# 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 1.2.0
## 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()
## Warning: Missing column names filled in: 'X10' [10]
## Parsed with column specification:
## cols(
##   Topic_Code = col_character(),
##   Chapter = col_character(),
##   Topic = col_character(),
##   `Q Size` = col_integer(),
##   `Sol Size` = col_integer(),
##   Difficulty = col_integer(),
##   Code = col_character(),
##   Status = col_character(),
##   Text = col_character(),
##   X10 = col_character()
## )
## # A tibble: 6 x 10
##         Topic_Code              Chapter        Topic `Q Size` `Sol Size`
##              <chr>                <chr>        <chr>    <int>      <int>
## 1             <NA>                 <NA>         <NA>       NA         NA
## 2 MTH-12-JEE-18-00 Inverse Trigonometry Introduction      131        336
## 3 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       81        497
## 4 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       79        349
## 5 MTH-12-JEE-18-00 Inverse Trigonometry Introduction      116       1006
## 6 MTH-12-JEE-18-00 Inverse Trigonometry Introduction       67        337
## # ... with 5 more variables: Difficulty <int>, Code <chr>, Status <chr>,
## #   Text <chr>, X10 <chr>
```

**Cleaning and adding Grade, Subject, Curriculum and Chapter No**

```
## # A tibble: 6 x 13
##         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 8 more variables: Difficulty <int>, Code <chr>, Status <chr>,
## #   Text <chr>, Grade <chr>, Subject <chr>, Curriculum <chr>, Ch_No <chr>
```

## Creating vocabulary by tokenizing text

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

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

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

```
## Number of docs: 232
## 0 stopwords:  ...
## ngram_min = 1; ngram_max = 1
## Vocabulary:
##                 term term_count doc_count
## 1:          opposite          3         3
## 2:              like          3         3
## 3:               out          3         3
## 4:         developed          3         3
## 5: roots.</li></ol>          3         3
## 6:         <p>leaves          3         3
```

## Creating Document Term Matices

```
## [1] 232 280
```

**Cosine similarity with tf-idf**

```
## 5 x 5 sparse Matrix of class "dgCMatrix"
##   opposite like out developed roots.</li></ol>
## 1        .    .   .         .                .
## 2        .    .   .         .                .
## 3        .    .   .         .                .
## 4        .    .   .         .                .
## 5        .    .   .         .                .
```

Calculate similarities between all rows of dtm_tfidf matrix

```
## 6 x 6 sparse Matrix of class "dsCMatrix"
##   1 2 3 4 5 6
## 1 1 . . . . .
## 2 . 1 . . . .
## 3 . . 1 . . .
## 4 . . . 1 . .
## 5 . . . . 1 .
## 6 . . . . . 1
```

## 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

Calculate similarities between all rows of dtm_tfidf_lsa matrix

```
##           1         2         3         4         5         6
## 1 1.0000000 0.7095634 0.9308843 0.7645943 0.8621513 0.6194950
## 2 0.7095634 1.0000000 0.5825346 0.9352255 0.9242858 0.8225262
## 3 0.9308843 0.5825346 1.0000000 0.5526154 0.6796274 0.3541298
## 4 0.7645943 0.9352255 0.5526154 1.0000000 0.9737433 0.9280813
## 5 0.8621513 0.9242858 0.6796274 0.9737433 1.0000000 0.8919710
## 6 0.6194950 0.8225262 0.3541298 0.9280813 0.8919710 1.0000000
```

### Tidying similarity matrix

```
##   Var1 Var2      Freq
## 1    1    1 1.0000000
## 2    2    1 0.7095634
## 3    3    1 0.9308843
## 4    4    1 0.7645943
## 5    5    1 0.8621513
## 6    6    1 0.6194950
```

### Filtering near duplicates

```
##   row column       sim
## 1  86      9 1.0000000
## 2  87     10 1.0000000
## 3  88     11 1.0000000
## 4  89     12 0.9999981
## 5  90     13 1.0000000
## 6  91     14 1.0000000
```

## Final list of Duplicate Qs

```
##    row_id
## 1 P014720
## 2 P014721
## 3 P014722
## 4 P014723
## 5 P014724
```

```
##
## 1
## 2
## 3
## 4 <p><font face="Fira Sans"><span>Fi</span></font><span style="color: rgb(51, 51, 51); font-style: n
## 5
##    col_id
## 1 P009938
## 2 P009941
## 3 P009943
## 4 P009945
## 5 P009946
##
## 1
## 2
## 3
## 4 <p><font color="#000000"><font face="Fira Sans"><span>Fi</span></font><span style="font-style: nor
## 5
```