

Text Mining 2

Contents

1. Document Representation
2. Document Similarity
3. Document Classification

Review

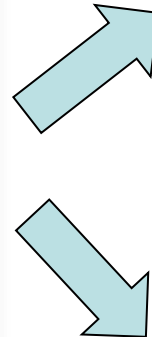
- 비정형 데이터를 ‘단어’ 단위로 처리
- 갖가지 전처리 기법 (stemming, lemmatization으로 각 단어 토큰화)

한화시스템

Morphological Analysis

- 시제, 수, 성별 등에 따라 다르게 표현되는 단어를 통일하는 과정
- 예) car → cars, give → gives, gave, given
- Stemming & Lemmatization

Word	stemming	lemmatization
Love	Lov	Love
Loves	Lov	Love
Loved	Lov	Love
Loving	Lov	Love
Innovation	Innovat	Innovation
Innovations	Innovat	Innovation
Innovate	Innovat	Innovate
Innovates	Innovat	Innovate
Innovative	Innovat	Innovative



한화시스템

Stemming

- 단어를 어근으로 변환
- 장점: 간단하고 빠르다 (형태소 분석이 불필요)
- 단점: 정보 손실이 너무 큼

Word	stemming	lemmatization
Love	Lov	Love
Loves	Lov	Love
Loved	Lov	Love
Loving	Lov	Love
Innovation	Innovat	Innovation
Innovations	Innovat	Innovation
Innovate	Innovat	Innovate
Innovates	Innovat	Innovate
Innovative	Innovat	Innovative

한화시스템

Lemmatization

- 단어를 lemma로 변환
- 장점: 정보 손실이 작음 (품사가 보존)
- 단점: 처리 시간이 오래 걸림 (형태소 분석이 포함)

Word	stemming	lemmatization
Love	Lov	Love
Loves	Lov	Love
Loved	Lov	Love
Loving	Lov	Love
Innovation	Innovat	Innovation
Innovations	Innovat	Innovation
Innovate	Innovat	Innovate
Innovates	Innovat	Innovate
Innovative	Innovat	Innovative

Review

- 비정형 데이터를 ‘단어’ 단위로 처리
- 갖가지 전처리 기법 (pos-tagging으로 각 단어의 품사 탐색)

POS Tagging



- Part of speech (POS) tagging
- Token이 같아도 tag는 다를 수 있음
 - 예) ✓ I love you. → "love" is a **verb**
 - ✓ All you need is love. → "love" is **noun**
- 좋은 형태소 분석기가 NLP 연구의 시작점

Natural language processing (NLP) is a field of computer science

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

JJ NN NN -LRB- NN -RRB- VBZ DT NN IN NN NN

POS Tagging



- Supervised learning
 - Input: sentence
 - Output: tag sequence
 - Model: SVM, Decision Tree, ...

Natural language processing (NLP) is a field of computer science

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

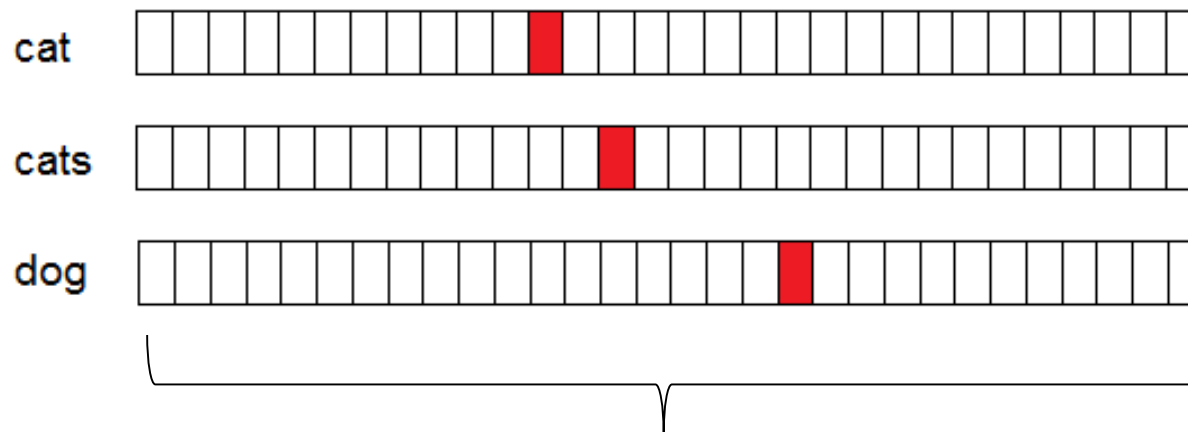
classifier classifier

"processing" = NN? VBG? JJ? "computer" = NN? VBG? JJ?

