STEP 01. 각 문서의 내용 지정

STEP 02. Document 생성

STEP 03. 말뭉치(Corpus) 생성및 전처리(소문자변경)

STEP 04. TF구하기 (Term Frquency 구하기)

STEP 05. TF-IDF 구하기

[Quiz] 아래 3개의 텍스트 문서에 대해 TF, TF-IDF 를 구해보자.

Lab05_TFIDF

STEP 01. 각 문서의 내용 지정

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 3.4.4
```

```
## Loading required package: NLP
```

```
## Warning: package 'NLP' was built under R version 3.4.1
```

```
doc1 <- "The fox chases the rabbit"
doc2 <- "The rabbit ate the cabbage"
doc3 <- "The fox caught the rabbit"
```

STEP 02. Document 생성

```
doc.list <- c(doc1, doc2, doc3)
n.docs <- length(doc.list)
names(doc.list) <- paste("doc", c(1:n.docs), sep="")
names(doc.list)
```

```
## [1] "doc1" "doc2" "doc3"
```

STEP 03. 말뭉치(Corpus) 생성및 전처리(소문자변경)

- source 의 종류

STEP 01. 각 문서의 내용 지정

STEP 02. Document 생성

STEP 03. 말뭉치(Corpus) 생성및 전처리(소문자변경)

STEP 04. TF구하기 (Term Frquency 구하기)

STEP 05. TF-IDF 구하기

[Quiz] 아래 3개의 텍스트 문서에 대해 TF, TF-IDF 를 구해보자.

- DirSource() : 디렉토리
- DataframeSource() : R데이터 프레임
- VectorSource() : R 벡터
- XMLSource() : XML 파일
- URISource() : URI

- VCorpus : Volatile(메모리에 저장되는) 코퍼스

```
my.corpus <- Corpus(VectorSource(doc.list))
my.corpus <- tm_map(my.corpus, tolower)
inspect(my.corpus)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
##                         doc1                         doc2
##  the fox chases the rabbit the rabbit ate the cabbage
##                         doc3
##  the fox caught the rabbit
```

STEP 04. TF구하기 (Term Frquency 구하기)

```
TDM <- TermDocumentMatrix(my.corpus,
                          control=list(weighting=weightTf))
m <- as.matrix(TDM)
m
```

STEP 01. 각 문서의 내용 지정

STEP 02. Document 생성

STEP 03. 말뭉치(Corpus) 생성및 전처리(소문자변경)

STEP 04. TF구하기 (Term Frquency 구하기)

STEP 05. TF-IDF 구하기

[Quiz] 아래 3개의 텍스트 문서에 대해 TF, TF-IDF 를 구해보자.

```
##          Docs
## Terms     doc1 doc2 doc3
##   chases     1    0    0
##   fox        1    0    1
##   rabbit     1    1    1
##   the        2    2    2
##   ate        0    1    0
##   cabbage    0    1    0
##   caught     0    0    1
```

STEP 05. TF-IDF 구하기

weighting 은 행렬의 원소를 나타내는 인수로서 기본값은 TF를 나타내는 weightTf이다.

TF-IDF : $\text{tfidf}(t, d, D) = tf(t, d) \times idf(t, D)$

- 단어 빈도, tf(t,d): 문서 내에 나타나는 해당 단어의 빈도
- 역문서빈도 : $idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$
  - $|D|$: 전체 문서의 수
  - $|\{d \in D : t \in d\}|$: 단어 t가 포함된 문서의 수

```
TDM <- TermDocumentMatrix(my.corpus,
           control=list(weighting=weightTfIdf))
m <- as.matrix(TDM)
m
```

STEP 01. 각 문서의 내용 지정

STEP 02. Document 생성

STEP 03. 말뭉치(Corpus) 생성및 전처리(소문자변경)

STEP 04. TF구하기 (Term Frquency 구하기)

STEP 05. TF-IDF 구하기

[Quiz] 아래 3개의 텍스트 문서에 대해 TF, TF-IDF 를 구해보자.

```
##          Docs
## Terms          doc1      doc2      doc3
##   chases  0.3169925 0.0000000 0.0000000
##   fox     0.1169925 0.0000000 0.1169925
##   rabbit  0.0000000 0.0000000 0.0000000
##   the     0.0000000 0.0000000 0.0000000
##   ate     0.0000000 0.3169925 0.0000000
##   cabbage 0.0000000 0.3169925 0.0000000
##   caught  0.0000000 0.0000000 0.3169925
```

[Quiz] 아래 3개의 텍스트 문서에 대해 TF, TF-IDF 를 구해보자.

```
doc1 : The game of life is a game of everlasting learning
doc2 : The unexamined life is not worth living
doc3 : Never stop learning
```