-1 that you reached. For each cross validation iteration, report (a) what matchers were you trying to evaluate using the cross validation, and (b) precision/recall/F-1 of those.
  - No debugging iterations were performed.
- Report the final best matcher that you selected, and its precision/recall/F-1.
  - (See above)
- **It is important to note that all precision/recall/F-1 numbers asked for in the aboves are supposed to be numbers obtained via CV on the set I. Do not yet use set J.**
- Now report these numbers:
  - For each of the six learning methods, train the matcher based on that method on I, then **report its precision/recall/F-1 on J.**

Method	**Precision**	**Recall**	**F1**
Decision Tree	90.16%	94.02%	92.05%
Random Forest	95.73%	95.73%	95.73%
SVM	81.69%	99.15%	89.58%
Linear Regression	96.64%	98.29%	97.46%
Naive Bayes	97.22%	59.83%	74.07%

  - For the final best matcher Y selected, train it on I, then **report its precision/recall/F-1 on J.**

Logistic Regression	95.76%	96.58%	96.17%

- Report approximate time estimates: (a) to do the blocking, (b) to label the data, (c) to find the best matcher.
  - Blocking: 3 hours
  - Label data: 1 hour
  - Best matcher: 2 hours
- Provide a discussion on why you didn't reach higher recall, and what you can do in the future to obtain higher recall.
  - To obtain higher recall, we should reduce our false negatives. There were only 4 false negatives produced by the logistic regression model. The major causes were: misspellings in data, short artist names which caused the overlap between artist (in the song table) and list of artists (in the tracks table) to be small, and a mislabeling by us.