Review

- 비정형 데이터를 ‘단어’ 단위로 처리
- 단어 representation

1) 단어를 벡터로 바꾸는 방법 (one-hot encoding)



처리하고자 하는 모든 단어의 수

Review

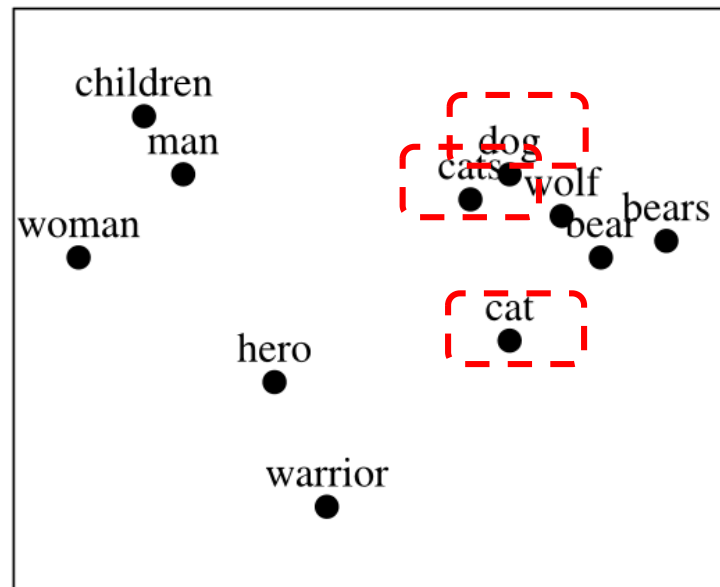
- 비정형 데이터를 ‘단어’ 단위로 처리
- 단어 representation

2) Word2Vec: 의미 공간에 각 단어의 좌표값을 지정

cat \rightarrow (1.4, 0)

cats \rightarrow (1.4, 0.8)

dog \rightarrow (1.4, 1)



