

# **TEXT MINING for PRACTICE**

Python을 활용한 비정형 데이터 분석 - WEEK 09

---

전병진 FINGEREDMAN (fingeredman@gmail.com)

# Part 8.

텍스트 감성분석과 활용  
문서요약과 키워드 추출

# 감성분석 (Sentiment Analysis)

## 단어의 감성수준을 수치화하는 분석방법

- ▶ 문장이 의미하는 감성의 극성을 판별하거나 그 수준을 점수로 매기는 방법
- ▶ 텍스트 데이터를 계량 데이터로 바꾸는 가장 좋은 방법 중 하나
- ▶ 사전(말뭉치) 기반 감성분석과 머신러닝을 활용한 감성분석이 있음

[ 사전기반 감성분석 ]

A = I am not interested in class and have no fun.  
B = Today class is very interesting and fun.



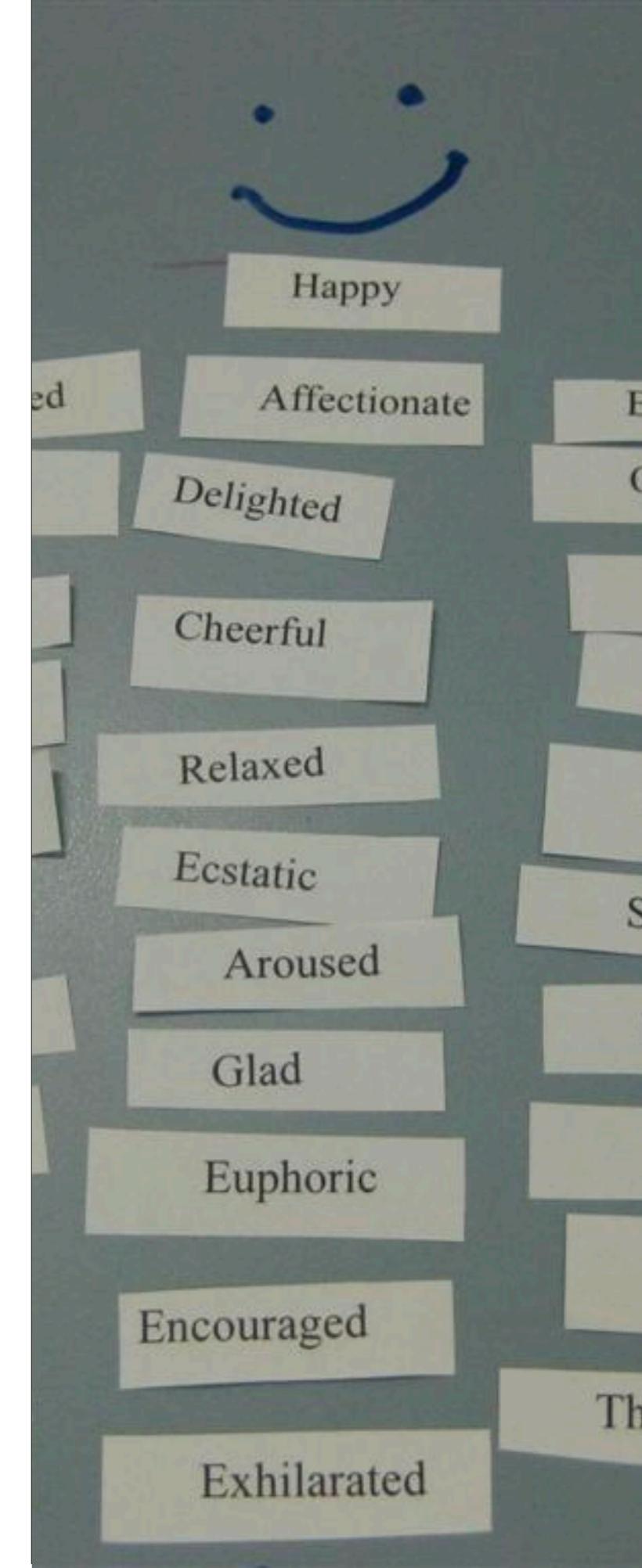
$A_{Senti} = [(i, 0), (be, 0), (not, x-1), (interest, 0.8), (in, 0), (class, 0), (and, 0), (have, 0), (no, x-1), (fun, 0.9)]$   
 $B_{Senti} = [(Today, 0), (class, 0), (be, 0), (very, x2), (interesting, 0.8), (and, 0), (fun, 0.9)]$



$$A_{Score} = -1 \times 0.8 + -1 \times 0.9 = -1.7$$
$$B_{Score} = 2 \times 0.8 + 0.9 = 2.5$$

[ 감성사전 예시 ]

Word	Polarity	Weight
not	-	negation
no	-	negation
...	...	...
interest	+	0.8
fun	+	0.9
...	...	...
sorry	-	0.9
sad	-	0.8
...	...	...

