University of Wisconsin-Madison

Computer Science 839

Data Science: Principles, Algorithms, and Applications

Information Extraction from Natural Text

Team Member:	Contact Email:
Jiatong Li	jli872@wisc.edu
Siyu Wang	swang739@wisc.edu
Xinyu Zhang	xzhang959@wisc.edu

March 10, 2019

1 Data

Text Documents Book reviews with well-formed sentences from *goodreads*, a website allowing free accesses.

Entity Type: Person Name.

Mark Examples:

- <>Peter</>, <>Harry Potter</>, <>John Howard Griffin</>
- Mr. <>E.B. White</>'s
- the <>Trueba</> family
- the <>Nobel</> Prize winner

Data Sets We generated 450 text documents with 904 mentions of names in total and split them by random order into development set I and test set J.

Table 1: Size of Data Sets

Number of :		Document	Mention
Set I:	Dev Set	300	583
Set J:	Test Set	150	332
Total:	Full Set	450	915

2 Model Development

2.1 Classifier M

Random Forest has the highest F1 and recall during the first time's cross validation on set I, thus is selected as classifier M in the first step. The detailed results are as follow.

Table 2: Performance of Classifiers on Set I.

(%)	Precision	Recall	F1
Random Forest (M)	72.55	**78.81**	**75.55**
Decision Tree	71.90	76.31	74.07
Support Vector Machine	72.58	73.17	72.87
Linear Regression	**76.92**	69.30	72.91
Logistic Regression	74.27	74.21	74.27

2.2 Classifier X

Random Forest still kept the highest F1 and recall, thus was selected as the best model on set J.

Table 3: Performance of Classifiers on Set J.

(%)	Precision	Recall	F1
Random Forest (X)	71.40	**83.10**	**76.81**
Decision Tree	71.29	82.98	76.69
Support Vector Machine	70.39	77.05	73.57
Linear Regression	**76.49**	69.74	72.96
Logistic Regression	73.34	73.90	73.62

2.3 Rules

Positive Rules for identification of person names.

- The first letter is uppercase, and the following letters, if exist, must be lowercase, such as "Peter".
  - **Location:** Regexp match in codes, "$r'[A-Z][-Z]*\$'$".
- With a prefix like "The/An/..." and a suffix like "family/Prize/...", the middle part should be a person name, such as "the Trueba family."
  - **Location:** Set [cutoff] in codes.

- When certain words like called/said/told appear, we think the former/latter word should be a person name.
    - **Location:** Sets [right_pre] and [right_suf] in codes.

Negative Rules for ruling out words that are not person names.

- With a prefix like "The/An/..." and a suffix without "family/Prize/...", the middle part should not be a person name.
    - **Location:** Set [cutoff] in codes.
- With suffix like "School/Hospital/...", the words is unlikely to be person names.
    - **Location:** Set [wrong_suf] in codes.

3 Result

Based on the selected Random Forest classifier and added post-processing rules. We test the performance of the final classifier Y on the test set J with the following results.

Table 4: Performance of Final Classifier Y on Set J.

(%)	Precision	Recall	F1
Random Forest & Rule (Y)	90.87	79.07	84.56