Auto Tagging Chapters on CMS

Pritesh Shrivastava

March 3, 2018

Reading data from CMS

```
## # A tibble: 6 x 10
##         Topic_Code              Chapter        Topic `Q Size` `Sol Size`
##              <chr>                <chr>        <chr>    <int>      <int>
## 1 MTH-12-JEE-18-00 Inverse Trigonometry Introduction      131        336
## 2 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       81        497
## 3 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       79        349
## 4 MTH-12-JEE-18-00 Inverse Trigonometry Introduction      116       1006
## 5 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       67        337
## 6 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       74        327
## # ... with 5 more variables: Difficulty <int>, Code <chr>, Status <chr>,
## #   `Problem ID` <chr>, Text <chr>
```

Cleaning and adding Grade, Subject, Curriculum and Chapter No

```
## # A tibble: 6 x 14
##         Topic_Code              Chapter        Topic `Q Size` `Sol Size`
##              <chr>                <chr>        <chr>    <int>      <int>
## 1 MTH-12-JEE-18-00 Inverse Trigonometry Introduction      131        336
## 2 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       81        497
## 3 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       79        349
## 4 MTH-12-JEE-18-00 Inverse Trigonometry Introduction      116       1006
## 5 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       67        337
## 6 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       74        327
## # ... with 9 more variables: Difficulty <int>, Code <chr>, Status <chr>,
## #   `Problem ID` <chr>, Text <list>, Grade <chr>, Subject <chr>,
## #   Curriculum <chr>, Ch_No <chr>
```

Summarizing Chapter wise # Qs in the entire JEE dataset

```
## # A tibble: 19 x 2
##                              Chapter count_qs
##                                <chr>    <int>
##  1           3 Dimensional Geometry       72
##  2      Applications of Derivatives      105
##  3   Bridge Intervention Curriculum        2
##  4               Conic Sections - I      127
##  5              Conic Sections - II      179
##  6 Continuity and Differentiability      120
##  7             Definite Integration      140
##  8           Differential Equations       52
##  9                       Functions 2       98
## 10      Fundamentals of Mathematics       90
## 11           Indefinite Integration       54
```

```
## 12             Inverse Trigonometry       71
## 13           Mathematical Reasoning       29
## 14        Matrices and Determinants      103
## 15                      Probability      101
## 16                   Selection Test       20
## 17              Sequence and Series      117
## 18    Sets, Relations and Functions       85
## 19                   Vector Algebra       98
```

Picking 2 chapters for classification

```
## # A tibble: 2 x 2
##                      Chapter count_qs
##                        <chr>    <int>
## 1 Matrices and Determinants      103
## 2               Probability      101
```

Changing Chapter labels to factors

The Chapter variable is currently a character vector. Since this is a categorical variable, it would be better to convert it to a factor.

```
##  Factor w/ 2 levels "Matrices and Determinants",..: 2 2 2 2 2 2 2 2 2 2 ...
```

Summary of test set

```
## # A tibble: 2 x 2
##                      Chapter count_qs
##                       <fctr>    <int>
## 1 Matrices and Determinants      103
## 2               Probability      101
```

Create text corpus

Let's view some Qs

```
## $`2`
## [1] "There are   n   houses are available in a locality which are applied by   n   people  Each appli
##
## $`3`
## [1] "The probability of event   A   occurring is   0 5   and event   B   occurring is   0 3    If   A
##
## $`4`
## [1] "The probability that at least one of the two events   A   and   B    occurs is   0 6    If   A
```

Cleaning text

Removing punctuations, numbers and stop words Converting to lower case Stemming words - learned, learning, and learns are transformed into the base form, learn Removing additional white spaces

Let's view some cleaned Qs

```
## $`2`
## [1] "n hous avail local appli n peopl appli one hous without consult other probabl appli hous"
##
## $`3`
## [1] "probabl event occur event b occur b mutual exclus event probabl neither b occur"
##
## $`4`
## [1] "probabl least one two event b occur b occur simultan probabl evalu p p b"
```

Bag of words - Tokenization

```
## <<DocumentTermMatrix (documents: 204, terms: 495)>>
## Non-/sparse entries: 2237/98743
## Sparsity           : 98%
## Maximal term length: 11
## Weighting          : term frequency (tf)
```

Data preparation - Creating training and test datasets

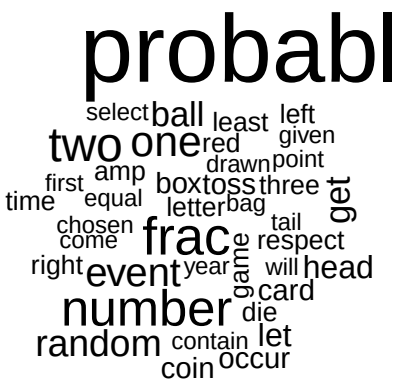
```
## text_train_labels
## Matrices and Determinants              Probability
##                   0.50625                  0.49375
```

Distribution of labels in test set

```
## text_test_labels
## Matrices and Determinants              Probability
##                       0.5                      0.5
```

Visualizing text data - word clouds

Word Cloud from Probability chapter



Word cloud from Matrices chapter

Words appearing at least a specified number of times.Filter our DTM to include only the terms appearing in a specified vector.

The Naive Bayes classifier is typically trained on data with categorical features. This poses a problem, since the cells in the sparse matrix are numeric and measure the number of times a word appears in a message. We need to change this to a categorical variable that simply indicates yes or no depending on whether the word appears at all. Train matrix :

```
##  chr [1:160, 1:69] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" ...
##  - attr(*, "dimnames")=List of 2
##   ..$ Docs : chr [1:160] "1" "4" "5" "6" ...
##   ..$ Terms: chr [1:69] "adj" "alpha" "amp" "bag" ...
```

Test matrix :

```
##  chr [1:44, 1:69] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" ...
##  - attr(*, "dimnames")=List of 2
##   ..$ Docs : chr [1:44] "2" "3" "7" "8" ...
##   ..$ Terms: chr [1:69] "adj" "alpha" "amp" "bag" ...
```

TRAIN NAIVE BAYES MODEL

PREDICTION

```
## Confusion Matrix and Statistics
```

```
##
##                           Reference
## Prediction                 Matrices and Determinants Probability
##   Matrices and Determinants                        22           0
##   Probability                                       0          22
##
##                Accuracy : 1
##                  95% CI : (0.9196, 1)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 5.684e-14
##
##                   Kappa : 1
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0
##             Specificity : 1.0
##          Pos Pred Value : 1.0
##          Neg Pred Value : 1.0
##              Prevalence : 0.5
##          Detection Rate : 0.5
##    Detection Prevalence : 0.5
##       Balanced Accuracy : 1.0
##
##        'Positive' Class : Matrices and Determinants
##
```

Sources:

[1] tm, e1071 - http://blog.thedigitalgroup.com/rajendras/2015/05/28/supervised-learning-for-text-classification/

[2] tm, e1071, wordcloud - https://rstudio-pubs-static.s3.amazonaws.com/194717_4639802819a342eaa274067c9dbb657e.html