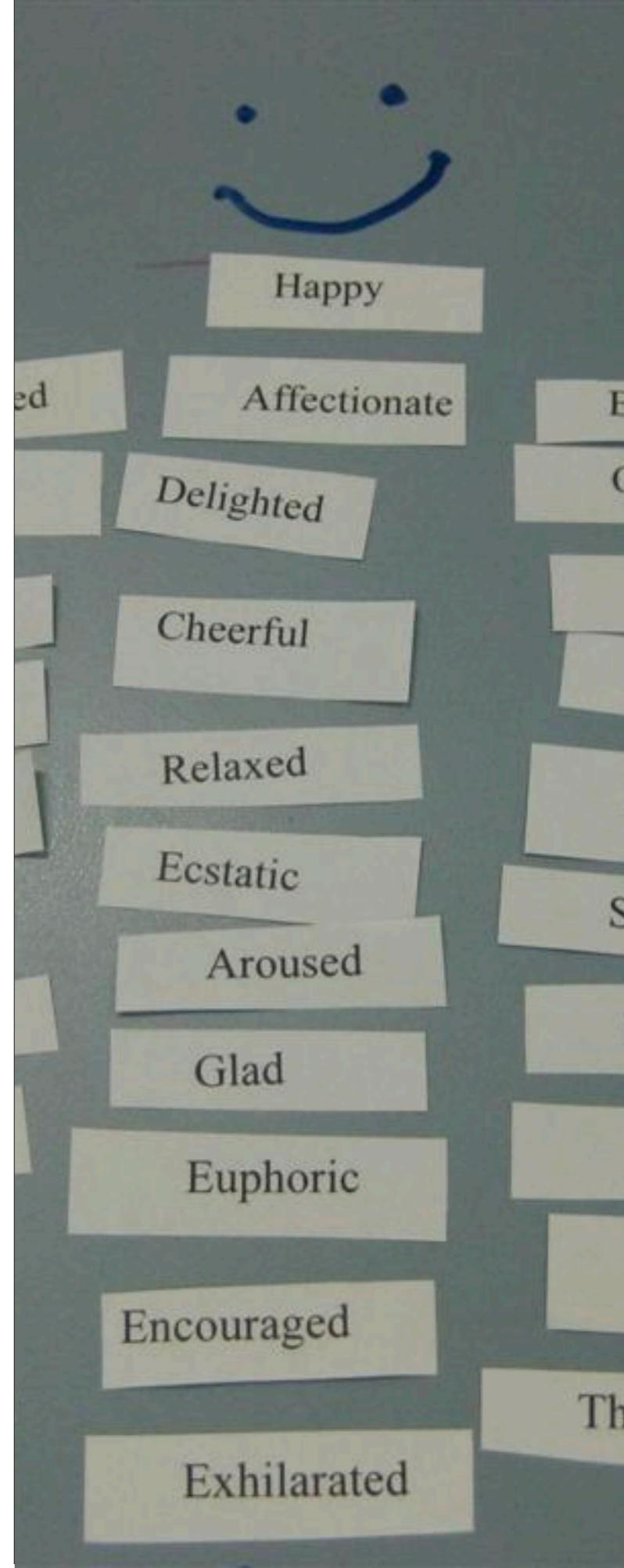


# 감성분석 (Sentiment Analysis)

## 상용/연구용 감성사전 종류

- ▶ Linguistic Inquiry and Word Count (LIWC) - <http://www.liwc.net/>
- ▶ MPQA Subjectivity Cues Lexicon - <http://www.cs.pitt.edu/>
- ▶ SentiWordNet - <http://sentiwordnet.isti.cnr.it/>
- ▶ KOSAC - <http://word.snu.ac.kr/kosac/icon.php>

