08 텍스트 정보를 활용한 예측

1 데이터 소스, 특성, 현황 및 필요성

1 데이터 소스

- Web/SNS 게시판 내용
- Web Log
- App Log
- Machine Log
- 통화내용

2 데이터 특성

Web/SNS/email

- 반 정형화된 형태로 원천 장소에 의존적임
- 대용량에서 대용량 데이터 추출됨

Web Log

- 반 정형화된 형태로 복잡함
- 대용량에서 소량규모의 데이터 추출됨

2 데이터 특성

App Log

- 반 정형화된 형태로 복잡하지만 web에 비해 다소 단순함
- 대용량에서 소량규모의 데이터 추출됨

Machine Log

- 반 정형화된 형태로 보다 단순함
- 지속적으로 소량 데이터 발생, 누적 시 규모 큼
- 다수의 장비에서 발생

2 데이터 특성

통화내용

- 비정형으로 정의하기에 따라 활용도가 매우 다름
- 중간규모의 복잡한 내용의 데이터 많이 발생함

3 데이터 현황

- Internet의 발달로 다양한 텍스트 정보가 전세계적으로 다양한 용도로 발생
- 대부분의 사람들이 Internet에 데이터 발생시키고 있음
- 인터넷 도임 후 20년 넘는 기간 동안 많은 데이터가 누적되고 있음
- IoT의 발전으로 더욱 많은 데이터가 발생할 것임

4 데이터 필요성

- **정형 데이터 활용의 한계성**으로 추가 데이터 필요
- **외부 데이터 활용**을 통한 정보의 한계극복 필요
- **종합적이고 시간의 변화에 따른 패턴**을 파악할 수 있는 데이터 필요

08 텍스트 정보를 활용한 예측

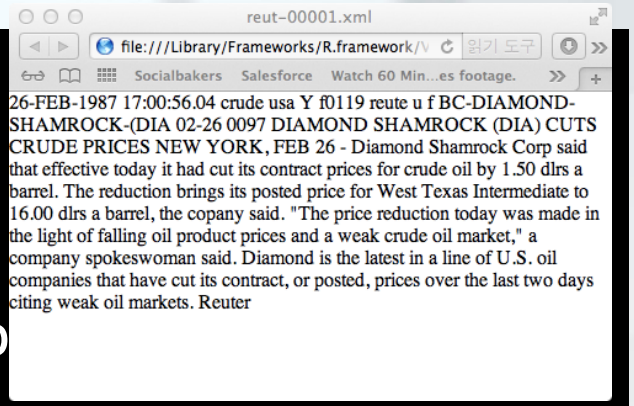
2 텍스트 처리 절차

1 처리절차

1. Data Scrap
2. Plain Text
3. Remove Whitespace
4-1. Remove Punctuation
4-2. Remove Number
5. Remove Stopword
6. Stemming
7. Document Term Matrix
8. Remove Sparse Term
9. Text Analysis

2 R - Text Mining

```
> library(tm)
> data(crude)
> inspect(crude[1])
Diamond Shamrock Corp said that
effective today it had cut its contract p
1.50 dlrs a barrel.
    The reduction brings its posted price for West Texas
Intermediate to 16.00 dlrs a barrel, the copany said.
    "The price reduction today was made in the light of falling
oil product prices and a weak crude oil market," a company
spokeswoman said.
    Diamond is the latest in a line of U.S. oil companies that
have cut its contract, or posted, prices over the last two days
citing weak oil markets.
 Reuter
```



2 R - Text Mining

```
> crude <- tm_map(crude, removePunctuation)
> inspect(crude[1])
Diamond Shamrock Corp said that
effective today it had cut its contract prices for crude oil by
150 dlrs a barrel
    The reduction brings its posted price for West Texas
Intermediate to 1600 dlrs a barrel the copany said
    The price reduction today was made in the light of falling
oil product prices and a weak crude oil market a company
spokeswoman said
    Diamond is the latest in a line of US oil companies that
have cut its contract or posted prices over the last two days
citing weak oil markets
 Reuter
```

