

# Project Report Stage Two

Hongyi Wang, Hao Fu, Miao Yang

March 12, 2017

## 1 Name of All Team Members

- Hongyi Wang
- Hao Fu
- Miao Yang

## 2 Entity Extraction

### 2.1 Entity Type

In the stage two of this project, we planed to extract the **name of university** from the text files we collected in stage one. The features we extract from all mentions contain:

- length of mentions (i.e. the string)
- if the word "university" contains in the mention
- if the word "state" contains in the mention
- if a name of state (e.g. Wisconsin, California, and etc) contains in the mention
- if dash (i.e. -) contains in the mention
- if the mention contain number
- if all letters in the mention are capital (e.g. MIT)

The labels in our extracted data instances are "True" (i.e. the mention is a university name) or "False" (i.e. the mention isn't a university name)

### 2.2 Entity Examples

We give several examples of mentions here:

MIT, Rochester Institute of Technology, University of California–Riverside, UCR, ASU-Temple, CUA, and etc. The those foregoing mentions were all labeled True in our data set.

Here are also several False labeled instances:

School's, Undergraduate, NCAA-division, Institution, and etc.

### 2.3 How we Implement Information Extraction (IE)

All mentions are label manually, the mentions that labeled "True" were always the real university names, the mentions that labeled "False" were chose randomly from each text files. We use the format "<p university name p>" to label positive mentions, and format "<n fake-university name n>" to label negative mentions. For further use, we also distinguished private university names and public university names. In specific, we labeled private university name using "<p1 private university p1>" while

public university name using "<p2 public university p2>". Here we show an example of the labeled text files:

**<p1 Harvard University p1>** is a private institution that was founded in 1636. It has a total undergraduate enrollment of 6,699, its setting is urban, and the campus size is 5,076 acres. It utilizes a semester-based academic calendar. **<p1 Harvard p1>**'s ranking in the 2017 edition of Best Colleges is National Universities, 2. Its tuition and fees are \$47,074 (2016-17).

In addition to the College, Harvard is made up of 13 other schools and institutes, including the top-ranked Business School and Medical School and the highly ranked Graduate Education School, School of Engineering and Applied Sciences, Law School and John F. Kennedy School of Government. Eight U.S. presidents graduated from Harvard College, including Franklin **<n Delano n>** Roosevelt and **<n John F. Kennedy n>**.

After labeling all text files, we wrote a script for IE that implements the hand-crafted rules, which use regexes and a dictionary-based method to extract labeled mentions, and feature information contained. All our code for this project can be found at: <https://github.com/samful1994/cs838webpage/tree/master/code>

### 3 About the Data Set (Mentions)

In our data set, we have entirely 1322 instances (i.e. mentions or data points). In addition, we have 931 mentions, which were extracted from 206 documents from our text files in set I (we call it dev set in the following), and 391 mentions, which were extracted from 102 documents from the text files in set J (we call it test set in the following).

## 4 Performance on Different Models and Model Selection

### 4.1 How We Implement Machine Learning

For dev set, the training set is shuffled at the beginning, instead of implementing simple cross validation, we implement ***k*-fold cross validation**, i.e. split the whole training set into  $k$  subsets, and then use  $k - 1$  of  $k$  subsets of training set to train the model, and test the trained model on the remaining set. Under this case, we trained  $k$  models for each ML algorithm, and we select the one with highest cross validation accuracy.

We tried following ML models:

- **SVM** with RBF kernel
- **Decision Tree** using CART Algorithm
- **Random Forest**
- **Logistic Regression**
- **Linear Regression** (set 0.5 as threshold, with no  $l_1$  (LASSO) or  $l_2$  (Ridge) regularization)
- **Neural Network** with one hidden layer, which contains 15 hidden units with an extra bias unit. We use SGD method with batch size of 64 data points, and constant learning rate ( $\gamma = 0.001$ )

