Module 3 followup

#Text Mining and TF-IDF in R

```
library(tm)
## Warning: package 'tm' was built under R version 3.0.3
Dir <- "C:/Users/shajaas/Google Drive/Edureka/somefiles"

Documents <- Corpus(DirSource(Dir))
## Warning: incomplete final line found on 'C:/Users/shajaas/Google
Drive/Edureka/somefiles/doc1.txt'
## Warning: incomplete final line found on 'C:/Users/shajaas/Google
Drive/Edureka/somefiles/doc2.txt'
inspect(Documents)
## A corpus with 3 text documents
##
## The metadata consists of 2 tag-value pairs and a data frame
## Available tags are:
##   create_date creator
## Available variables in the data frame are:
##   MetaID
##
## $doc1.txt
## this is my test document
## hello how are you
##
## $doc2.txt
## all is well
## health is wealth
##
## $doc3.txt
## Knowledge is wealth
## I am fine
## this is my test document
## hello how are you
X <- tm_map(Documents, tolower)
inspect(X)
## A corpus with 3 text documents
##
## The metadata consists of 2 tag-value pairs and a data frame
## Available tags are:
##   create_date creator
## Available variables in the data frame are:
##   MetaID
##
## $doc1.txt
## this is my test document
## hello how are you
##
## $doc2.txt
## all is well
## health is wealth
##
## $doc3.txt
## knowledge is wealth
## i am fine
## this is my test document
```

```
## hello how are you
Y <- tm_map(Documents, removeWords, stopwords("english"))
inspect(Y)
## A corpus with 3 text documents
##
## The metadata consists of 2 tag-value pairs and a data frame
## Available tags are:
##   create_date creator
## Available variables in the data frame are:
##   MetaID
##
## $doc1.txt
##    test document
## hello
##
## $doc2.txt
##   well
## health  wealth
##
## $doc3.txt
## Knowledge  wealth
## I  fine
##    test document
## hello
Z <- tm_map(Documents, removeWords, c("fine", "wealth"))
inspect(Z)
## A corpus with 3 text documents
##
## The metadata consists of 2 tag-value pairs and a data frame
## Available tags are:
##   create_date creator
## Available variables in the data frame are:
##   MetaID
##
## $doc1.txt
## this is my test document
## hello how are you
##
## $doc2.txt
## all is well
## health is
##
## $doc3.txt
## Knowledge is
## I am
## this is my test document
## hello how are you
X <- DocumentTermMatrix(Documents)
inspect(X)
## A document-term matrix (3 documents, 13 terms)
##
## Non-/sparse entries: 21/18
## Sparsity           : 46%
## Maximal term length: 9
## Weighting          : term frequency (tf)
##
##           Terms
## Docs       all are document fine health hello how knowledge test this
##   doc1.txt   0   1        1    0      0     1   1         0    1    1
```

```
##   doc2.txt   1   0        0    0      1     0   0         0    0    0
##   doc3.txt   0   1        1    1      0     1   1         1    1    1
##           Terms
## Docs       wealth well you
##   doc1.txt      0    0   1
##   doc2.txt      1    1   0
##   doc3.txt      1    0   1
inspect(TermDocumentMatrix(Documents))
## A term-document matrix (13 terms, 3 documents)
##
## Non-/sparse entries: 21/18
## Sparsity           : 46%
## Maximal term length: 9
## Weighting          : term frequency (tf)
##
##            Docs
## Terms       doc1.txt doc2.txt doc3.txt
##   all               0        1        0
##   are               1        0        1
##   document          1        0        1
##   fine              0        0        1
##   health            0        1        0
##   hello             1        0        1
##   how               1        0        1
##   knowledge         0        0        1
##   test              1        0        1
##   this              1        0        1
##   wealth            0        1        1
##   well              0        1        0
##   you               1        0        1
termFreq(Documents[[1]])
##      are document    hello      how     test     this      you
##        1        1        1        1        1        1        1
## attr(,"class")
## [1] "term_frequency" "integer"
dtm_tfidf <- weightTfIdf(TermDocumentMatrix(Documents))
inspect(dtm_tfidf)
## A term-document matrix (13 terms, 3 documents)
##
## Non-/sparse entries: 21/18
## Sparsity           : 46%
## Maximal term length: 9
## Weighting          : term frequency - inverse document frequency
(normalized) (tf-idf)
##
##            Docs
## Terms       doc1.txt doc2.txt doc3.txt
##   all        0.00000   0.3962   0.0000
##   are        0.08357   0.0000   0.0585
##   document   0.08357   0.0000   0.0585
##   fine       0.00000   0.0000   0.1585
##   health     0.00000   0.3962   0.0000
##   hello      0.08357   0.0000   0.0585
##   how        0.08357   0.0000   0.0585
##   knowledge  0.00000   0.0000   0.1585
##   test       0.08357   0.0000   0.0585
##   this       0.08357   0.0000   0.0585
##   wealth     0.00000   0.1462   0.0585
##   well       0.00000   0.3962   0.0000
```

```
##   you        0.08357   0.0000   0.0585
dissimilarity(Documents[[1]], Documents[[3]], method="cosine")
##          doc1.txt
## doc3.txt   0.1633
summary(Documents)
## A corpus with 3 text documents
##
## The metadata consists of 2 tag-value pairs and a data frame
## Available tags are:
##   create_date creator
## Available variables in the data frame are:
##   MetaID
```