2 R - Text Mining

```
> crude <- tm_map(crude, function(x) removeWords(x,stopwords()))
> inspect(crude[1])
Diamond Shamrock Corp
effective    cut  contract prices  crude oil
150 dlrs  barrel
    The reduction brings  posted price  West Texas
Intermediate  1600 dlrs  barrel  copany
    The price reduction      light  falling
oil product prices   weak crude oil market  company
spokeswoman
    Diamond      line  US oil companies
 cut  contract  posted prices     days
citing weak oil markets
 Reuter
```

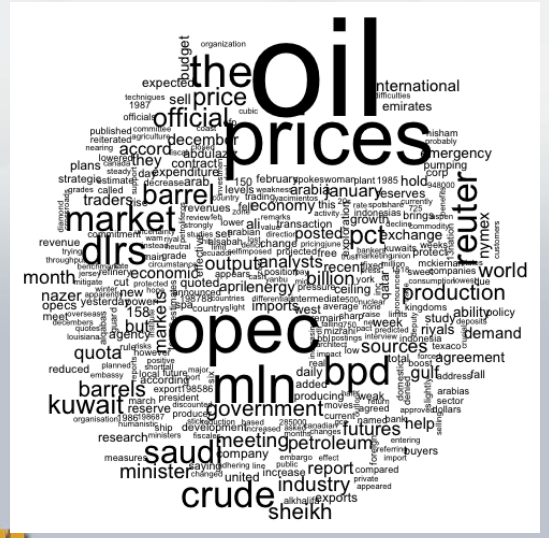
2 R - Text Mining

```
> tdm <- TermDocumentMatrix(crude)
> inspect(tdm[100:105,10:15])
A term-document matrix (6 terms, 6 documents)

Non-/sparse entries: 4/32
Sparsity           : 89%
Maximal term length: 11
Weighting          : term frequency (tf)

              Docs
Terms          248 273 349 352 353 368
  architect      1   0   0   1   0   0
  argentine      0   0   0   0   0   0
  arrangement    0   0   0   0   0   0
  asia           0   0   0   0   0   0
  asian          0   0   0   0   0   0
  asked          1   0   0   1   0   0
```

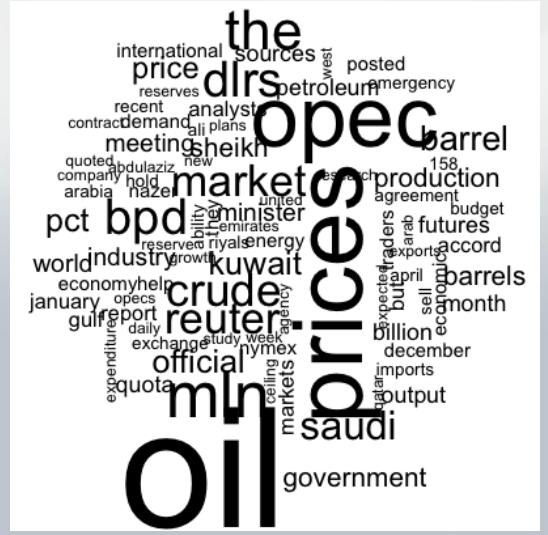
2 R - Text Mining

```
> m <- as.matrix(tdm)
> v <- sort(rowSums(m),decreasing=TRUE)
> d <- data.frame(word = names(v),freq=v)
> wordcloud(d$word,d$freq,c(8,.3),2)
> wordcloud(d$word,d$freq,c(8,.5),2,,FALSE,.1)
```

