Project 4 - Author Name Disambiguation

Marie Schiltz

4/7/2017

In this report we will study different methods of **Author Name Disambiguation**. It's the problem of determining whether records in a publications database refer to the same person.

There are two major challenges in author name disambiguation, synonyms and homonyms. In this project we focuses on the second challenge.

We will use domain specific knowledge such as co-aurthors, title of publications and title of journals to perform this task.

The goal of this report is to implement and compare two scientific publications. (Paper 2) Two supervised learning approaches for name disambiguation in author citations (Han et al. [2004]) - we will study the SVM part of this paper (Paper 5) Author disambiguation using error-driven machine learning with a ranking loss function (Culotta et al. [2007]) - we will study the C/E/Pc part os this paper Those two papers can be found in the repository under doc/papers

Step 0: Load Pakages and Functions

```
packages.used=c("stringr", "tex2vec", "caret", "gmum.r", "e1071", "plyr", "tidyr", "gridExtra")

# Check packages that need to be installed.
packages.needed=setdiff(packages.used, intersect(installed.packages()[,1],
                                                 packages.used))

# Install packages
if(length(packages.needed)>0){
  install.packages(packages.needed, dependencies = TRUE,
                   repos='http://cran.us.r-project.org')
}
```

```
## Warning: package 'tex2vec' is not available (for R version 3.3.2)
```

```
# Load packages
library(stringr)
library(text2vec)
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(gmum.r)
library(e1071)
library(plyr)
library(tidyr)
library(gridExtra)

# Source Functions
source("../lib/helper_load.R")
source("../lib/helper_model.R")
```

Step 1: Load and Process data

The dataset is downloaded from http://clgiles.ist.psu.edu/data/ - There are 14 .txt files in the data folder. Each file is a collection of ambiguous names and associtated citations. e.g. AGupta.txt is the citation files of 26 “A. Gupta”s. The 14 canonical names are top ranked ambiguous names, such as “J. Lee”, “J Smith”, “S. Lee”and “Y. Chen” from the DBLP bibliography. - The datasets are pre-processed as follows. All the author names in the citations were simpli- fied to first name initial and last name. For example, “Yong-Jik Kim” was simplified to “Y. Kim”. A reason for such simplification is that the first name initial and last name format is popular in citation records. Publication dates are eliminate from citations. - All citations in the raw data are in the format of clusterid citationid authors;authors;…<>paper title<>publication venue title, where clusterid indicates the canonical author id.

```
### Dataset - Extract Files' Names
files <- list.files(path = "../data/nameset", pattern = "*.txt")
files <- substr(files, 1, nchar(files)-4)
# Create list of auhtors with good format "F Lastname"
authors <- paste(substring(files, 1, 1), " ",
                 substring(files, 2, nchar(files)), sep="")

### Upload & Clean Dataset
# Format: Nested List - Upper Level: per homonym - Lower Level: one list per record
# Initialize dataset
dataset <- list(length(files))
for(i in 1:length(files)){
  temp <- as.list(readLines(paste0("../data/nameset/", files[i],".txt")))
  dataset[[i]] <- lapply(temp, clean.record, author=authors[i])
}
```

Step 2: Implement Paper 2

Brief Description of the Paper This paper is using both SVM and Naive Bayes to perform Name Disambiguation. We will be studying the SVM part.
The author consider each individual author has a class, and perform multiclass one-versus-all SVM to separate the differetn classes. The feeatures used are multiples. They first consider the co-authors, then the title of the paper and finally the journal. Additionally they test an hybrid method using all those features.
We will reproduce those tests on the datasets that we have.

Feature Creation The very first step is to process the clean data and to extract the necessary features. At first, we will conduct the evaluation of the paper on the first dataset only (the name set is A Gupta)

We will process the data in a dataframe.

```
# Change Format - Data Frame
df <- data.frame(matrix(unlist(dataset[[1]]),
                        nrow=length(dataset[[1]]), byrow=T),
                 stringsAsFactors=FALSE)

# Update columns' names
names(df) <- c("author.id", "paper.id", "coauthors", "paper", "journal")
```

Let’s first create a vocabulary-based DTM. Here we collect unique terms from all records and mark each of them with a unique ID using the `create_vocabulary()` function. We use an iterator to create the vocabulary.

```
it <- itoken(df$paper,
             preprocessor = tolower,
             tokenizer = word_tokenizer,
```

```r
             ids = df$paper.id,
             # turn off progressbar because it won't look nice in rmd
             progressbar = FALSE)
vocab <- create_vocabulary(it, stopwords = c("a", "an", "the", "in", "on",
                                             "at", "of", "above", "under"))
```

Here, we remove pre-defined stopwords, the words like "a", "the", "in", "I", "you", "on", etc, which do not provide much useful information.

