Text Processing with RTextTools

Karen Mazidi

The RTextTools package integrates text processing and machine learning. Read more about it in this paper. We are going to look at using RTextTools for processing the amazon reviews data.

Load the data

```
library(RTextTools)
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
reviews <- read.csv("data/reviews.csv", header=TRUE, stringsAsFactors=F)
```

Create a document term matrix

This uses the tm package under the hood.

```
dtm <- create_matrix(reviews$Review, language="english", removeNumbers=TRUE,
                     removeStopwords=TRUE, stemWords=TRUE, removeSparseTerms=.998)
```

Create a container

The container will hold train and test observations as well as labels.

```
container <- create_container(dtm, reviews$Rating, trainSize=1:3000,
                              testSize=3001:4139, virgin=FALSE)
```

Train model

There are several algorithms to choose from, we just selected 3 of them.

```
svm <- train_model(container, "SVM")
glmnet <- train_model(container, "GLMNET")
maxent <- train_model(container, "MAXENT")
```

Classify

Now apply the models to the test data.

```
svm_classify <- classify_model(container, svm)
glmnet_classify <- classify_model(container, glmnet)
maxent_classify <- classify_model(container, maxent)
```

Analytics

Interpreting the results.

```
analytics <- create_analytics(container, cbind(
      svm_classify, glmnet_classify, maxent_classify))
summary(analytics)
```

```
## ENSEMBLE SUMMARY
##
##        n-ENSEMBLE COVERAGE n-ENSEMBLE RECALL
## n >= 1                1.00              0.83
## n >= 2                1.00              0.83
## n >= 3                0.78              0.89
##
##
## ALGORITHM PERFORMANCE
##
##        SVM_PRECISION           SVM_RECALL           SVM_FSCORE
##                0.820                0.815                0.815
##     GLMNET_PRECISION        GLMNET_RECALL        GLMNET_FSCORE
##                0.815                0.810                0.810
## MAXENTROPY_PRECISION    MAXENTROPY_RECALL    MAXENTROPY_FSCORE
##                0.805                0.805                0.800
```

Create ensemble aggreement

Calculate coverage, the percentage of cases on which the n cases agree, for n >= 1, 2, 3 models.

```
create_ensembleSummary(analytics@document_summary)
```

```
##        n-ENSEMBLE COVERAGE n-ENSEMBLE RECALL
## n >= 1                1.00              0.83
## n >= 2                1.00              0.83
## n >= 3                0.78              0.89
```