minimum frequency=2



minimum frequency=5



3 American Customer Satisfaction Index

ACSI

The American Customer Satisfaction Index™

X close

Print

Scores By Industry

Hotels

	Base-line	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	Previous Year % Change	First Year % Change
Hilton	75	75	75	75	72	74	77	74	76	74	77	76	78	76	78	79	80	80	0.0	6.7
Starwood	NM	NM	NM	NM	NM	NM	73	71	69	73	73	75	75	76	74	74	77	79	2.6	8.2
Marriott	80	76	77	76	76	77	74	77	76	76	76	76	75	79	78	77	80	79	-1.3	-1.3
Hotels	**75**	**73**	**72**	**71**	**71**	**72**	**72**	**71**	**71**	**73**	**72**	**73**	**75**	**71**	**75**	**75**	**75**	**77**	**2.7**	**2.7**
Hyatt	76	75	77	77	75	73	74	73	75	77	74	74	75	77	78	74	79	77	-2.5	1.3
All Others	NM	73	71	71	70	71	72	70	70	72	71	73	76	70	76	76	74	77	4.1	5.5
InterContinental	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	74	75	78	76	-2.6	2.7
Best Western	74	70	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	70	75	76	76	0.0	2.7
Choice	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	71	76	74	74	0.0	4.2
Wyndham	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	70	70	70	73	4.3	4.3
Holiday Inn	69	69	NM	NM	69	68	71	71	69	72	73	69	72	72	#				N/A	N/A
Promus Hotel	82	80	83	77	78	79	#												N/A	N/A
Ramada	70	69	70	64	67	67	69	66	67	70	67	66	70	69	#				N/A	N/A

Score tables print best in landscape.

Legend

NA	Not available
#	Company merger
†	Company defunct
NM	Not measured
^	Industry aggregated

4 Positive / Negative Word

[Positive Words]

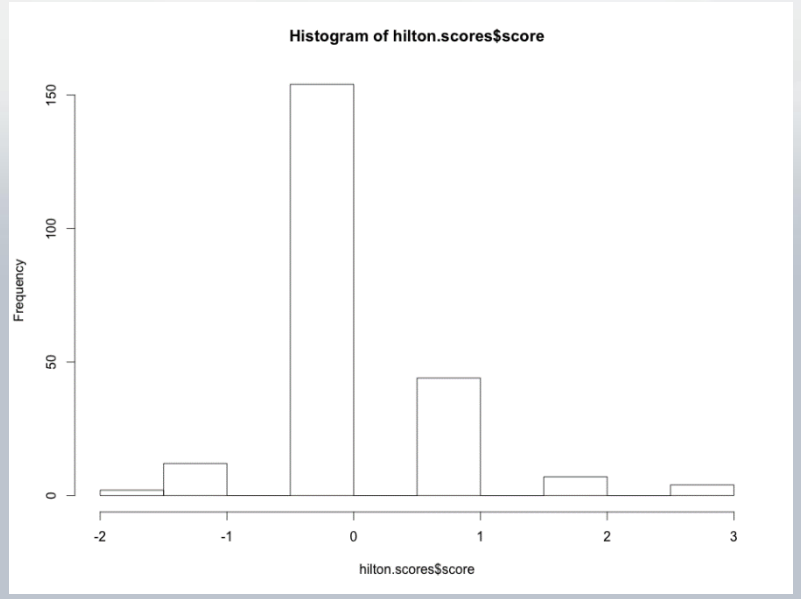
a+
abound
abounds
abundance
abundant
accessable
accessible
acclaim
acclaimed
acclamation
accolade
accolades
accommodative
accomodative
accomplish
accomplished

[Negative Words]

2-faced
2-faces
abnormal
abolish
abominable
abominably
abominate
abomination
abort
aborted
aborts
abrade
abrasive
abrupt
abruptly
abscond

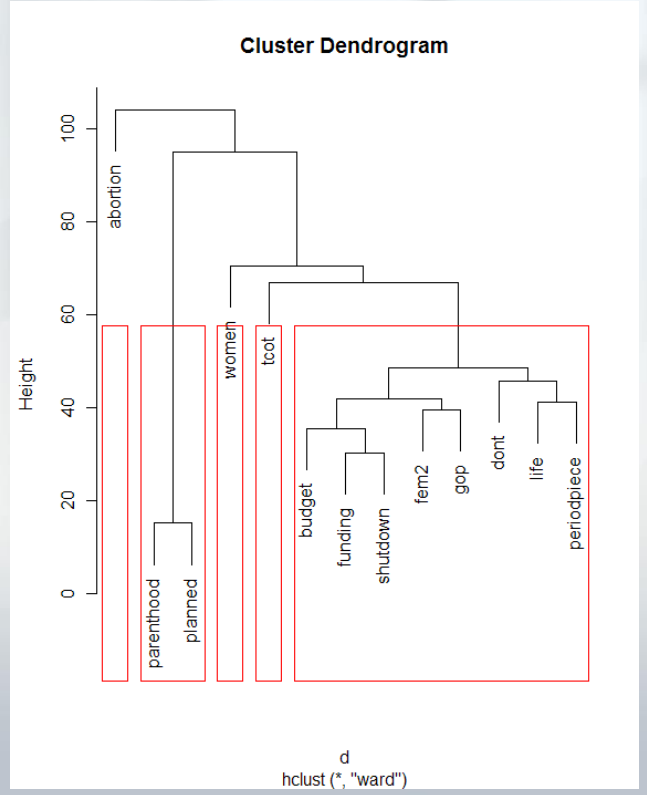
5 R - Sentiment Analysis

```
> library(twitteR)
> hilton.tweets <- searchTwitter("@hilton",n=1500,cainfo='cacert.pem')
> hilton.text <- laply(hilton.tweets, function(t)t$getText())
> pos.word=scan("positive-words.txt",what="character",comment.char=";")
> neg.word=scan("negative-words.txt",what="character",comment.char=";")
> hilton.scores=score.sentiment(hilton.text,pos.words,neg.words,.progress='text')
> hist(hilton.scores$score)
```



