De-duplicating CMS repositry

Pritesh Shrivastava

January 22, 2018

Reading data from CMS

```
## Warning: package 'tidyverse' was built under R version 3.4.3

## -- Attaching packages ----------------------------------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.3.4     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.3.0
## v readr   1.1.1     v forcats 0.2.0

## Warning: package 'tidyr' was built under R version 3.4.3

## Warning: package 'readr' was built under R version 3.4.3

## Warning: package 'purrr' was built under R version 3.4.3

## Warning: package 'dplyr' was built under R version 3.4.3

## Warning: package 'forcats' was built under R version 3.4.3

## -- Conflicts -------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Attaching package: 'SparseM'

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

## Warning: package 'tm' was built under R version 3.4.3

## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate

## Warning: package 'caret' was built under R version 3.4.3

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## Warning: Missing column names filled in: 'X15' [15], 'X16' [16],
## 'X17' [17], 'X18' [18], 'X19' [19]
```

```
## Parsed with column specification:
## cols(
##   Topic_Code = col_character(),
##   Chapter = col_character(),
##   Topic = col_character(),
##   `Q Size` = col_character(),
##   `Sol Size` = col_character(),
##   Difficulty = col_integer(),
##   Code = col_character(),
##   Status = col_character(),
##   `Problem ID` = col_character(),
##   Text = col_character(),
##   A = col_character(),
##   B = col_character(),
##   C = col_character(),
##   D = col_character(),
##   X15 = col_character(),
##   X16 = col_character(),
##   X17 = col_character(),
##   X18 = col_character(),
##   X19 = col_character()
## )
```

```
## # A tibble: 6 x 19
##         Topic_Code              Chapter        Topic `Q Size` `Sol Size`
##              <chr>                <chr>        <chr>    <chr>      <chr>
## 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 14 more variables: Difficulty <int>, Code <chr>, Status <chr>,
## #   `Problem ID` <chr>, Text <chr>, A <chr>, B <chr>, C <chr>, D <chr>,
## #   X15 <chr>, X16 <chr>, X17 <chr>, X18 <chr>, X19 <chr>
```

Adding Grade, Subject, Curriculum and Chapter No

```
## # A tibble: 6 x 19
##         Topic_Code              Chapter        Topic `Q Size` `Sol Size`
##              <chr>                <chr>        <chr>    <chr>      <chr>
## 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 14 more variables: Difficulty <int>, Code <chr>, Status <chr>,
## #   `Problem ID` <chr>, Text <chr>, A <chr>, B <chr>, C <chr>, D <chr>,
## #   Grade <chr>, Subject <chr>, Curriculum <chr>, Ch_No <chr>,
## #   fullText <chr>
```

Creating vocabulary by tokenizing text - using text2vec

```
## Warning: package 'text2vec' was built under R version 3.4.3
```

```
##
## Attaching package: 'text2vec'
```

```
## The following object is masked from 'package:dplyr':
##
##     collect
```

Creating Document Term Matices - text2vec

```
## [1] 7603 2948
```

Applying tf-idf on dtm

```
## 5 x 5 sparse Matrix of class "dgCMatrix"
##   servant modular blood rise send
## 1       .       .     .    .    .
## 2       .       .     .    .    .
## 3       .       .     .    .    .
## 4       .       .     .    .    .
## 5       .       .     .    .    .
```

Cosine similarities between all rows of dtm_tfidf matrix

```
## 6 x 6 sparse Matrix of class "dsCMatrix"
##           1          2         3         4          5          6
## 1 1.0000000 0.21361347 0.2177698 0.35203835 0.19269810 0.36768955
## 2 0.2136135 1.00000000 .         0.07972315 0.11576713 .
## 3 0.2177698 .          1.0000000 0.39865275 0.67205308 0.16566470
## 4 0.3520383 0.07972315 0.3986527 1.00000000 0.51369938 0.19562040
## 5 0.1926981 0.11576713 0.6720531 0.51369938 1.00000000 0.09468772
## 6 0.3676895 .          0.1656647 0.19562040 0.09468772 1.00000000
```

Cosine similarity with Latent Semantic Analysis

Usually tf-idf/bag-of-words matrices contain a lot of noise. Applying LSA model can help with this problem, so you can achieve better quality similarities

```
##           1           2          3          4          5           6
## 1 1.0000000  0.35854123  0.1861758 0.83493359 0.33287054  0.88858241
## 2 0.3585412  1.00000000 -0.1088990 0.03600821 0.08588808 -0.07890175
## 3 0.1861758 -0.10889901  1.0000000 0.64927613 0.97716351  0.23808092
## 4 0.8349336  0.03600821  0.6492761 1.00000000 0.72057118  0.86088282
## 5 0.3328705  0.08588808  0.9771635 0.72057118 1.00000000  0.30361609
## 6 0.8885824 -0.07890175  0.2380809 0.86088282 0.30361609  1.00000000
```

Tidying similarity matrix

```
##   Var1 Var2      Freq
## 1    1    1 1.0000000
## 2    2    1 0.3585412
## 3    3    1 0.1861758
## 4    4    1 0.8349336
## 5    5    1 0.3328705
## 6    6    1 0.8885824
```

Filtering near duplicates

```
##    row column Similarity_Measure
## 1   14      2                  1
## 2  453      4                  1
## 3 2964     24                  1
## 4   60     43                  1
## 5 2953     52                  1
## 6  428     77                  1
```

Final list of Duplicate Qs

```
##    row_id
## 1 P006292
## 2 P057177
## 3 P022392
## 4 P044860
## 5 P017252
##
## 1                                                                                    If \\(\\cos^{-
## 2                                        If \\(f\\left(x\\right)=2\\:\\tan^{-1}x+\\sin^{-1}\\lef
## 3
## 4 Let \\(\\tan^{-1}y=\\tan^{-1}x+\\tan^{-1}\\left(\\frac{2x}{1-x^2}\\right)\\), where\\(\\left|x\\ri
## 5                                                                        \\(\\sin\\left[2\\cos^{-1}\
##   row_status  col_id
## 1      final P003479
## 2      final P005140
## 3      final P009145
## 4      final P027896
## 5      final P031390
```