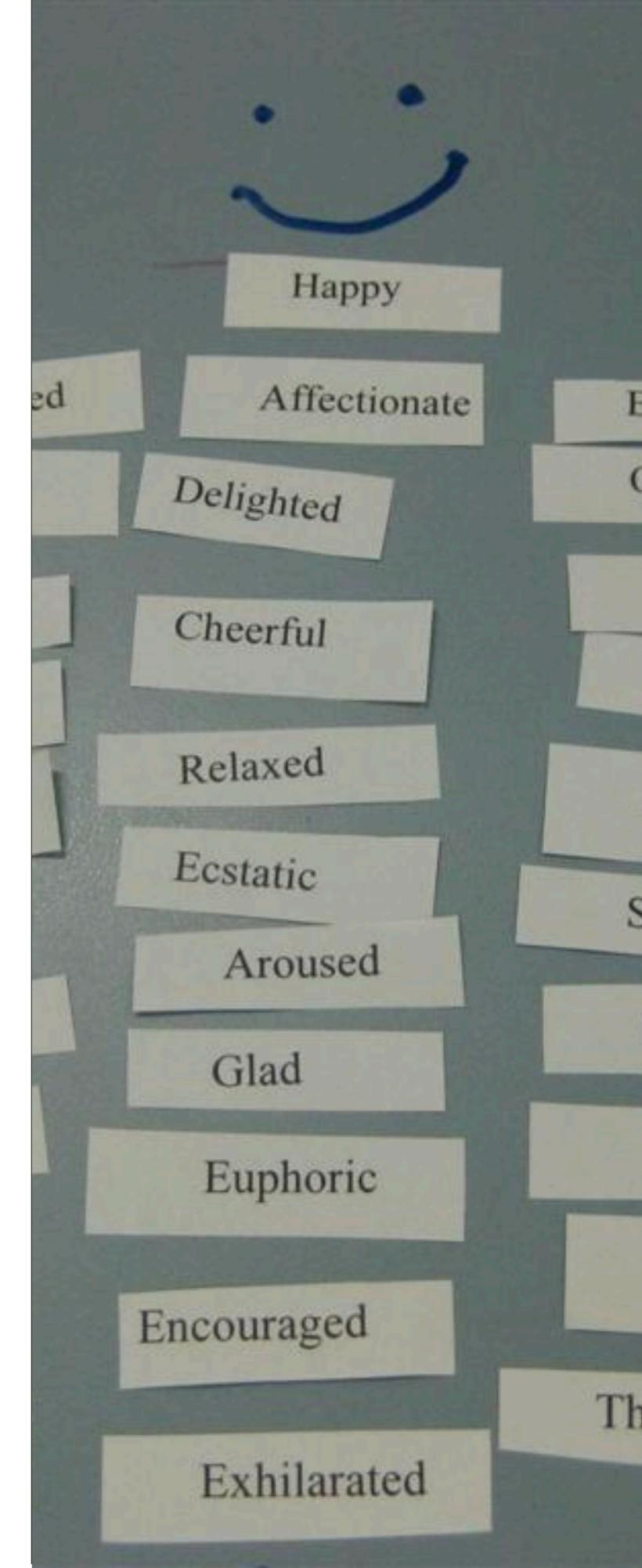
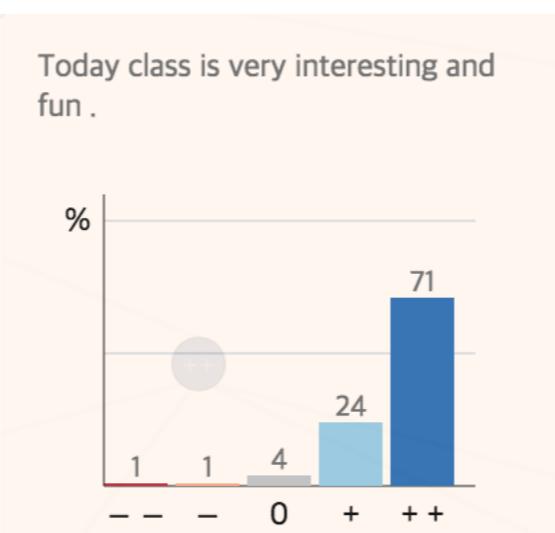
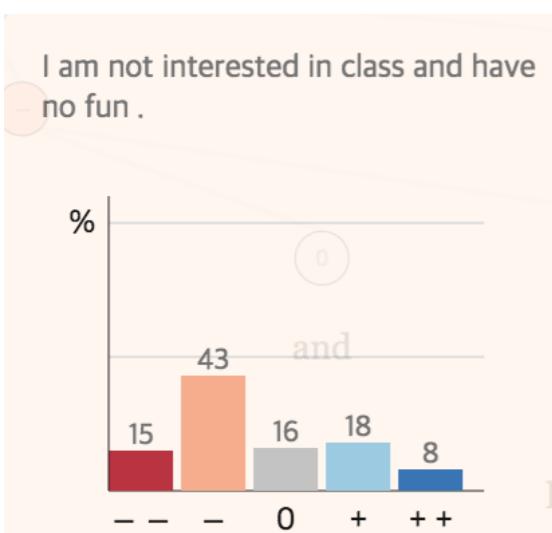
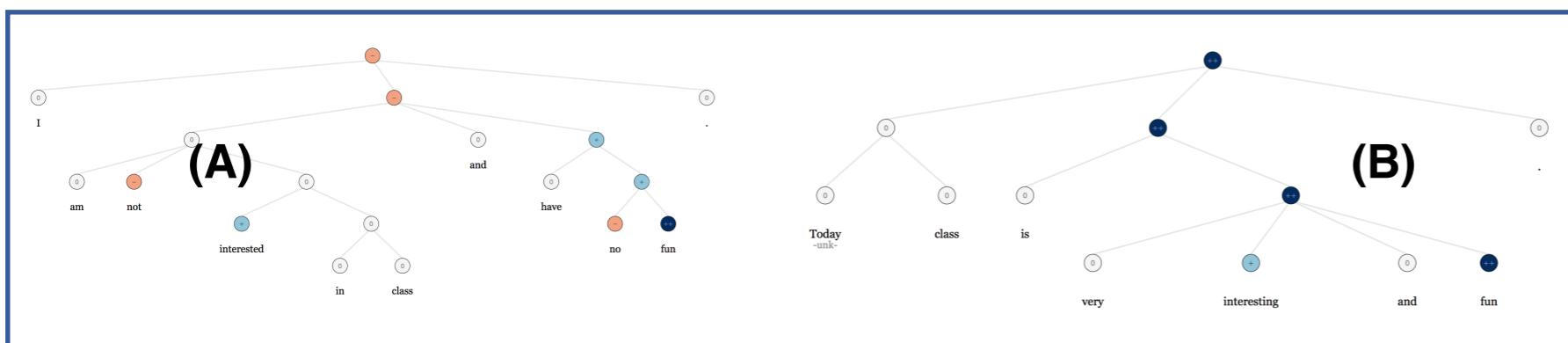
ngram	freq	COMP	NEG	NEUT	None	POS	max.value	max.prop
싸구려/NNG	1	0	1	0	0	0	NEG	1
싸구려/NNG;로/JKB	1	0	1	0	0	0	NEG	1
싸구려/NNG;로/JKB;둔갑/NNG	1	0	1	0	0	0	NEG	1
싸늘/XR	1	0	1	0	0	0	NEG	1
싸늘/XR;하/XSA	1	0	1	0	0	0	NEG	1
싸움/NNG	1	0	1	0	0	0	NEG	1
싸움/NNG;을/JKO	1	0	1	0	0	0	NEG	1
싸움/NNG;을/JKO;일으키/VV	1	0	1	0	0	0	NEG	1
써먹/VV	1	0	1	0	0	0	NEG	1
써먹/VV;지/EC	1	0	1	0	0	0	NEG	1
써먹/VV;지/EC;못하/VX	1	0	1	0	0	0	NEG	1
기대/NNG;되/XSV	2	0	0	0	0	1	POS	1
기대/NNG;를/JKO	2	0	0	0	0	1	POS	1
기대/NNG;하/XSV	2	0	0	0	0	1	POS	1
기량/NNG	2	0	0	0	0	1	POS	1
기뻐하/VV	2	0	0	0	0	1	POS	1
기회/NNG;를/JKO;주/VV	2	0	0	0	0	1	POS	1
길/VA	2	0	0	0	0	1	POS	1
꼭/MAG;필요/NNG	2	0	0	0	0	1	POS	1
꼭/MAG;필요/NNG;하/XSA	2	0	0	0	0	1	POS	1
꼽/VV	2	0	0	0	0	1	POS	1
꼽히/VV	2	0	0	0	0	1	POS	1
꽃/NNG;을/JKO	2	0	0	0	0	1	POS	1
꽃/NNG;을/JKO;피우/VV	2	0	0	0	0	1	POS	1



# 감성분석 (Sentiment Analysis)

[ 머신러닝(딥러닝) 기반 감성분석 ]

A = I am not interested in class and have no fun.  
B = Today class is very interesting and fun.