6 R - Text Mining

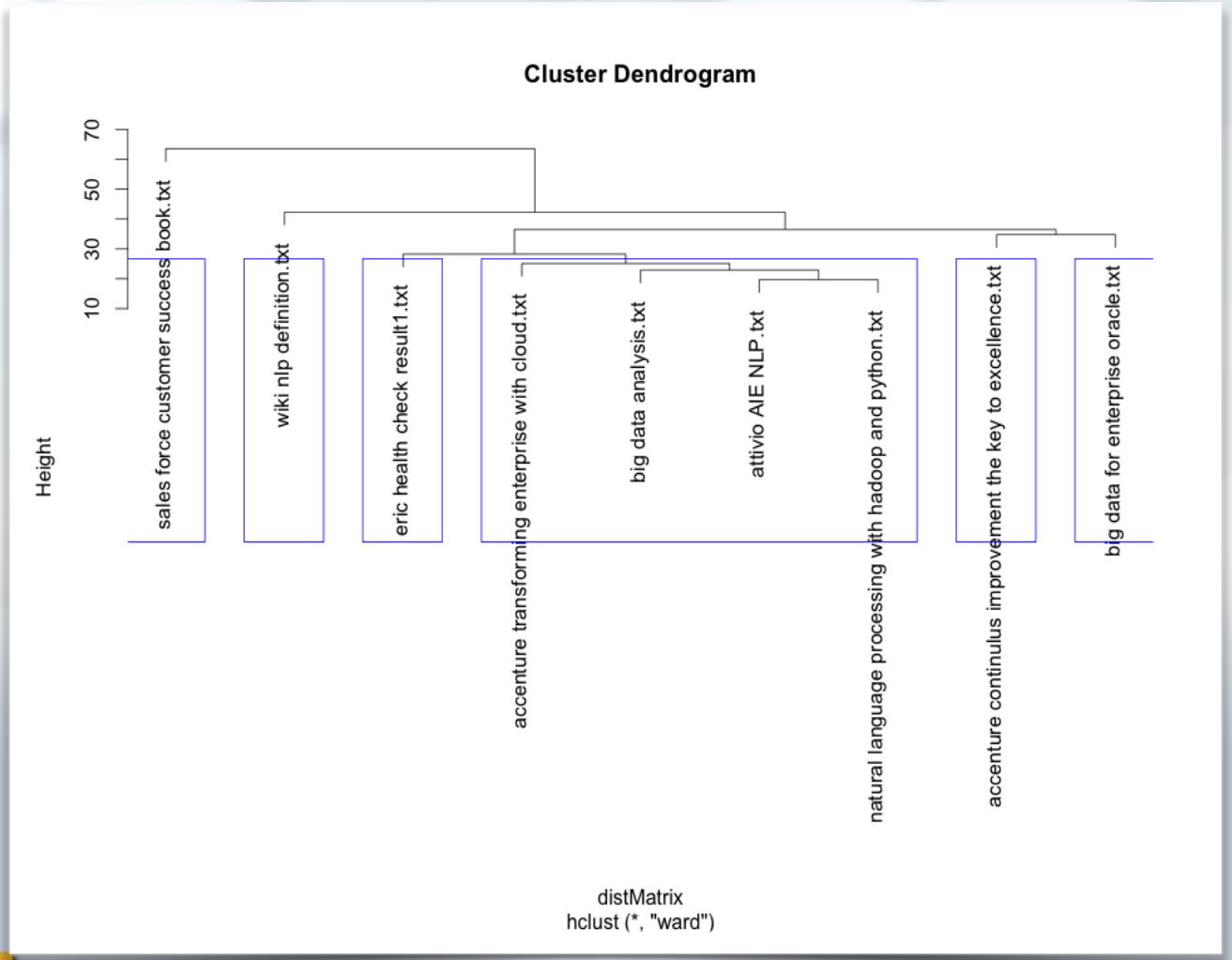
```
> findFreqTerms(mydata.dtm, lowfreq=30)
> findAssocs(mydata.dtm, 'fetus', 0.20)
> mydata.dtm2 <- removeSparseTerms(mydata.dtm,
sparse=0.95)
> mydata.df <- as.data.frame(inspect(mydata.dtm2))
> nrow(mydata.df)
> ncol(mydata.df)
> mydata.df.scale <- scale(mydata.df)
> d <- dist(mydata.df.scale, method = "euclidean") # distance
matrix
> fit <- hclust(d, method="ward")
> plot(fit)
> groups <- cutree(fit, k=5) # cut tree into 5 clusters
> rect.hclust(fit, k=5, border="red")
```