의미 공간

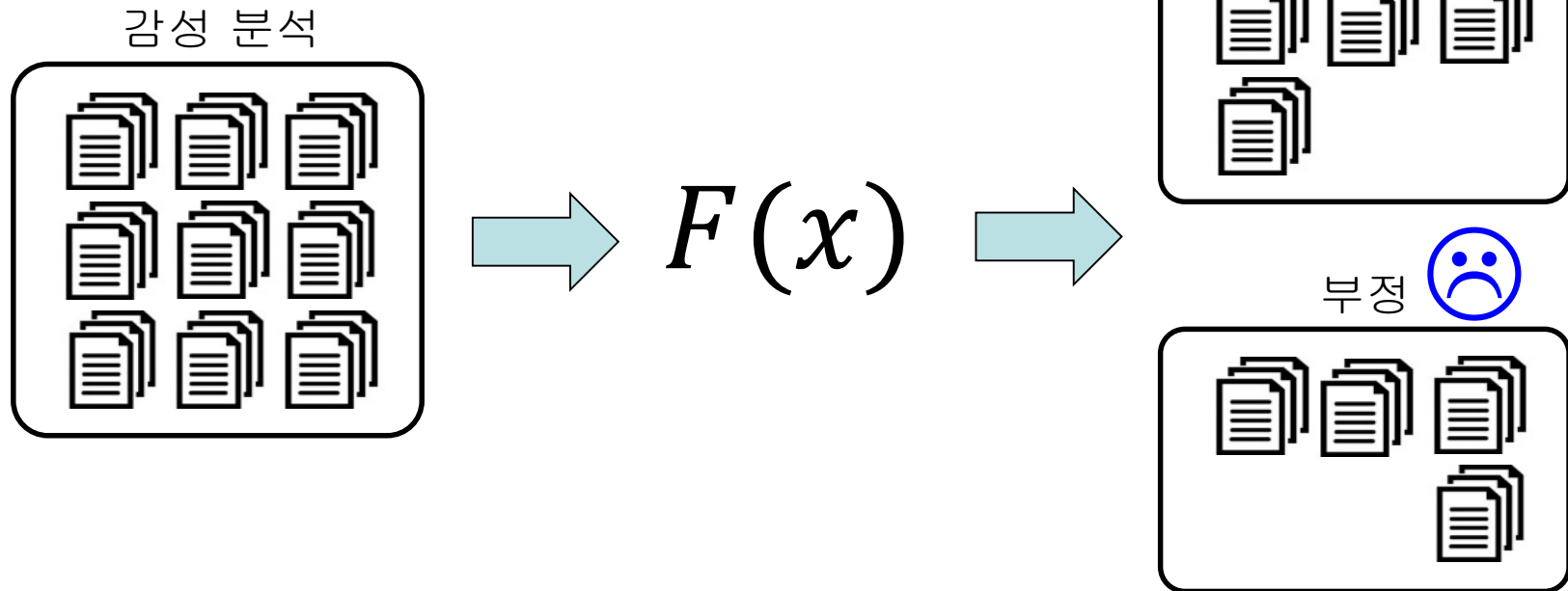
Review

- 왜?
- 이것을 가지고 실생활에서 어떤 것을 풀 수 있을까
 - Document Classification
 - Document Clustering
 - Document Summarization
 - Machine Translation
 - Question Answering
 - Machine Reading Comprehension
 - ...