# 감성분석 (Sentiment Analysis)

## Article Review: DBR

SR1. 감성 분석 활용 사례

“구매후기 한 줄에 고객의 이런 속마음이”

마케팅 난제, 속 시원히 풀어주는 분석

261호 (2018년 11월 Issue 2)



### Article at a Glance

최근 ‘글에 내재해 있는 사람들의 주관적 태도나 감성을 추출해 내는 분석 기법’인 ‘감성 분석’에 대한 관심이 높아지고 있다. 감성 분석은 소셜미디어와 같은 웹사이트/매체에서 정보를 수집하는 ‘데이터 수집’ 단계, 수집된 정보에서 텍스트 작성자의 주관이 드러난 부분만을 걸러내는 ‘주관성 탐지’ 과정, 마지막으로 ‘주관성의 극성’이나 ‘정도’를 측정하고 분류하는 과정으로 나눌 수 있다. 대표적 성공 사례로 초코바 스니커즈의 소비자 감성 변화에 따른 가격변동 마케팅, 국내 유명 화장품 브랜드 에뛰드하우스의 감성 분석 등을 꼽을 수 있다. 감성 분석에 성공하기 위해서는 적용 분야별 특성을 살린 사전을 구축하고, 데이터 수집 전략을 세우며, 다른 데이터와 연계해 다양한 분석을 수행할 수 있어야 한다.

# 감성분석 (Sentiment Analysis)

## Article Review: DBR

SR2. 감성 분석 잠재력과 한계

### 트럼프 당선 예측했던 그 분석 인간의 언어에서 감정을 읽어내

261호 (2018년 11월 Issue 2)



#### Article at a Glance

감성 분석은 초기 ‘워드클라우드’라는 이미지가 너무 굳어져 화려하기만 하고 별다른 성과는 내지 못하는 것으로 오해를 받았다. 그러나 ‘진짜’ 감성 분석은 소비자, 유권자 마음속에 숨은 감성을 파악할 수 있는 분석 기술이다. 트럼프 당선 예측이나 구글의 상품 검색 활용에서 그 힘을 보여준 감성 분석은 현재 많은 기업이 사활을 걸고 연구하는 분야이기도 하다. 감성 분석의 가능성과 한계를 정확히 이해하고 ‘어떤 감성이 돈이 되는지’ ‘돈이 되는 그 감성을 다른 범주의 감성과 칼 같이 나눌 수 있는지’ ‘그 감성은 어디에서 시작해 어디를 향하고 있는지’를 점검해보고 방향을 잡는다면 기업의 미래를 바꿀 수 있을지도 모른다.

#### 트럼프 당선을 정확히 예측한 감성 분석 기술

지난 2016년 11월9일, 전 세계 이목은 미국에 쏠렸다. 누가 미국 대통령직을 거머쥐게 될 것인가. 당시 대다수는 민주당 힐러리 클린턴이 공화당 도널드 트럼프를 제치고 당선될 것이라고 내다봤다. 실제 미국 주요 언론사/여론조사 기관 19곳 가운데 17개가 힐러리의 당선을 예측했다.

# 감성분석 (Sentiment Analysis)

## Article Review: DBR

SR3. 송민 연세대 교수 인터뷰

### ‘감성 분석 뜯다는데 해볼까’는 위험 어떤 문제에 왜 필요한지 정의 먼저

261호 (2018년 11월 Issue 2)



#### Article at a Glance

기업들이 최근 가장 많은 관심을 갖고 있는 데이터 분석인 ‘감성 분석’은 생각보다 무척 어렵고 복잡하다. 특히 표현이 다양하고 SNS나 인터넷상에서 온갖 형태로, 때로는 거의 아무런 맥락 없이 변형돼 쓰이는 한국어를 대상으로 텍스트 마이닝을 하고 감성 분석을 하는 건 더욱 어려운 일이다. 그럼에도 제대로 데이터를 수집하고 분석한다면 ‘소비자의 진짜 마음’을 읽을 수 있기에 기업 입장에서 결코 포기할 수 없는 기술이자 방법론이다. 이 분야 대가인 송민 교수는 기업인들에게 “데이터를 무작정 쌓지 말고 최소 한의 분류를 해 놓거나 태그는 달아 놓을 것”과 “남이 하니까 우리도 하자라는 생각으로 시작하지 말고, 자신의 회사가 어떤 이유로 감성 분석을 해야 하는지 명확하게 정립하고 문제 정의부터 할 것”을 주문했다.

#### 편집자주

이 기사의 제작에는 동아일보 미래전략연구소 인턴연구원 홍석영(연세대 불어불문학·경영학과 4학년) 씨가 참여했습니다.

# 감성분석 (Sentiment Analysis)

## Article Review: DBR

SR4. 소셜 분석 방법론

### 극도로 복잡한 한국어 소셜 버즈 분석 필요 없는 데이터 잘 버리는 게 핵심

261호 (2018년 11월 Issue 2)



#### Article at a Glance

소셜 버즈 분석은 소셜 모니터링, 소셜 트렌드, 타깃 마이닝으로 크게 나눌 수 있는데 그 유용성과 정확한 분석력으로 최근 급격히 각광받고 있으나 복잡한 한국어의 구조와 어려운 텍스트 분석 난이도로 인해 여전히 해결해야 할 과제가 많은 분석 기법이다. 소셜 분석 정확도를 높이기 위해서는 정확한 수집 대상을 선정해야 하며 필요 없는 데이터를 효과적으로 버리는 작업도 필수적이다. 특히 한국어 소셜 버즈가 가진 특징과 한계를 정확히 파악하고 극복하려는 노력도 절실하다.

