# 텍스트 마이닝 중간 정리_Summary_0319

# Summary

06-01 nchar() : 글자수를 반환
06-02 length() : 벡터의 길이를 반환
06-03 paste() : 몇몇의 문자열을 결합시킨다.
06-04 substr() : 문자열의 일부를 가져온다.
06-05 strsplit() : 기준이 되는 것에 따라 문자를 나누기

07-01 오늘의 날짜를 알자. Sys.Date()
07-02 문자열을 날짜로 변경. as.Date("2010-12-31")
07-03 날짜를 문자열로 변경. as.character(Sys.Date())
07-04 년월일을 날짜로 변경. ISOdate(year, month, day), as.Date(ISOdate(year, month, day))
07-05 날짜의 일부값을 추출하기
d <- as.Date("2010-03-15")
p <- as.POSIXlt(d)
p$mday, p$mon, p$year

useSejongDic()
extractNoun(sentence)
mergeUserDic( )

**형태소 분석하기**
MorphAnalyzer(Sen) : 형태소 분석 함수

**품사 달기**
SimplePos09 : 9개의 품사 태그
SimplePos22 : 22개의 품사 태그

**텍스트에서 말뭉치로 변환**
VectorSource(x) --> Corpus( ) --> tm_map( ) --> Ter,DocumentMatrix( ) --> colSum( ), sort( )

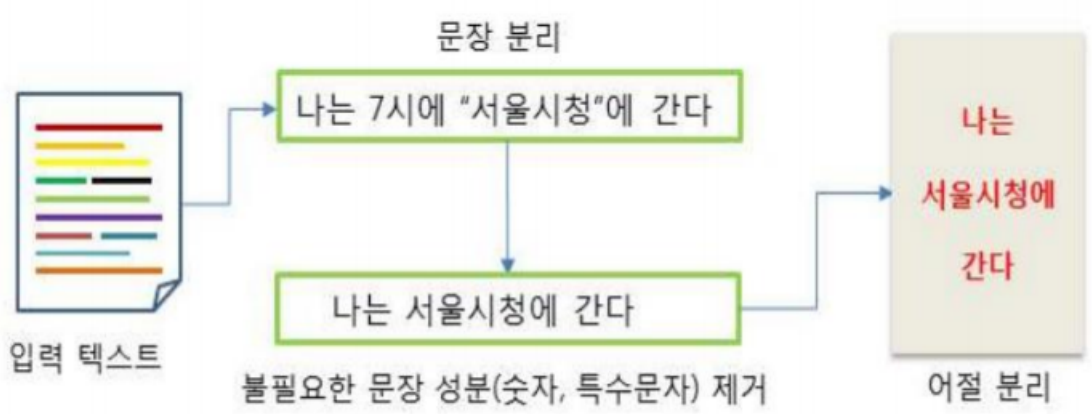
---

### 텍스트 마이닝이란?
비정형 텍스트 데이터 로부터 유용한 정보를 추출하는 기술

### 형태소 분석?
주어진 텍스트를 단어와 문법적 특성에 맞추어 명사, 동 사, 꾸밈어, 조사 등의 형태소로 분리할 수 있다.

### 형태소 분석이란?
주어진 단어 또는 어절을 구성하는 각 형태소를 분리한 후, 분리된 형태소의 기본형 및 품사 정보를 추출.

### 텍스트 전처리(text pre-processing)
텍스트 전처리 과정은 텍스트 분석을 위해 **문장 분리**, **불필요한 문장 성분을 제거**하는 과정이다




[그림 2-3] 전처리(pre-processing) 결과

### 품사 태깅 (POS tagging, Part-Of-Speech tagging)

하나의 단어가 여러 품사를 갖는다. 따라서 품사의 모호성 (혹은 중의성)을 제거하는 과정이 필요.
이를 수행하는 과정 을 품사 태깅이라 한다.

### 말뭉치란? (Corpus)

컴퓨터가 판독할 수 있는 형태(machine-readable form)

### 단어와 문서 관계 표현

말뭉치로부터 단어와 문서의 관계를 표현하기 위해 문서-단어 행렬 혹은 단어-문서 행렬을 작성

▶ 단어-문서 행렬예시

단어가 행을 구성,
문서가 열을 구성한다.
단어 포함(1), 불포함(0)

입력 텍스트 1 : "파스타 먹방, 강남 파스타 데이트"

입력 텍스트 2 : "강남 버스 파스타 맛집"

입력 텍스트 3 : "강남 버스, 강남 파스타, 강남 맛집

| | 입력텍스트 1 | 입력텍스트 2 | 입력텍스트 3 |
|---|---|---|---|
| 파스타 | 1 | 1 | 1 |
| 먹방 | 1 | 0 | 0 |
| 강남 | 1 | 1 | 1 |
| 데이트 | 1 | 0 | 0 |
| 버스 | 0 | 1 | 1 |
| 맛집 | 0 | 1 | 1 |
| ... | ... | ... | ... |

대상 텍스트

말뭉치에서 추출된 단어들

[그림 2-12] 단어-문서 행렬 예시

### TF-IDF
### 단어빈도(term frequency, TF)
### 역문서 빈도(inverse document frequency, IDF)

# TF-IDF – 버스

단어와 문서 관계 표현

| | 입력텍스트 1 | 입력텍스트 2 | 입력텍스트 3 |
|---|---|---|---|
| 파스타 | 0 | 0 | 0 |
| 데이트 | 0.24 | 0 | 0 |
| 버스 | 0 | 0.044 | 0.029 |
| 맛집 | 0 | 0.044 | 0.029 |
| ... | ... | ... | ... |

대상 텍스트

말뭉치에서 추출된 단어들

입력 텍스트 1 : "파스타 먹방, 강남 파스타 데이트"
입력 텍스트 2 : "강남 버스 파스타 맛집"
입력 텍스트 3 : "강남 버스, 강남 파스타, 강남 맛집"

| 입력 텍스트 2 | 입력 텍스트 3 |
|---|---|
| TF = 1/4, | TF = 1/6, |

**IDF = log(3/2) = 0.176**

**0.25 * 0.176= 0.044** **0.17 * 0.176= 0.029**

$$TF = \frac{\text{문서 내 단어 수}}{\text{문서 내 모든 단어 수}}$$

$$IDF = \log\left(\frac{\text{전체 문서 수}}{\text{단어를 포함한 문서 수}}\right)$$