7 Text Mining & Clustering for KMS

- Conventional KMS is not flexible
- Document categorization need to improved
- Hard to find proper information
- Redundant analysis should be decreased

8 R - Text Mining & Clustering



8 R - Text Mining & Clustering

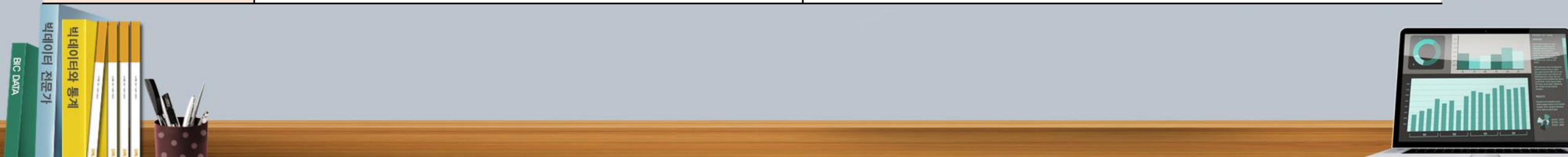
- cluster 1: data database enterprise
- cluster 2: data analytics language
- cluster 3: accenture business value
- cluster 4: language nlp evaluation
- cluster 5: level date values
- cluster 6: customer solution business

08 텍스트 정보를 활용한 예측

3 활용방안

1 활용방안

사례	필요성	혜택
감성 분석	기업에 대한 고객들의 반응을 긍정적인지, 일반적인지, 부정적인지 파악하여 대응	실시간 데이터 수집 및 추이 분석 통해 급작스런 문제에 조기 대응하거나 중장기적 방안 수립하는데 활용
주제 분석	특정분야 관련 내용이 존재하는지 자동으로 자료 선별해서 응용	논문, 특허, 범죄기록 등 단순검색으로 처리할 수 없는 내용에 대해 활용



1 활용방안

사례	필요성	혜택
영업 기회 획득	실시간으로 소셜 미디어에 언급되는 자동차 구매 니즈 고객 확보 하여 매출 증대	소셜미디어에 있는 자동차 구매니즈가 있는 언급들을 수집 및 분석하여 영업에 활용
분류 식별	많은 정보에서 거짓정보 식별하여 의미있는 정보 획득	인터넷 상에 있는 정보, 소송관련 내용, 이력서 및 자기소개서의 진실/거짓을 식별하는데 활용

1 활용방안

사례	필요성	혜택
성향 파악	콜센터에서 발생하는 많은 데이터 중 대화내용을 활용하여 추가적인 서비스나 서비스 개선에 활용	콜센터 대화내용을 텍스트 전환 통해 고객성향 구분, 구매의지 강도 확인→ 매출증대, 고객만족에 활용