#### 들어가며: 소셜 분석, ‘한국어 소셜 버즈’ 분석

많은 기업이 소셜 버즈 분석 1을 비즈니스에 적용하고 있다. 분석 결과를 활용해 새로운 상품과 서비스를 기획, 개발하고, 신제품에 대한 소비자 반응을 모니터링하며, 타깃 소비자들의 숨겨진 니즈를 발굴하고 있다.

# 감성분석 (Sentiment Analysis)

## Article Review: DBR

SR5. 감성 분석 활용 비즈니스 전략

‘감’으로 여겨졌던 영역을 수치화

확장 가능한 상상력, 새 비즈니스 기회

261호 (2018년 11월 Issue 2)



### Article at a Glance

감성컴퓨팅은 ‘사람이 보는 것을 기계가 볼 수 있도록 하라’는 것부터 시작됐다. 사물의 위치를 파악하는 인지 및 지각 분야만큼이나 인간의 감성 인식 분야는 상당한 양의 데이터와 기계 학습을 필요로 한다. 하지만 최근에는 적은 양의 데이터만 가지고도 기계가 자체적으로 학습할 수 있도록 하는 연구가 진행되고 있다. 감성 인식이 일상생활 모든 곳에서 공기처럼 존재할 날이 그만큼 가까워지고 있다는 뜻이다. 기계와 사용자의 심리적 친밀도 상승은 새로운 비즈니스 기회가 될 수 있다. 보유하고 있는 데이터들에 대한 중요도를 높이고, 보안과 관련해 더 막중한 책임감을 지녀야 하며, 확장 가능한 상상력을 기반으로 ‘기계 감성시대’를 준비할 필요가 있다.

# 키워드 추출 (Keyword Extraction)

## 키워드 추출이란?

- ▶ 문서의 주제를 가장 잘 설명하는 단어를 자동으로 식별하는 작업
- ▶ 키워드란 영문으로 Keywords, Key-phrases, Key-terms, Key-segments 등으로 표현할 수 있으며, 모두 문서를 가잘 잘 설명하는 용어를 의미하는 단어로 사용됨
- ▶ 키워드 추출은 텍스트 마이닝, 정보검색 (information retrieval) 및 자연어처리 (NLP) 분야에서 오랫동안 중요한 문제로 인식되어 왔으며, 다양한 알고리즘이 제안됨
  - TF-IDF (Term Frequency & Inverse Document Frequency)
  - TextRank (Google)
  - Practical Key-phrase Extraction Algorithm (PKEA)
  - Rapidly Key-phrase Extraction Algorithm (RKEA)



# 키워드 추출 (Keyword Extraction)

## 키워드 추출이란?

-----text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate \*\*\*\* text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate

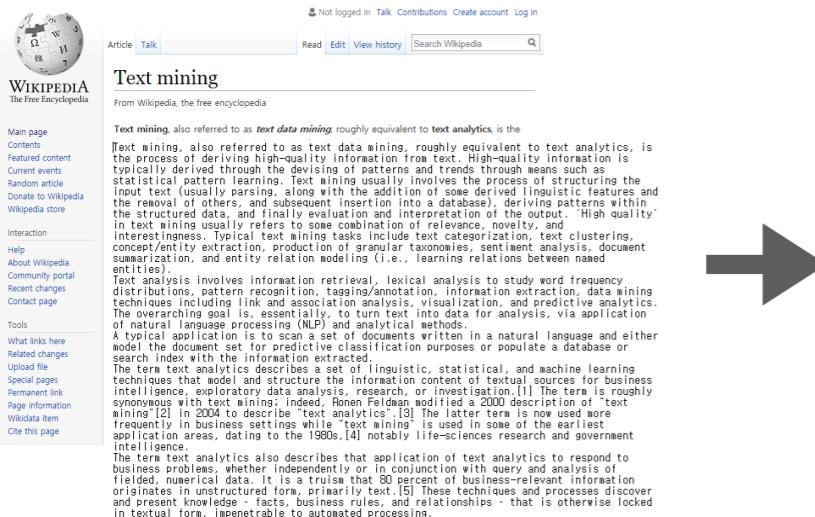
-- Extracting Keyphrases...  
-- Reading instance  
-- Converting instance

Word	Score
mining	0.5171
text mining	0.5171
data	0.5171
text data	0.5171
data mining	0.5171
text	0.5171
information	0.262
analytics	0.262
process	0.262
deriving	0.262
high-quality	0.262
high-quality information	0.262
derived	0.1455
involves	0.1455
analysis	0.0561
document	0.0561
database	0.0561
Typical	0.0561
text mining tasks include	0.0561
extraction	0.0561
predictive	0.0561
application	0.0561
natural	0.0561
natural language	0.0561
language	0.0561
set	0.0561
set of documents	0.0561
patterns	0.0238
concept/entity	0.0008

# Practical Key-phrase Extraction Algorithm (PKEA)

## Algorithm: Step 1. 키워드 후보군 선정

- ▶ (1) 텍스트 전처리 : 텍스트를 문장과 단어 단위로 구분함



	Index	Sentence	Words
1		Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the text (structured or unstructured), and the derivation of the term 'high quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).	Text, mining, *, referred, *, text, data, mining, *, equivalent, *, text, analytics, *, process, *, deriving, ...
2		Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of ...	Text, mining, *, involves, *, process, *, structuring, *, input, text, *, parsing, along, *, addition, *, ...
3		High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.	*, High, quality, *, text, mining, *, refers, *, combination, *, relevance, *, novelty, interestingness, ...