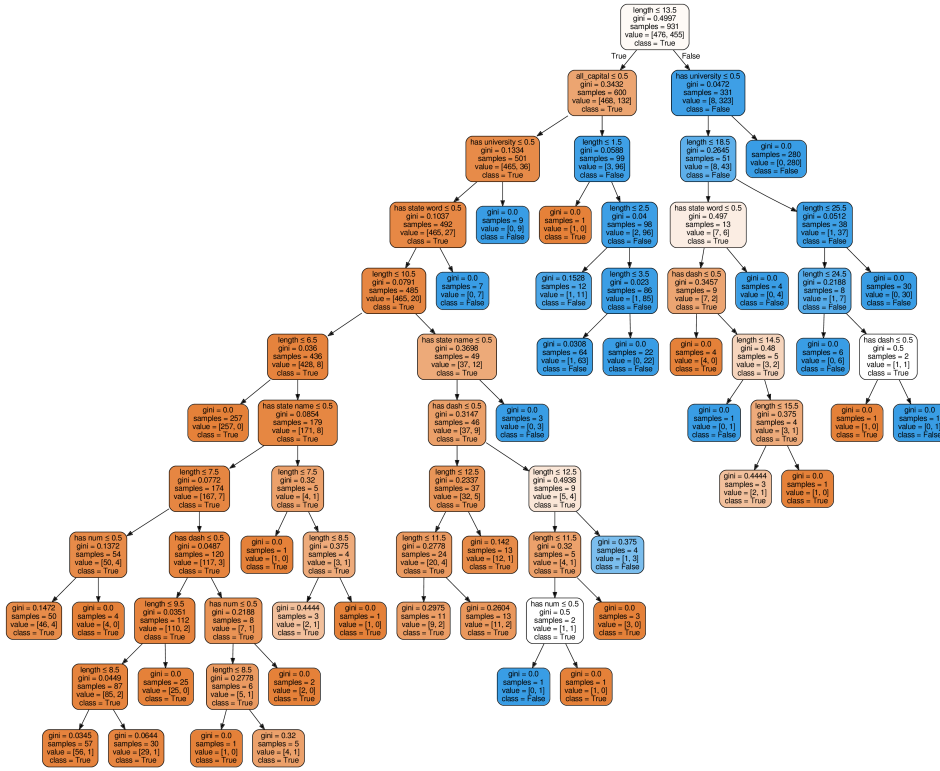


Figure 1: Trained Decision Tree

To give a result of trained models, it's relatively easy to plot the trained decision tree, while it's quite hard to visualize other trained models. Therefore, we give visualized our trained decision tree in Fig1. The results on different models are given below, also visualized in Fig2:

Evaluations Among ML Models			
ML Models	Precision	Recall	F_1 Score
SVM	0.9648	0.9357	0.9500
Decision Tree	0.9649	0.9406	0.9526
Random Forest	0.9739	0.9455	0.9595
Logistic Regression	0.9546	0.9369	0.9457
Linear Regression	0.9647	0.9307	0.9474
Neural Network	0.9649	0.9405	0.9525

Due to the result, obviously the result from **random forest** is the best. Thus we select the result from random forest.

## 4.2 Any Rule-based Post-processing?

Clearly, since the foregoing results are nearly "perfect" we do not need any rule-based post-processing for our information extraction work.

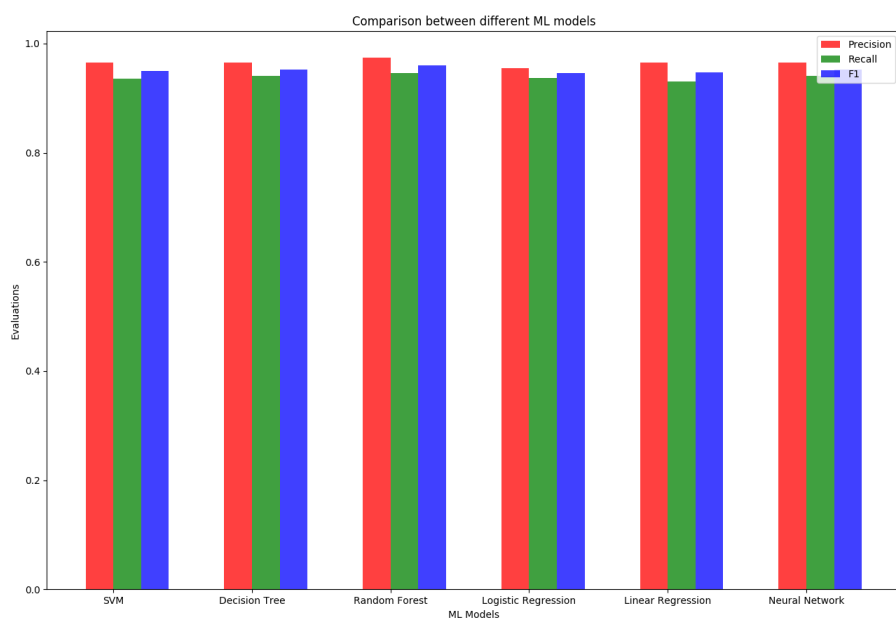


Figure 2: Comparison among ML Models