Before processing the features, we need to split the training and the testing sets. Each author is considered as a different class. To unsure balanceness, we will split data per class. We will put 80% of each record per author (not per name set) and put it on the training set.

```r
# Split training & testing set
df$author.id <- factor(df$author.id)
set.seed(123) # for reproducibility
inTrain <- createDataPartition(df$author.id, p=0.8, list=FALSE)
df.train <- df[inTrain,]
df.test <- df[-inTrain,]
```

Now that we have a vocabulary list and specific train and test sets, we can construct document-term matrices.

```r
vectorizer <- vocab_vectorizer(vocab)
# Train set
it.train <- itoken(df.train$paper,
             preprocessor = tolower,
             tokenizer = word_tokenizer,
             ids = df.train$paper.id,
             # turn off progressbar because it won't look nice in rmd
             progressbar = FALSE)
dtm.train <- create_dtm(it.train, vectorizer)

# Test set
it.test <- itoken(df.test$paper,
             preprocessor = tolower,
             tokenizer = word_tokenizer,
             ids = df.test$paper.id,
             # turn off progressbar because it won't look nice in rmd
             progressbar = FALSE)
dtm.test <- create_dtm(it.test, vectorizer)
```

Now we have DTM and can check its dimensions.

```r
dim(dtm.train)
```

```
## [1]  474 1261
```

As you can see, the DTM has 474 rows, equal to the number of citations, and 1261, equal to the number of unique terms excluding stopwords.

It's always easier to work with dataframe to pass them to a machine learning function, so we'll change the format of the document-term matrices.

```r
dtm.test <- as.data.frame(as.matrix(dtm.test))
dtm.train <- as.data.frame(as.matrix(dtm.train))
# Add labels to the trainnig set
dtm.train <- cbind(df.train$author.id, dtm.train)
names(dtm.train)[1] <- "author.id"
```

SVM Evaluation

The paper uses a SVM classifier to differentiate the differetn homonyms. They use the scheme "one-versus-all". We will evaluate this version and also the "one-versus-one version". We will also try different type of SVM classifier and tune the parameters by using cross-validation.

Note that we will scale the data and we are using the default value of epsilon (0.1) which ensures that if a feature is not seen in the training set, it will not be associated a probability of 0 when discovered in the test set. [TO DO - More specific explanation of this point]

One-Versus-All Method

```
# Fit Linear SVM
start.time <- Sys.time()
svm.linear.all <- cv.svm.all(dtm.train, K=5)
end.time <- Sys.time()
time.linear.all <- end.time - start.time

svm.linear.all$best.performance
```

```
## [1] 0.2696245
```

```
svm.linear.all$best.parameter
```

```
## [1] 10
```

One-Versus-One Method We use the e1071 library, this library implemetns the one-versus-one method for SVM. The tune function uses by default 10-folds cross validation.

```
# Fit Linear SVM
start.time <- Sys.time()
tune.svm.linear <- tune(svm, author.id~.,
                        data=dtm.train, kernel="linear",
end.time <- Sys.time()
time.linear.one <- end.time - start.time

tune.svm.linear$best.performance
```

```
## [1] 0.2844858
```

```
tune.svm.linear$best.parameters
```

```
##   cost
## 2    1
```

```
# Fit Radial SVM
start.time <- Sys.time()
tune.svm.radial <- tune(svm, author.id~., data=dtm.train, kernel="radial",
                        ranges=list(cost=c(0.1,1,10,100,1000),
                                    gamma=c(0.5,1,2,3,4)))
end.time <- Sys.time()
time.radial.one <- end.time - start.time

tune.svm.radial$best.performance
```

```
## [1] 0.5821365
```

```
tune.svm.radial$best.parameters
```

```
##   cost gamma
## 3   10   0.5
```

```
print("Summary of the running times")
```

```
## [1] "Summary of the running times"
```

```
print("SVM - One-versus-all - Linear")
```

```
## [1] "SVM - One-versus-all - Linear"
```

```
time.linear.all
```

```
## Time difference of 25.30137 secs
```

```
print("SVM - One-versus-one - Linear")
```

```
## [1] "SVM - One-versus-one - Linear"
```

```
time.linear.one
```

```
## Time difference of 40.99126 secs
```

```
print("SVM - One-versus-one - Radial")
```

```
## [1] "SVM - One-versus-one - Radial"
```

```
time.radial.one
```

```
## Time difference of 3.458169 mins
```