# Practical Key-phrase Extraction Algorithm (PKEA)

## Algorithm: Step 1. 키워드 후보군 선정

- (2) 키워드 식별 : 불용어를 제외한 문장내 모든 단어들의 연속된 조합 (N-gram)을 식별함 (불용어를 기준으로 키워드 구분)



Index	Sentence	Words	Index	Candidate Phrase
1	Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving ...	Text, mining, *, referred, *, text, data, mining, *, equivalent, *, text, analytics, *, process, *, deriving, ...	1	Text
2	Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of ...	Text, mining, *, involves, *, process, *, structuring, *, input, text, *, parsing, along, *, addition, *, ...	2	mining
3	High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.	*, High, quality, *, text, mining, *, refers, *, combination, *, relevance, *, novelty, interestingness, ...	3	Text mining

# Practical Key-phrase Extraction Algorithm (PKEA)

## Algorithm: Step 1. 키워드 후보군 선정

- ▶ (3) 원형복원 및 대/소문자 통일 : Stemming 또는 Lemmatization을 활용하여 단어의 원형을 복원하고 대/소문자를 통일함

Index	Candidate Phrase	Index	Candidate Phrase
1	Text	1	text
2	mining	2	mining
3	Text mining	3	text mining
4	referred	4	refer
5	text	5	data
6	data	6	text data
7	mining	7	data mining
8	Text data	8	text data mining
9	data mining	9	equivalent
10	text data mining	10	Analytics
11	...	11	...

# Practical Key-phrase Extraction Algorithm (PKEA)

## Algorithm: Step 2. 특성 계산 (Feature Calculation)

- ▶ TF-IDF : 단어가 출현하는 빈도와 그 희박성 (sparsness)에 의해 측정되는 척도
- ▶ First Occurrence : 단어가 문장의 첫 단어로부터 등장하는 거리 (가까울 수록 중요하다고 가정함)

Index	Candidate Phrase	Index	Candidate Phrase	TF-IDF	First Occurrence
1	Text	1	text	1.0833	0
2	mining	2	mining	1.1000	1
3	Text mining	3	text mining	1.4555	1
4	referred	4	refer	0.7500	3
5	text	5	data	0.7500	7
6	data	6	text data	1.4555	7
7	mining	7	data mining	1.4666	8
8	Text data	8	text data mining	...	...
9	data mining	9	equivalent	...	...
10	text data mining	10	Analytics	...	...
11	...	11	...	...	...

# Practical Key-phrase Extraction Algorithm (PKEA)

## Algorithm: Step 3. 학습 (Traning)

- ▶ 학습 데이터 (training data) : 학습될 문서와 그 문서의 키워드 후보군, 그리고 각 후보군의 특성값이 학습데이터로 활용됨
- ▶ 머신러닝 기반의 분류 알고리즘 중 나이브베이스 기법에 의해 모델을 학습함

Index	Candidate Phrase	TF-IDF	First Occurrence	Label (Author's Key-phrase)
1	text	1.0833	0	0
2	mining	1.1000	1	0
3	text mining	1.4555	1	1
4	refer	0.7500	3	0
5	data	0.7500	7	0
6	text data	1.4555	7	1
7	data mining	1.4666	8	1
8	text data mining	...	...	...
9	equivalent	...	...	...
10	Analytics	...	...	...
11	...	...	...	...

-- Extracting Keyphrases...  
 -- Reading instance  
 -- Converting instance  
 mining 0.5171  
 text mining 0.5171  
 data 0.5171  
 text data 0.5171  
 data mining 0.5171  
 text 0.5171  
 information 0.262  
 analytics 0.262  
 process 0.262  
 deriving 0.262  
 high-quality 0.262  
 high-quality information 0.262  
 derived 0.1455  
 involves 0.1455  
 analysis 0.0561  
 document 0.0561  
 database 0.0561  
 Typical 0.0561  
 text mining tasks include 0.0561  
 extraction 0.0561  
 predictive 0.0561  
 application 0.0561  
 ...

# Practical Key-phrase Extraction Algorithm (PKEA)

## Algorithm: Step 4. 테스트 (Testing)

- ▶ 학습된 모델을 바탕으로 새로운 문서에 대해 키워드를 추출함
- ▶ 모델이 추출한 키워드와 문서의 작성자가 선정한 키워드와 일치횟수를 측정하여 키워드 여부를 확인함
  - 테스트 결과 저자가 선정한 키워드와 20~30% 일치하는 결과를 보임
  - 20개의 적은 학습데이터 셋으로도 좋은 성능의 모델을 만들 수 있음

Protocols for secure, atomic transaction execution in electronic commerce	Neural multigrid for gauge theories and other disordered systems	Proof nets, garbage, and computations
anonymity	<i>atomicity</i>	<i>disordered</i>
<i>atomicity</i>	<i>auction</i>	<i>gauge</i>
<i>auction</i>	<i>customer</i>	<i>gauge fields</i>
<i>electronic</i>	<i>electronic</i>	<i>multigrid</i>
<i>commerce</i>	<i>commerce</i>	neural multigrid
privacy	intruder	neural networks
real-time	merchant	length scale
<i>security</i>	protocol	<i>multigrid</i>
<i>transaction</i>	<i>security</i>	smooth
	third party	
	<i>transaction</i>	

**Figure 1** Examples of author- and Kea-assigned keyphrases

# Rapidly Key-phrase Extraction Algorithm (RKEA)

## Algorithm: Step 1. 키워드 후보군 선정

- ▶ 각 텍스트를 파싱하여 키워드 후보군을 선정하는 과정
- ▶ 문서를 구분기호로 분리하고 단어의 배열로 분할함

Compatibility of systems of linear constraints over the set of natural numbers

Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types.

