CS 784
Project Part 4: Matching
Kaashyapee Jha and Margaret Pearce

• *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.*

Learning method	Precision	Recall	F1
Decision tree	0.715649	0.799944	0.794831
Random forest	0.858442	0.755518	0.789085
SVM	0.803683	0.774510	0.813249
Naive Bayes	0.667024	0.877759	0.763993
Logical Regression	0.837121	0.824090	0.83705
Linear Regression	0.741746	0.822465	0.825121

• *Report which learning based matcher you selected after that cross validation.*

We debugged both random forest and linear regression matchers.

• 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.

Debugging random forest:

- Iteration 1
    - a) Random forest, precision 0.858442, recall 0.755518, f-1 0.789085
    - b) Problems found: The random forest method repeatedly creates a tree from randomly selected features, then takes the majority vote among the set of trees. Because the features are randomly selected, we saw many trees created that used different similarity measures for the same attribute. The resulting trees sometimes don't have enough variety in their attributes to make accurate predictions. To fix this, we tried to use a limited set of features (one similarity measure per feature).
    - c) Precision: 0.7353, recall = 0.7813, f-1 0.7576 (not much improvement, did not end up using this)
- Iteration 2
    - a) Random forest, precision 0.858442, recall 0.755518, f-1 0.789085

    - b) Problems found: Many trees did not consider ISBN13. If ISBN13 was an exact match for two tuples, then the entities should be matched. To fix this, we added a trigger to check if match ISBN13 values matched (false negatives).
    - c) Precision: 0.7692, recall: 0.9375, f-1: 0.8451
- Iteration 3
    - a) Random forest, precision 0.858442, recall 0.755518, f-1 0.789085
    - b) Problems found: If title, author, and page count matched, then the tuples should be considered matched (false negatives).
    - c) Precision: 0.75, recall: 0.9375, f-1: 0.8333
- Iteration 4
    - a) Random forest, precision 0.858442, recall 0.755518, f-1 0.789085
    - b) Problems found: If title, author, and publisher matched, then the tuples should be considered matched (false negatives).
    - c) Precision: 0.7632, recall: 0.9063, f-1: 0.8286
- Iteration 5
    - a) Random forest, precision 0.858442, recall 0.755518, f-1 0.789085
    - b) Problems found: If title differs by a reasonable amount between the two tuples, the entities should not be matched (false positive).
    - c) Precision: 0.8235, recall: 0.875, f-1: 0.8485
- Iteration 6
    - a) Random forest, precision 0.858442, recall 0.755518, f-1 0.789085
    - b) Problems found: If page count, publisher, and publish date don't match within a reasonable amount, then the tuples should not be considered matched (false positives).
    - c) Precision: 0.8, recall: 0.875, f-1: 0.8358
- Iteration 7
    - a) Random forest, precision: 0.8, recall: 0.875, f-1: 0.8358
    - b) Problems found: Publish dates have different formats between table A and table B. For example, table A says "January 1st 2015" and table B says "1/1/15". To effectively compare them, three new features were created that return 0/1 if month matches, day matches, and year matches between the tuple pairs.
    - c) Precision: 0.7568, recall: 0.875, f-1: 0.8116 (not much improvement without the triggers)
- Iteration 8
    - a) Random forest, precision 0.858442, recall 0.755518, f-1 0.789085
    - b) Problems found: Some labels were incorrect in the manually labeled data. To fix this, we revised some labels, read this in as a separate file, and re-fit the matchers with triggers applied.
    - c) Precision: 0.9032, recall: 0.7568, f-1: 0.8235 (improvements on all measures)

Debugging logistic regression:

- Iteration 1
    - a) Log Reg, precision: 0.846825, recall: 0.791320, f1: 0.808742
    - b) Problems found: feature selection- some features- like SecondAuthor, ThirdAuthor, and Publication date (beause of mismatched format) are not that relevant to matching so exclude those features and create new subset of features
    - c) precision: 0.90, recall: 0.871 f1: 0.8852

- Iteration 2
    - a) Log Reg, precision:0.863492, recall: 0.838022, f1: 0.841158

    - b) Problems found: To try to eliminate false negatives, add a trigger with a rule that states that of ISBN13 is an exact match, then the label will be 1
    - c) precision: 0.931 , recall: 0.871, f1: 0.90

- Iteration 3
    - a) Log Reg, precision 0.846825, recall: 0.791320, f1: 0.808742
    - b) Problems: Focus on false positives- add triggers to place more stringent rule of when a pair should be declared a match by setting thresholds for Title, FirstAuthor, and Publisher
    - c) precision: 0.9333, recall: 0.9032 , f1: 0.918

