CS 638 Project Report: Stage-4

Project Members (Group #23):

#	Name	E-Mail ID
1.	Ajay Joseph Thomas	ajosephthoma@wisc.edu
2.	Rahul Singh	rsingh53@wisc.edu
3.	Ting Lei	tlei@wisc.edu

Q1: For each of the five learning methods (Decision Tree, Random Forest, SVM, Naive Bayes, Logistic 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.

Answer: (for k=5)

Model\Accuracy	Precision	Recall	F-1
Naïve Bayes	0.966063	0.890288	0.926359
Linear Regression	0.954245	0.943296	0.94861
Logistic Regression	0.96708	0.934969	0.950637
Decision Tree	0.933395	0.957763	0.944311
Random Forest	0.958165	0.936598	0.947073
Support Vector Machine	0.948669	0.912234	0.929629

Q2: Report which learning based matcher you selected after that cross validation.

Answer: Linear regression had highest precision and high recall and high F-1 score for k=5 fold. However, Logistic Regression did better for k=10, k=20 and k=50 so we picked Logistic Regression as our best matcher.

Q3: 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,

Answer: Although we had reasonably high accuracy (96.7% precision, 93.5% recall and 95.1% F-1 score) with Logistic Regression matcher, we still went ahead and tried all matchers (i.e. Naïve Bayes, Linear Regression, Logistic Regression, Decision Tree, Random Forest and Support Vector Machine) for debugging iteration as well.

(b) what kind of problems you found, and what you did to fix them?

Answer: Name of directors had following issues:

- Abbreviated (first/middle) name in one table and full name in another table
- Order of directors' name was different in left and right table.
- Extra/Missing non-alphabetic symbols such as dot, and non-English symbols.
- Inconsistency in space after comma.

Solution: We copy-pasted the names from left table to right table for these cases.

(c) the final precision/recall/F-1 that you reached.

Answer:

Iteration-1 (different run than the above):

Model\Accuracy	**Precision**	**Recall**	**F-1**
Naïve Bayes	92.05% (81/88)	93.41% (85/91)	92.72%
Linear Regression	94.44% (85/90)	93.41% (85/91)	93.92%
Logistic Regression	95.65% (88/92)	95.60% (87/91)	95.63%
Decision Tree	91.21% (83/91)	92.31% (84/91)	91.75%
Random Forest	92.31% (84/91)	94.51% (86/91)	93.39%
Support Vector Machine	93.26% (83/89)	94.51% (86/91)	93.88%

Iteration-2:

Model\Accuracy	Precision	Recall	F-1
Naïve Bayes	95.4% (83/87)	92.22% (83/90)	93.79%
Linear Regression	96.55% (84/87)	93.33% (84/90)	94.92%
Logistic Regression	96.77% (90/93)	100.0% (90/90)	98.36%
Decision Tree	92.47% (86/93)	95.56% (86/90)	93.99%
Random Forest	96.67% (87/90)	96.67% (87/90)	96.67%
Support Vector Machine	95.6% (89/93)	94.44% (85/90)	95.01%

(d) For each cross-validation iteration, report

a. what matchers were you trying to evaluate using the cross validation,

Answer: All of them.

b. and precision/recall/F-1 of those.

Iteration-1:

Model\Accuracy	Precision	Recall	F-1
Naïve Bayes	94.83% (110/116)	90.91% (110/121)	92.83%
Linear Regression	94.92% (112/118)	92.56% (112/121)	93.72%
Logistic Regression	95.0% (114/120)	94.21% (114/121)	94.61%
Decision Tree	94.92% (112/118)	92.56% (112/121)	93.72%
Random Forest	94.26% (115/122)	95.04% (115/121)	94.65%
Support Vector Machine	94.02% (110/117)	90.91% (110/121)	92.44%

Iteration-2:

Model\Accuracy	Precision	Recall	F-1
Naïve Bayes	95.19% (99/104)	84.87% (101/119)	89.74%
Linear Regression	97.09% (100/103)	84.03% (100/119)	90.09%
Logistic Regression	97.32% (109/112)	91.6% (109/119)	94.37%
Decision Tree	97.27% (107/110)	89.92% (107/119)	93.45%

Random Forest	96.49% (110/114)	92.44% (110/119)	94.42%
Support Vector Machine	96.3% (104/108)	87.39% (104/119)	91.63%

Q4: Report the final best learning-based matcher that you selected, and its precision/recall/F-1.

Answer:

Best-Matcher\Accuracy	Precision	Recall	F-1
Logistic Regression	97.14% (102/105)	91.89% (102/111)	94.44%

Randomization in multiple stages (such as while splitting datasets into train and test) are causing minor variations in each iteration (which is expected).

Q5: Now report the following:

– a) For each of the five learning methods, train it on I, then report its precision/recall/F-1 on J.

Answer:

Model\Accuracy	Precision	Recall	F-1
Naïve Bayes	96.97% (96/99)	86.49% (96/111)	91.43%
Linear Regression	97.09% (100/103)	90.09% (100/111)	93.46%
Logistic Regression	97.14% (102/105)	91.89% (102/111)	94.44%
Decision Tree	97.12% (101/104)	90.99% (101/111)	93.95%
Random Forest	96.19% (101/105)	90.99% (101/111)	93.52%
Support Vector Machine	89.19% (99/111)	89.19% (99/111)	89.19%

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

Answer:

Best-Matcher\Accuracy	Precision	Recall	F-1

Logistic Regression	97.14% (102/105)	91.89% (102/111)	94.44%

- c) List the final set of features that you are using in your feature vectors.

Answer: Initially we planned to create feature vectors from 4 features, however considering the correlation between feature and actual accuracy, we decided to use only "Directors" as the feature and the feature vector we used was as follows:

- Directors_Directors_jac_qgm_3_qgm_3
- Directors_Directors_cos_dlm_dc0_dlm_dc0
- Directors_Directors_jac_dlm_dc0_dlm_dc0
- Directors_Directors_mel
- Directors_Directors_lev_dist
- Directors_Directors_lev_sim
- Directors_Directors_nmw

Q6: Report an approximate time estimate:

(a) how much did it take to label the data

Answer: We labeled the data manually and it took 5-10 minutes to write an excel formula for labeling and around 30-45 minutes to verify our labeling. We couldn't use py_entitymatching module for labeling the data as it seemed to go in infinite loop (I waited for 30 minutes for it to stop. It wasn't showing progress bar either. On running 'top' command, system showed that there were python processes running in the background.)

(b) and to find the best learning-based matcher.

Answer: It took around 60-90 minutes.

Q7: Discuss why you can't reach higher precision, recall, F-1.

Answer:

1. We couldn't think of any issue in our data anymore.

2. Accuracy (Precision/Recall/F-1) is very high in our final stage.