Review

- 왜?
- 이것을 가지고 실생활에서 어떤 것을 풀 수 있을까

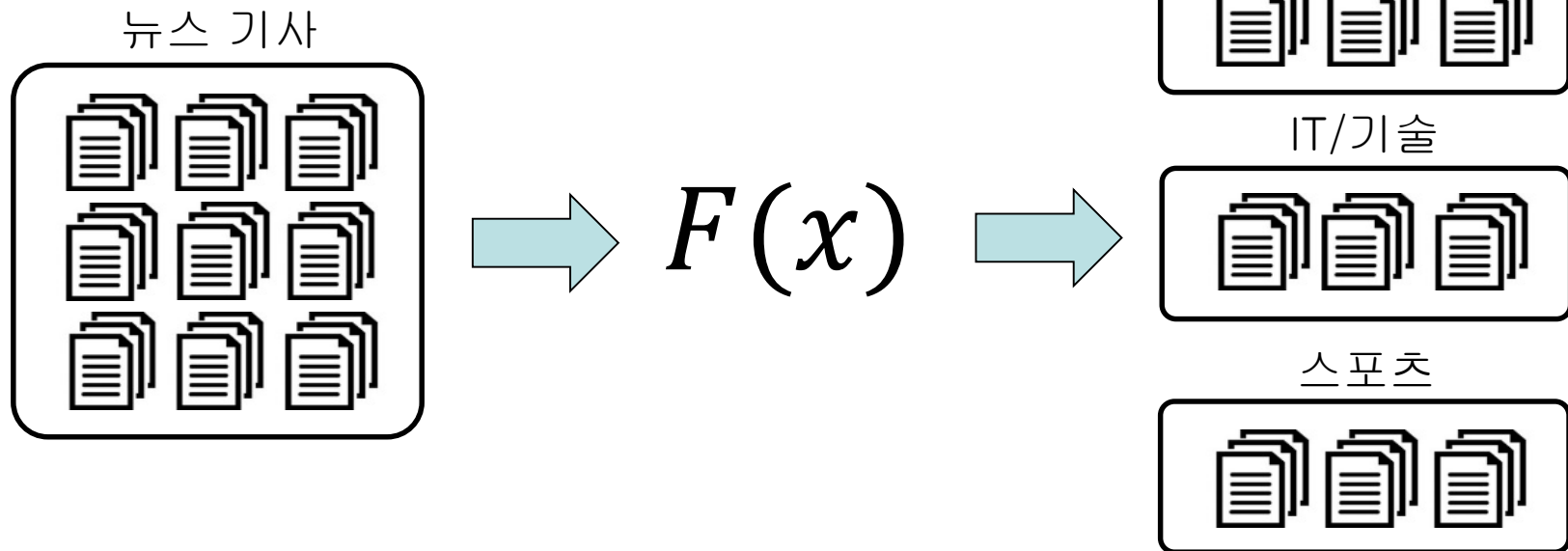
- Document Classification



Review

- 왜?
- 이것을 가지고 실생활에서 어떤 것을 풀 수 있을까

- Document Classification

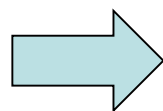
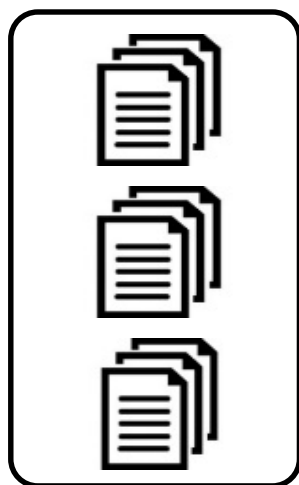


Review

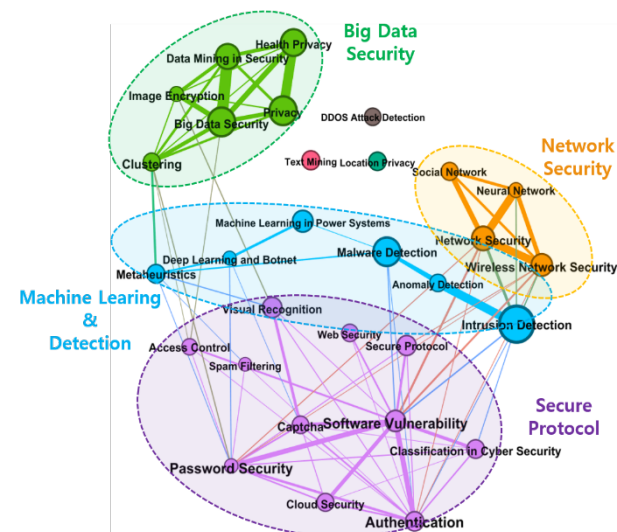
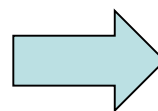
- 왜?
- 이것을 가지고 실생활에서 어떤 것을 풀 수 있을까