The one versus all scheme is perfomring better, it has both a more efficient running time and a lower crosss validation error rate. Hence we will be keeping this method. Given the running time and the result of the radial kernel we didn't bother trying it with the one versus all shceme. Given the results of this analysis, we will be using the one versus all scheme with a linear kernel. We will keep using this scheme until the end of the analysis but we will adapt the parameter C for each name set (using cross-validation).

Run model on all datasets for all attributes

For more clarity, we put everything into one function `run.svm` which take into parameters the id of one name set, the dataset, and the attibute you want to use.

```
results <- data.frame(paper=double(),
                      journal=double(),
                      coauthors=double(),
                      hybrid=double())
for (attribute in c("paper", "coauthors", "journal", "hybrid")){
  for (name_set in 1:length(dataset)){
    results[name_set, attribute] <- run.svm(name_set, attribute, dataset)
  }
}
```

Now that we have the results for all dataset, let's draw a table of those results.

```
table <- results
table <- lapply(table[,], round, 3)
table <- as.data.frame(table)
table <- cbind(authors, table)
table
```

```
##       authors paper journal coauthors hybrid
## 1     A Gupta 0.825   0.650     0.883  0.922
## 2     A Kumar 0.837   0.767     0.721  0.907
## 3      C Chen 0.707   0.564     0.699  0.880
## 4   D Johnson 0.897   0.765     0.735  1.000
## 5        J Lee 0.734   0.512     0.709   0.857
## 6     J Martin 0.786   0.786     0.714   1.000
## 7   J Robinson 0.793   0.724     0.931   0.931
## 8      J Smith 0.839   0.759     0.672   0.920
## 9     K Tanaka 0.846   0.673     0.865   0.981
## 10     M Brown 0.958   0.750     0.833   0.917
## 11     M Jones 0.778   0.800     0.733   0.911
## 12    M Miller 0.885   0.859     0.910   0.962
## 13       S Lee 0.739   0.609     0.715   0.893
## 14      Y Chen 0.780   0.619     0.852   0.951
```