- Iteration 4
    - a) Log Reg, precision: 0.846825, recall: 0.791320, f1: 0.808742
    - b) Problems: Add last two triggers into one trigger to handle both FP and FN
    - c) precision: 0.9333, recall: 0.9032, f1: 0.918

- Iteration 5
    - a) Log Reg, precision: 0.846825, recall: 0.791320, f1: 0.808742
    - b) Problems: try to decrease FP, add a blackbox feature that returns 1 if difference between page counts is less than 10
    - c) precision: 0.90, recall: 0.871, f1: 0.8852

- Iteration 6
    - a) Log Reg, precision: 0.846825, recall: 0.791320, f1: 0.808742
    - b) Problems: make a new trigger setting thresholds for Title, FirstAuthor, and PageCount; add this trigger to the previous two
    - c) precision: 0.8788, recall: 0.9355, f1: 9063

Cross Validation

- CV 1
    - a) Random forests with all features and all triggers
    - b) Precision: 0.822842, recall: 0.933532, f-1: 0.872731
- CV 2
    - a) Random forests with added features for publish dates, no triggers
    - b) Precision: 0.917143, recall: 0.758214, f-1: 0.823167
- CV 3
    - a) Random forests with added features for publish dates and all triggers
    - b) Precision: 0.861732, recall: 0.947899, f1: 0.898437
- CV 4
    - a) Random forests with added features for publish dates, all triggers except one about author/ title/ page count, revised labels
    - b) Precision: 0.910714, recall: 0.844786, f-1: 0.872804
- CV 5
    - a) Log regression with reduced features
    - b) Precision: 0.87000, recall: 0.829654, f1: 0.839628
- CV 6:
    - a) Log regression with ISBN trigger, reduced features
    - b) Precision: 0.891818, recall: 0.868889, f1: 0.872294

- CV 7
    - a) Log regression with ISBN trigger and Publisher trigger, reduced features
    - b) Precision: 0.907143, recall: 0.869231, f1: 0.876875
- CV 8
    - a) Log regression with ISBN trigger, Publisher trigger, and Page count trigger, reduced features
    - b) Precision: 0.832949, recall: 0.862143, f1: 0.844562

• Report the final best learning-based matcher that you selected, and its precision/recall/F-1.

Final best: Random forests with custom features for publish dates, all triggers except one applied, and corrections to the original labels. Precision: 0.910714, recall: 0.844786, f-1: 0.872804.

• Report all the rules that you added to this matcher, and the final precision/recall/F-1 that you reach. When reporting these rules, report in which order the rules are applied. 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.

From first to last applied:

- If Title_Title_jac_qgm_3_qgm_3(ltuple, rtuple) <= 0.6, don't match
- If PageCount_PageCount_anm(ltuple, rtuple) <= 0.8 and Publisher_Publisher_jac_dlm_dc0_dlm_dc0(ltuple, rtuple) <= 0.35 and PublishDate_PublishDate_lev(ltuple, rtuple) <= 0.14, don't match
- If ISBN13_ISBN13_lev(ltuple, rtuple) == 1, match
- If Title_Title_lev(ltuple, rtuple) >= 0.9 and FirstAuthor_FirstAuthor_jac_qgm_3_qgm_3(ltuple, rtuple) >= 0.9 and Publisher_Publisher_lev(ltuple, rtuple) >= 0.9, match

• 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.
– For the final best learning method Y selected, train it on I, then report its precision/recall/F-1 on J.
– For the final best matcher (that is, Y ∗, which is the learning-based method Y plus the rules), train it on I then report its precision/recall/F-1 on J.

Training all methods on I, evaluating on J:

Learning method	Precision	Recall	F1
Decision tree	100%	75.76%	86.21%
Random forest	**100%**	**84.85%**	**91.8%**
SVM	0	0	0
Naive Bayes	93.75%	90.91%	92.31%

Logical Regression	93.33%	84.85%	88.89%
Linear Regression	96.0%	72.73%	82.76%

Final best learning method trained on I, evaluated on J:

Learning method	Precision	Recall	F1
Random forest	100.0%	84.85%	91.8%

Final best matcher trained on I, evaluated on J:

Learning method	Precision	Recall	F1
Random forest	100.0%	84.85%	91.8%

• Report an approximate time estimate: (a) how much did it take to label the data, (b) to find the best learning-based matcher, and (c) to add rules to the learning-based matcher.

a. Labeling the data: 3 hours initially, 1 hour reviewing together, 1 hour revising labels during matching process
b. 8 hours each (estimate)
c. 8 hours each (estimate)