Document Clustering

10,000 건



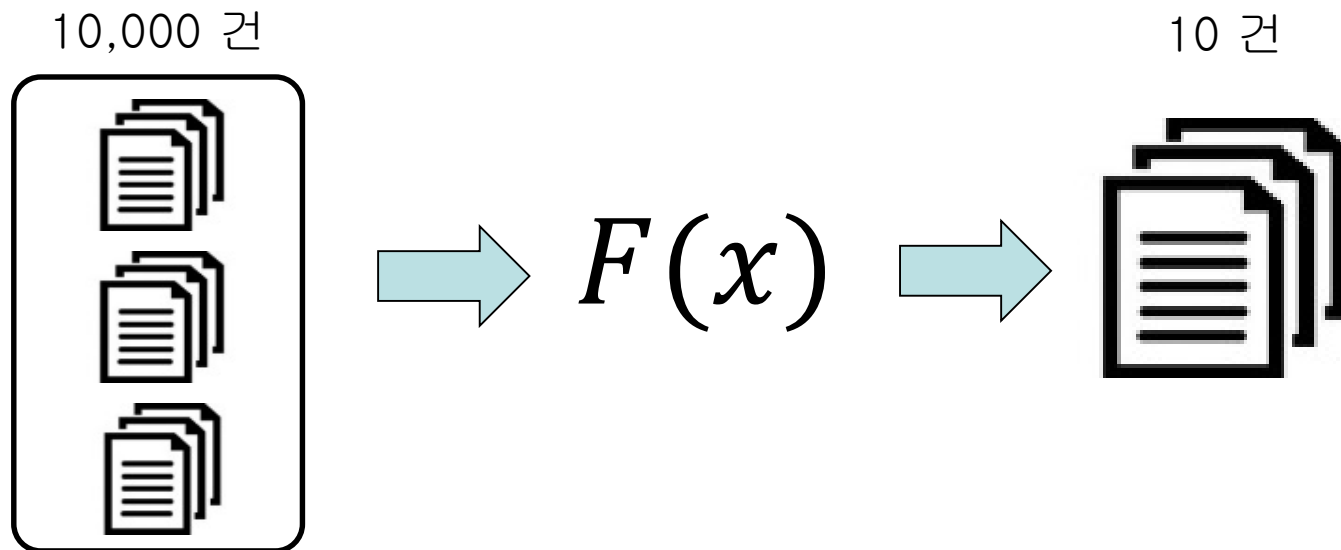
$$F(x)$$



Review

- 왜?
- 이것을 가지고 실생활에서 어떤 것을 풀 수 있을까

- Document Summarization



Review

- 왜?
- 이것을 가지고 어떤 것을 풀 수 있을까

- Document Classification
- Document Clustering
- Document Summarization
- Machine Translation
- Question Answering
- ...



문서 단위로 처리 되어야하는
문제 상황들이 많음



문서를 정형 데이터로 표현해보자

Text Mining Task

- 여러 방식으로 접근해 볼 수 있음
 - 1) 특정 단어의 공통 빈도 수
 - 2) 빈도 수 변형
 - 3) 딥러닝 기반 문서 임베딩
 - 4) ...

Document Representation

- 문서 집합에서의 단어의 빈도 수 사용
- 이유: 같은 의미 공간 내에서 단어의 의미를 찾는 것이 가능

배

	단어 1	단어 2	단어 3	단어 4
문서 1	배	타다		화물
문서 2	배		먹다	맛있게

중국집 메뉴

	단어 1	단어 2	단어 3	단어 4
문서 1	중국집	메뉴	짜장면	짬뽕
문서 2	?	?	짜장면	짬뽕

Document Representation

- 문서 집합에서의 단어의 빈도 수 사용
- 이유: 같은 의미 공간 내에서 단어의 의미를 찾는 것이 가능

배

	단어 1	단어 2	단어 3	단어 4
문서 1	배	타다		화물
문서 2	배		먹다	맛있게

중국집 메뉴

	단어 1	단어 2	단어 3	단어 4
문서 1	중국집	메뉴	짜장면	짬뽕
문서 2	?	?	짜장면	짬뽕

각 문서의 **배**가 어떤 것을 뜻하는 것인지 알 수 있음

Document Representation

- 문서 집합에서의 단어의 빈도 수 사용
- 이유: 같은 의미 공간 내에서 단어의 의미를 찾는 것이 가능

배

	단어 1	단어 2	단어 3	단어 4
문서 1	배	타다		화물
문서 2	배		먹다	맛있게

중국집 메뉴

	단어 1	단어 2	단어 3	단어 4
문서 1	중국집	메뉴	짜장면	짬뽕
문서 2	?	?	짜장면	짬뽕

? 에 들어가는 것이 무엇인지 추측 가능

Document Representation

- 빈도 수를 기반한 term-document matrix
- 각 문서를 벡터로 변환

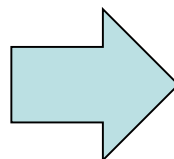
the complic evol landscap of cancer mutat pose a
 for mine textual pattern in news tweet paper and mani
 list oth
 prio tex
 to a dep
 too the
 base on
 curv and
 curat da
 oncosco
 oncosco
 priorit o

문서 3

문서 2

문서 1

this paper is a tutori on formal concept analysi fca and
 it applic fca is an appli branch of lattic theori a
 mathemat disciplin which enabl formalis of concept as
 basic unit of human think and analys data in the
 objectattribut form origin in earli s dure the last three
 decad it becam a popular humancentr tool for
 knowledg represent and data analysi with numer applic
 sinc the tutori was special prepar for russir the cover
 fca topic includ inform retriev with a focus on visualis
 aspect machin learn data mine and knowledg discoveri
 text mine and sever other