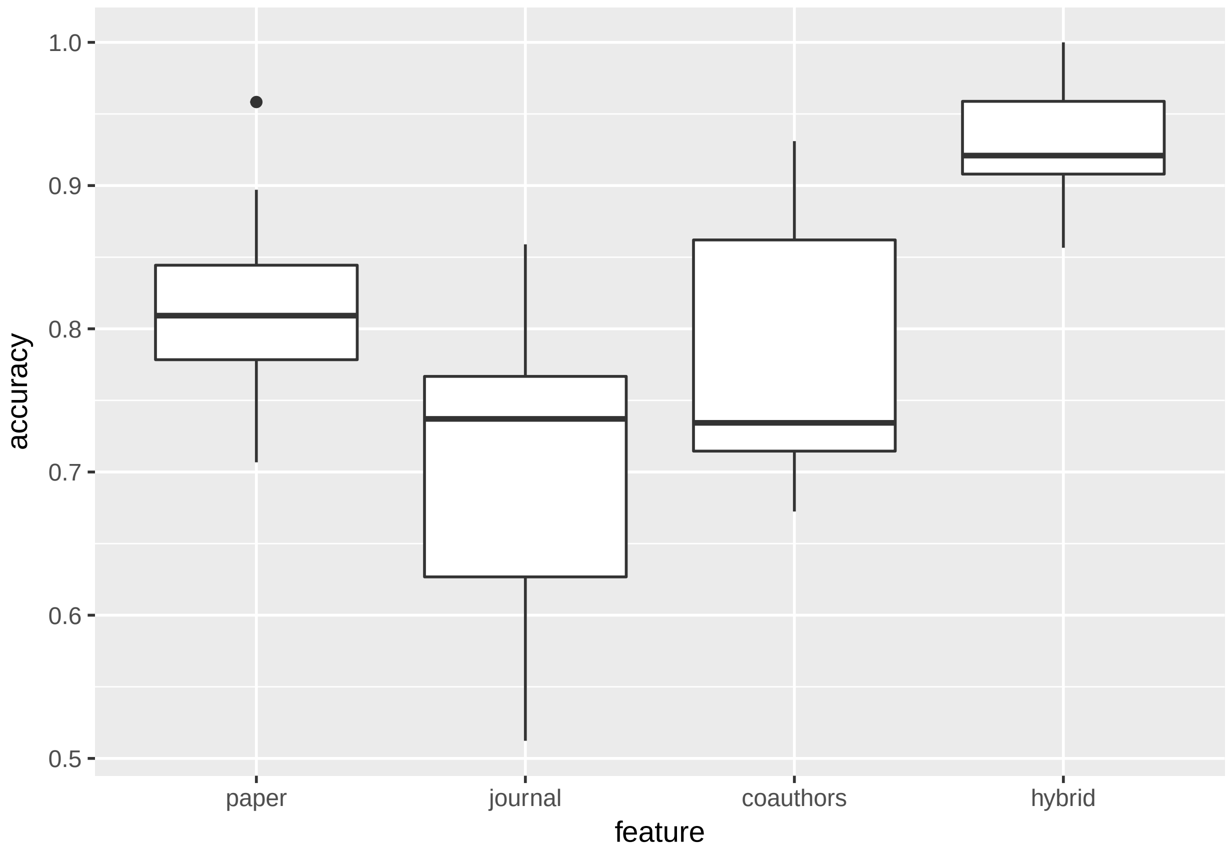
```
png(filename = "../output/svm_results.png", width = 400, height = 150, units = "mm", res=200)
grid.table(table, rows=NULL)
dev.off()
```

```
## pdf
##   2
```

We can also display a box plot, showing the accuracy depending on the features used.

```
results_long <- gather(results, feature, accuracy, paper:hybrid, factor_key=TRUE)

boxplot.svm <- ggplot(results_long, aes(x = feature, y = accuracy)) +
  geom_boxplot()
boxplot.svm
```



```
png(filename = "../output/svm_boxplot.png", width = 300, height = 150, units = "mm", res=200)
boxplot.svm
dev.off()
```

```
## pdf
##   2
```

When analysing those results, we have to keep in mind that the datasets have different sizes.

```
info <- data.frame(author=character(),
                      variations=double(),
                      train=double(),
                      test=double())
for (i in 1:length(authors)){
  info[i, "variations"] <- length(dataset[[i]])
  info[i, "train"] <- round(length(dataset[[i]])*0.80)
  info[i, "test"] <- length(dataset[[i]])-round(length(dataset[[i]])*0.80)
}

info$author <- authors
info
```

```
##         author variations train test
## 1      A Gupta        577   462  115
## 2      A Kumar        244   195   49
## 3       C Chen        801   641  160
## 4    D Johnson        368   294   74
## 5        J Lee       1419  1135  284
## 6     J Martin        112    90   22
## 7   J Robinson        171   137   34
## 8      J Smith        927   742  185
## 9     K Tanaka        280   224   56
## 10     M Brown        153   122   31
## 11     M Jones        260   208   52
## 12    M Miller        412   330   82
## 13       S Lee       1464  1171  293
## 14      Y Chen       1265  1012  253
```

```
png(filename = "../output/dataset_info.png", width = 400, height = 150, units = "mm", res=200)
grid.table(info, rows=NULL)
dev.off()
```

```
## pdf
##   2
```

Comments on the results Anaylizing the boxplot, we see that the hybrid method gets the best results. This methods takes as features, the title of the journal, the coauthors and the title of the paper. If you want a faster method by using only one of the features, then the title of the paper is the best choice.

Overall Comments on the Paper This paper is easily reproducible. All steps are clearly stated in the paper. The features and the model used are entirely explained.

Step 3: Implement Paper 5

Brief Description of the Paper This paper is using a completely different method. It is based on clustering and error-driven learning. The authors are using a clusterig based method, but instead of using a completely unsupervised learning they used the groundtruth for each record to update teh weight of the parameters. That's the reason why this method is called "Error-Driven".

Step 4: Comparison of the two methods