7. 데이터 전처리-2

tm패키지를 이용한 데이터 전처리

I. tm패키지 주요 구성과 전처리 과정

- ◆ 텍스트 마이닝을 위한 패키지로 가장 빈번히 사용됨
- ◆ Corpus라는 문서의 집합구조를 기반으로 분석을 수행

how to) tm을 이용한 분석과정

➢ Corpus -문서의 집합을 의미하는 것으로 tm에서 텍스트를 분석하기 위해서는 기본적으로 corpus로 텍스트를 변경해주는 작업을 수행

```
> install.packages(c("tm","wordcloud"))
--- 현재 세션에서 사용할 CRAN 미러를 선택해 주세요 ---
URL 'http://cran.nexr.com/bin/windows/contrib/3.4/tm_0.7-3.zip'을 시도합니다
Content type 'application/zip' length 1275403 bytes (1.2 MB)
downloaded 1.2 MB

URL 'http://cran.nexr.com/bin/windows/contrib/3.4/wordcloud_2.5.zip'을 시도$
Content type 'application/zip' length 568561 bytes (555 KB)
downloaded 555 KB

패키지 'tm'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다
패키지 'wordcloud'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다

다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
        C:\Users\Administrator\AppData\Local\Temp\RtmpcBj71L\downloaded_pack$
> library(tm)
필요한 패키지를 로딩중입니다: NLP
> library(wordcloud)
필요한 패키지를 로딩중입니다: RColorBrewer
> a1<-readLines("d:\\machine r\\poem.txt")
> msg1<-Corpus(VectorSource(a1))
```

➢ 문서의 본문내용을 확인하기 위해 inspect()함수를 사용

```
> inspect(msg1)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 12

 [1] 癤\xbf it was many and many a year ago,In a kingdom by the sea,That a m$
 [2] And this maiden she lived with no other thought Than to love and be lov$
 [3] I was a child and she was a child, In this kingdom by the sea, But we l$
 [4] With a love that the winged seraphs of Heaven Coveted her and me.      $
 [5] And this was the reason that, long ago, In this kingdom by the sea, A w$
 [6] So that her highborn kinsmen came And bore her away from me, To shut he$
 [7] The angels, not half so happy in Heaven, Went envying her and me Yes! t$
 [8] That the wind came out of the cloud by night, Chilling and killing my A$
 [9] But our love it was stronger by far than the love Of those who were old$
[10] For the moon never beams, without bringing me dreams Of the beautiful A$
[11] And the stars never rise, but I feel the bright eyes Of the beautiful A$
[12] And so, all the night-tide, I lie down by the side Of my darling my dar$
> inspect(msg1[3:5])
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 3

[1] I was a child and she was a child, In this kingdom by the sea, But we lo$
[2] With a love that the winged seraphs of Heaven Coveted her and me.      $
[3] And this was the reason that, long ago, In this kingdom by the sea, A wi$
```

msg1[start:end] 형식처럼 범위를 지정하여 특정 문서내용만 확인할 수 있어요.

- 문서에서 문장부호를 제거하거나 대소문자 구분을 없애고 모두 소문자로 변경, 불필요한 공백을 제거하고자 할 때 tm_map()함수를 이용
- 형식: tm_map(문서집합,사용할 함수)
- 문서변경을 위한 함수
  - ✓ stripWhitespace – 불필요한 공백을 제거
  - ✓ stemDocument – 문장단어들 예를 들어, live, lived가 있을 때 이들을 하나의 단어로 인식하게끔 어근형태로 바꿔주는 함수
  - ✓ tolower – 영문 소문자로 변경
  - ✓ removeNumbers – 불필요한 숫자를 제거
  - ✓ removePunctuation-불필요한 문장부호를 제거
  - ✓ removeWords – 불필요한 단어를 제거

```
> msg2<-tm_map(msg1,stripWhitespace)
> msg2<-tm_map(msg2,tolower)
> msg2<-tm_map(msg2,removeNumbers)
> msg2<-tm_map(msg2,removePunctuation)
```

- 문서의 의미있는 표현 – 문서행렬 표시 TermDocumentMatrix()함수
- TermDocumentMatrix() – 주어진 Corpus에서 단어를 행, 문서를 열로 표시하는 행렬구조를 만들어줌.
- 형식 : TermDocumentMatrix(Corpus,옵션)

```
> msg3<-TermDocumentMatrix(msg2)
> inspect(msg3)
<<TermDocumentMatrix (terms: 101, documents: 12)>>
Non-/sparse entries: 184/1028
Sparsity           : 85%
Maximal term length: 15
Weighting          : term frequency (tf)
Sample             :
           Docs
Terms       1 11 12 2 3 5 6 7 8 9
  and       1  1  2 2 2 1 1 1 1 1
  annabelle 1  1  0 0 1 1 0 0 1 1
  her       0  0  2 0 0 0 3 1 0 0
  kingdom   1  0  0 0 1 1 1 1 0 0
  love      0  0  0 2 3 0 0 0 0 2
  sea       0  0  2 0 1 1 1 1 0 1
  that      0  0  0 0 1 1 1 1 1 0
  the       2  3  4 0 1 2 1 3 2 6
  this      0  0  0 1 1 2 1 1 0 0
  was       1  0  0 0 3 1 0 1 0 1
```

- 문서에서 빈번하게 출현하는 단어표시–findFreqTerms()함수
- findFreqTerms()–단어-문서의 행렬구조로부터 빈번하게 출현하는 단어들을 화면에 표시
- 형식 : findFreqTerms(행렬,lowfreq=최소출현횟수값,highfreq=최대출현횟수값)

```
> msg_result<-findFreqTerms(msg3,lowfreq=2)
> msg_result
 [1] "and"      "annabelle" "kingdom"   "know"      "live"      "maiden"
 [7] "mani"     "the"       "there"     "was"       "love"      "she"
[13] "than"     "this"      "with"      "but"       "child"     "sea"
```

- 단어와 단어간의 연관성표시–findAssocs()함수
- findAssocs() : 단어 -문서의 행렬구조로부터 주어진 단어와의 상관관계가 높은 단어들을 화면에 표시
- 형식 : findAssocs(행렬,terms키워드단어,corlimit상관계수하한값)

```
> msg_assoc<-findAssocs(msg3,'she',0.3)
> msg_assoc
$she
   love    with     and   other thought   child    more     was    live
   0.80    0.77    0.67    0.67    0.67    0.67    0.67    0.48    0.40
 maiden    this    than
   0.40    0.35    0.30
```