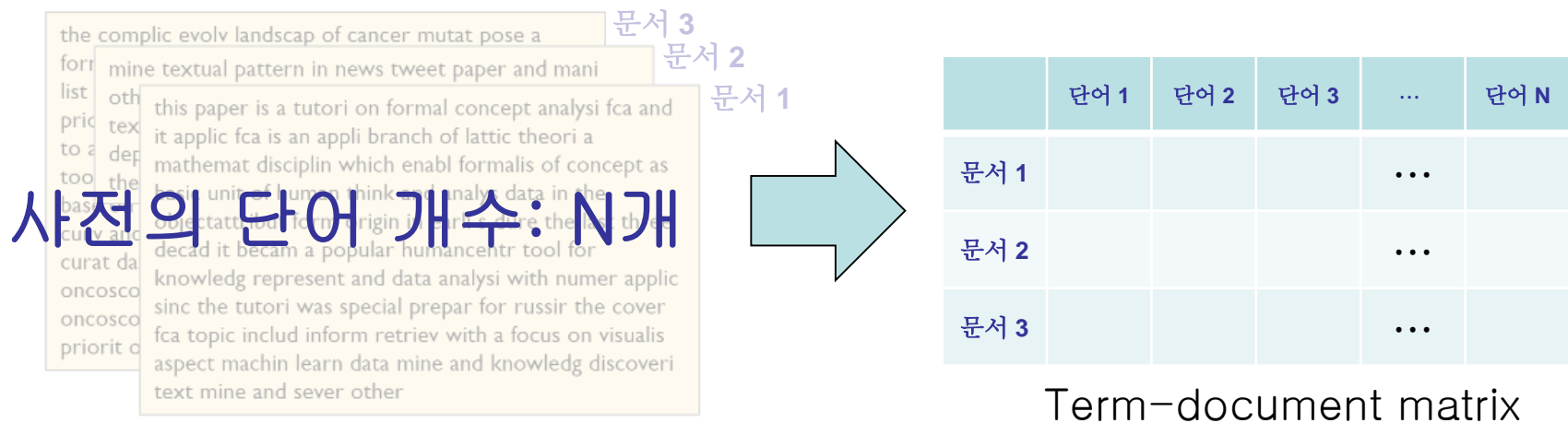


	단어 1	단어 2	단어 3	...	단어 N
문서 1				...	
문서 2				...	
문서 3				...	