Manually assigned keywords:

linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets

*Figure 1.1 A sample abstract from the Inspec test set and its manually assigned keywords.*

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

*Figure 1.2 Candidate keywords parsed from the sample abstract.*

# Rapidly Key-phrase Extraction Algorithm (RKEA)

## Algorithm: Step 2. 키워드 후보군 점수 계산

- ▶ 각 키워드 후보군에 대해 점수를 계산하며 점수는 키워드의 하위 단어 점수의 합으로 계산됨
- ▶ (1) 각 키워드의 동시출현 매트릭스 생성
- ▶ (2) 각 키워드 후보군의 하위 단어에 대해 점수를 계산하고 그 합계를 최종 점수로 계산함
  - 단어 빈도 : 단어의 동시출현과 관계없이 단순히 단어가 출현한 빈도
  - 단어 정도중심성 (word degree centrality) : 단어가 함께 쓰인 단어의 수 (길이가 긴 키워드에 대해 더 높게 나타남)
  - 단어의 빈도 대비 정도중심성 비율 : 단어 정도중심성 / 빈도

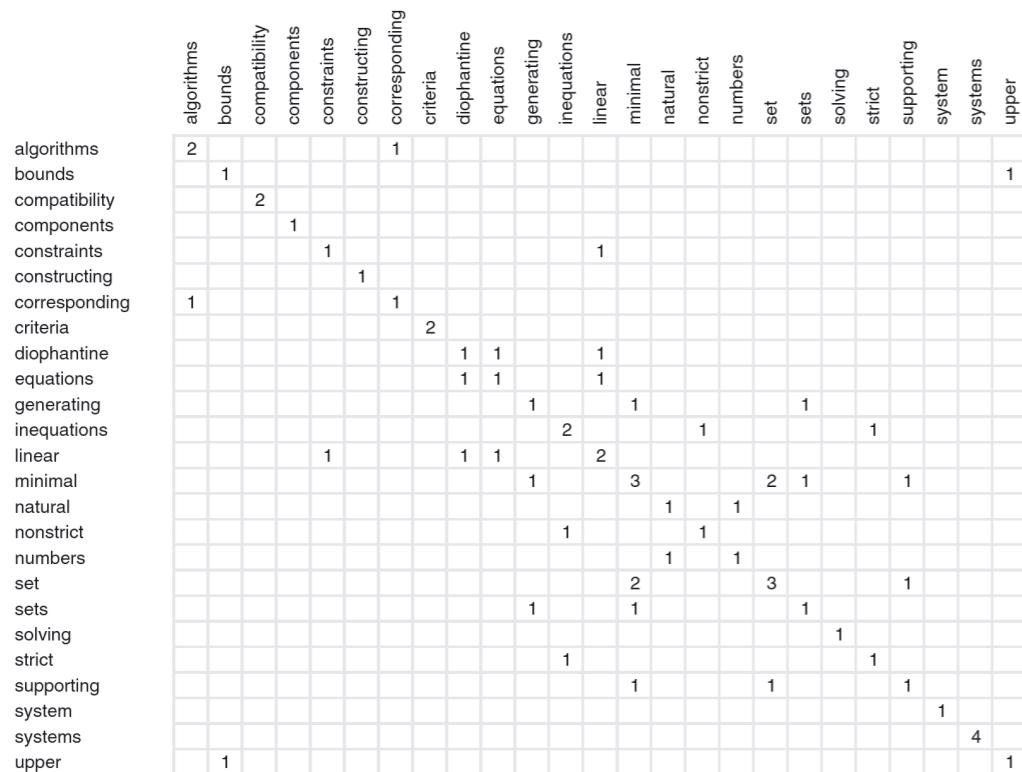


Figure 1.3 The word co-occurrence graph for content words in the sample abstract.

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper		
deg(w)	3	2	2	1	2	1	2	2	2	3	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	1	2	1	1	1	2	2	2	3	1	1	1	3	1	1	1	1	4	1	
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	3	1	2	3	1	1	2	

Figure 1.4 Word scores calculated from the word co-occurrence graph.