Term-document matrix

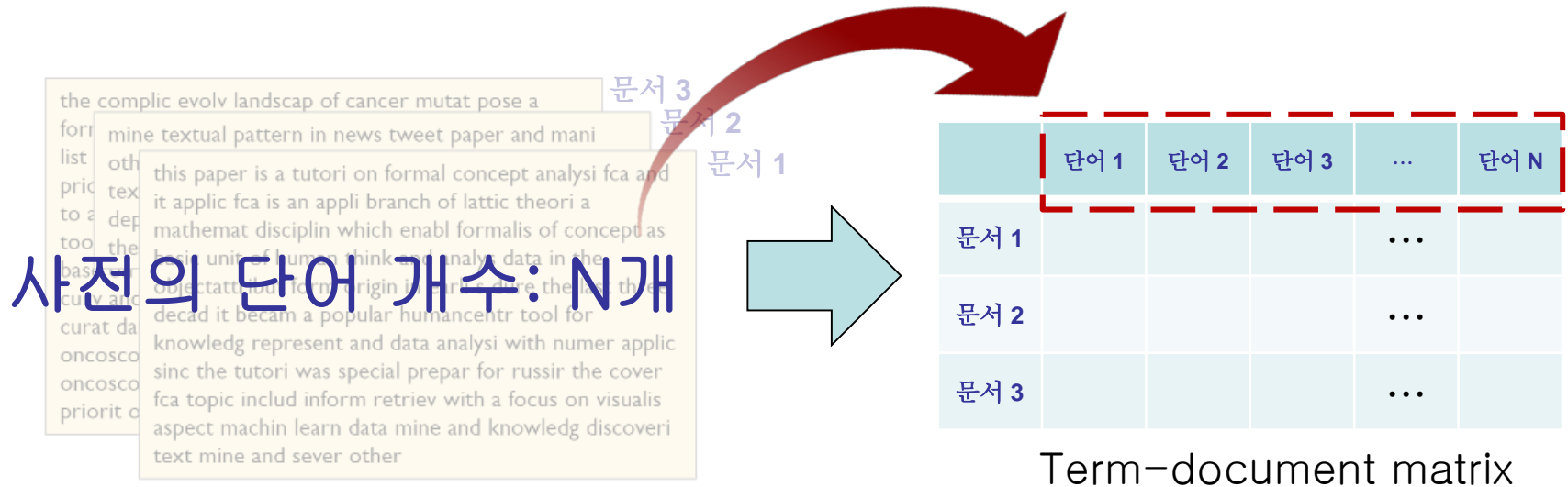
Document Representation

- 빈도 수를 기반한 term-document matrix
- 각 문서를 벡터로 변환



Document Representation

- 빈도 수를 기반한 term-document matrix
- 모든 문서의 단어가 변수가 됨



Document Representation

- 각 문서에 나타나는 단어 빈도수에 따라 문서 표현
- Bag-of-Word(BoW)라고도 불림

the complic evolv landscap of cancer mutat pose a
form mine textual pattern in news tweet paper and mani
list oth
prio tex
to a dep
too the
base on
curv and
curat da
oncosco
oncosco
priorit c

this paper is a tutori on formal concept analysi fca and
it applic fca is an appli branch of lattic theori a
mathemat disciplin which enabl formalis of concept as
basic unit of human think and analys data in the
objectattribut form origin in earli s dure the last three
decad it becam a popular humancentr tool for
knowledg represent and data analysi with numer applic
sinc the tutori was special prepar for russir the cover
fca topic includ inform retriev with a focus on visualis
aspect machin learn data mine and knowledg discoveri
text mine and sever other

Term-document matrix

	단어 1	단어 2	단어 3	...	단어 N
문서 1	2	0	1	...	30
문서 2	0	7	0	...	12
문서 3	2	0	0	...	8

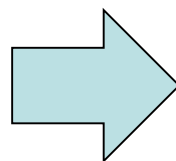


Document Representation

- 각 문서에 나타나는 단어 빈도수에 따라 문서 표현
- Bag-of-Word(BoW)라고도 불림

the complic evolv landscap of cancer mutat pose a
form mine textual pattern in news tweet paper and mani
list oth
prio tex
to a dep
too the
base on
curv and
curat da
oncosco
oncosco
priorit o

this paper is a tutori on formal concept analysi fca and
it applic fca is an appli branch of lattic theori a
mathemat disciplin which enabl formalis of concept as
basic unit of human think and analys data in the
objectattribut form origin in earli s dure the last three
decad it becam a popular humancentr tool for
knowledg represent and data analysi with numer applic
sinc the tutori was special prepar for russir the cover
fca topic includ inform retriev with a focus on visualis
aspect machin learn data mine and knowledg discoveri
text mine and sever other



Term-document matrix

	단어 1	단어 2	단어 3	...	단어 N
문서 1	2	0	1	...	30
문서 2	0	7	0	...	12
문서 3	2	0	0	...	8



문서 1 = [2, 0, 1, ..., 0]