# Rapidly Key-phrase Extraction Algorithm (RKEA)

## Algorithm: Step 2. 키워드 후보군 점수 계산

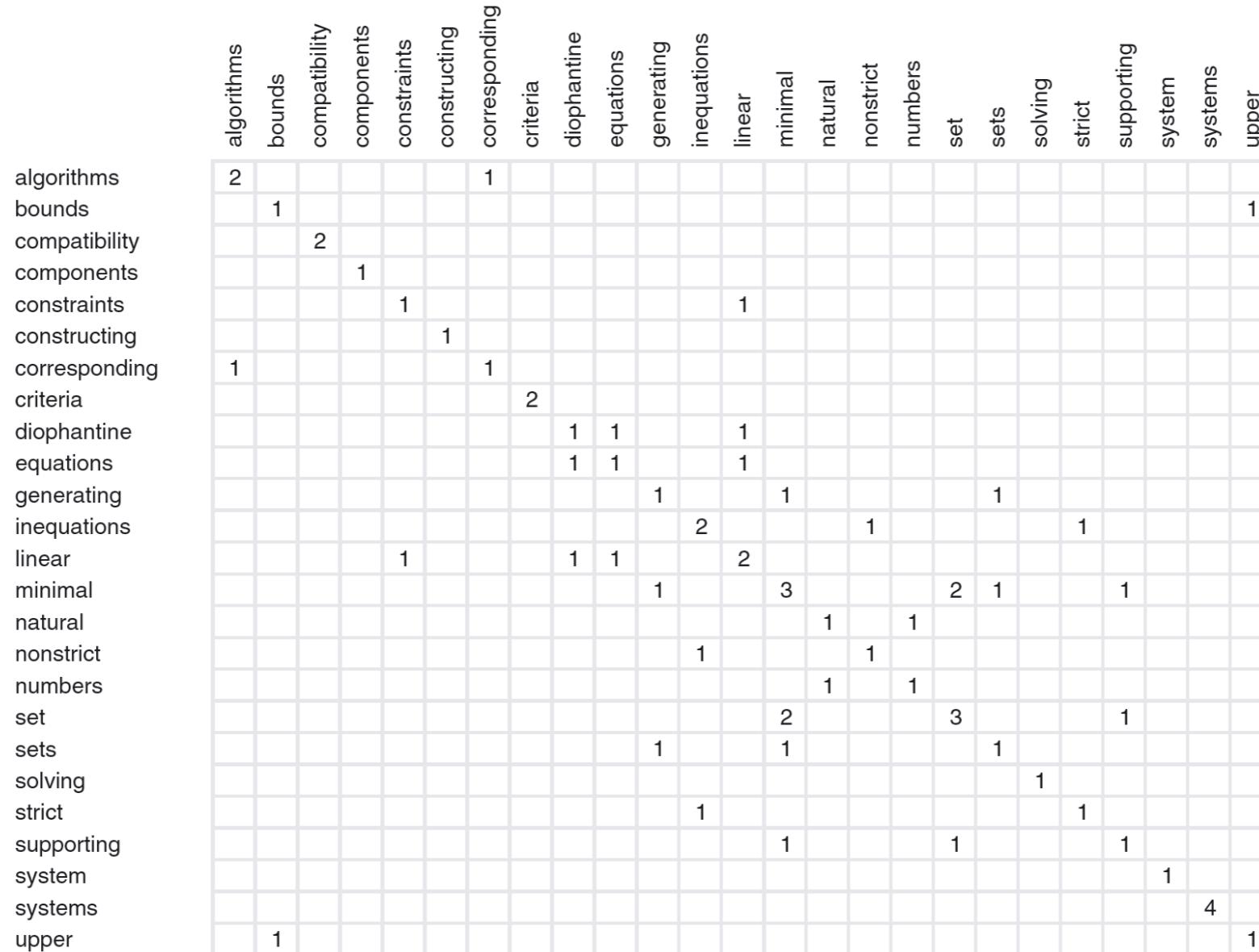


Figure 1.3 The word co-occurrence graph for content words in the sample abstract.

# Rapidly Key-phrase Extraction Algorithm (RKEA)

## Algorithm: Step 2. 키워드 후보군 점수 계산

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

Figure 1.4 Word scores calculated from the word co-occurrence graph.

minimal generating sets (8.7), linear diophantine equations (8.5), minimal supporting set (7.7), minimal set (4.7), linear constraints (4.5), natural numbers (4), strict inequations (4), nonstrict inequations (4), upper bounds (4), corresponding algorithms (3.5), set (2), algorithms (1.5), compatibility (1), systems (1), criteria (1), system (1), components (1),constructing (1), solving (1)

Figure 1.5 Candidate keywords and their calculated scores.

# Rapidly Key-phrase Extraction Algorithm (RKEA)

## Algorithm: Step 3. 최종 키워드 선정

- ▶ 각 키워드 후보군에 부여된 점수를 기준으로 최종 키워드 선정

Table 1.1 Comparison of keywords extracted by RAKE to manually assigned keywords for the sample abstract.

Extracted by RAKE	Manually assigned
minimal generating sets	minimal generating sets
linear diophantine equations	linear Diophantine equations
minimal supporting set	
minimal set	
linear constraints	linear constraints
natural numbers	
strict inequations	strict inequations
nonstrict inequations	nonstrict inequations
upper bounds	upper bounds
	set of natural numbers

# PAKE VS RAKE

구분	PAKE	RAKE
장점	<ul style="list-style-type: none"><li>- 정확도가 높음</li><li>- 학습된 문서의 특성을 반영함</li><li>- 매우 짧은 문서에도 성능이 좋음</li></ul>	<ul style="list-style-type: none"><li>- 연산이 매우 빠름</li><li>- 어구를 키워드로 취급하는 경우에 유리함</li><li>- 모든 어구 조합에 대한 가중치 계산 가능</li><li>- 언어와 문서의 특성에 관계없이 적용 가능</li></ul>
단점	<ul style="list-style-type: none"><li>- 학습을 통해 모델 구축이 필요함</li><li>- 학습에 시간이 소요되고 추출과정이 비교적 느림</li></ul>	<ul style="list-style-type: none"><li>- 어구 추출을 위해 정교한 불용어 사전 구축이 필요함</li></ul>
결과 비교	<pre>mining 0.5171 text mining 0.5171 data 0.5171 text data 0.5171 data mining 0.5171 text 0.5171 information 0.262 analytics 0.262 process 0.262 deriving 0.262 high-quality 0.262 high-quality information 0.262 derived 0.1455 involves 0.1455 analysis 0.0561 document 0.0561 database 0.0561 Typical 0.0561 text mining tasks include 0.0561</pre>	<pre>text mining task 1.841666666667 text data mining 1.466666666667 text mining 1.455555555556 input text 1.388888888889 entity relation 1.333333333333 entity extraction 1.333333333333 classification purpose 1.333333333333 search index 1.333333333333 information extraction 1.2 high-quality information 1.2 information retrieval 1.2 language processing 1.166666666667 association analysis 1.133333333333 sentiment analysis 1.133333333333 pattern learning 1.111111111111 mining 1.1 text 1.083333333333 document set 1.0 information 0.8 data 0.75</pre>



**E.O.D**