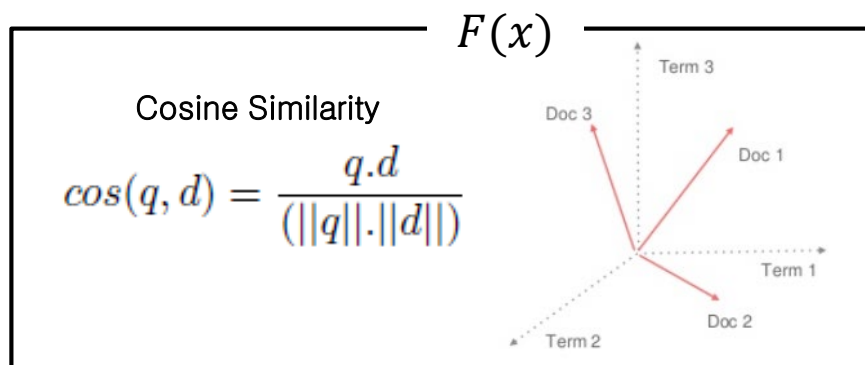
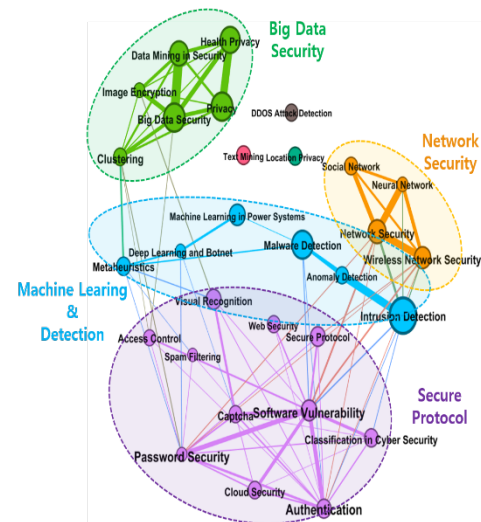
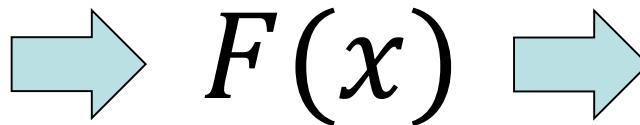
문서 2 = [0, 7, 0, ..., 12]

문서 3 = [2, 0, 0, ..., 8]

Document Representation

- 빈도수로 벡터화된 문서들이 학습되어 모델이 만들어짐
- 활용 예제 1
 - Document Clustering

문서 1 = [2, 0, 1, ..., 0]
 문서 2 = [0, 7, 0, ..., 12]
 문서 3 = [2, 0, 0, ..., 8]
 ⋮
 문서 N = [1, 0, 0, ..., 16]



Document Representation

- 빈도수로 벡터화된 문서들이 학습되어 모델이 만들어짐
- 활용 예제 2
 - Document Summarization

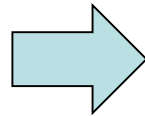
문서 **1** = [2, 0, 1, \dots , 0]

문서 **2** = [0, 7, 0, \dots , 12]

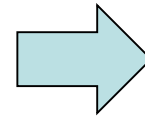
문서 **3** = [2, 0, 0, \dots , 8]

\vdots

문서 **N** = [1, 0, 0, \dots , 16]



$F(x)$



“Hanhwa”

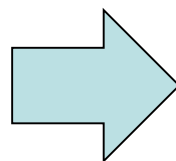
가장 많이 나온 단어가 가장 대표적인 단어이라는 가정

Document Representation

- 가장 많이 나온 단어의 수로 전체 문서를 표현 가능
- Term frequency per document

the complic evolv landscap of cancer mutat pose a
form mine textual pattern in news tweet paper and mani
list oth
prio tex
to a dep
too the
base on
curv and
curat da
oncosco
oncosco
priorit o

this paper is a tutori on formal concept analysi fca and
it applic fca is an appli branch of lattic theori a
mathemat disciplin which enabl formalis of concept as
basic unit of human think and analys data in the
objectattribut form origin in earli s dure the last three
decad it becam a popular humancentr tool for
knowledg represent and data analysi with numer applic
sinc the tutori was special prepar for russir the cover
fca topic includ inform retriev with a focus on visualis
aspect machin learn data mine and knowledg discoveri
text mine and sever other



Term-document matrix

	단어 1	단어 2	단어 3	...	단어 N
문서 1	2	0	1	...	30
문서 2	0	7	0	...	12
문서 3	2	0	0	...	8
빈도 수	4	7	1		50

위 3개의 문서는 N 번째 단어로 표현 될 수 있음

Document Representation

- Term frequency의 문제점
 - 1) Is, can, the, of 와 같은 단어들은 빈도수가 높아도 중요치 않음
 - 2) term-document matrix가 너무 sparse 해짐

1) 중요하지 않은 단어

	text	can	mining	...	is ✓
문서 1	2	0	1	...	30
문서 2	0	7	0	...	12
문서 3	2	0	0	...	8

2) Sparsity 문제

	text	can	mining	...	is
문서 1	2	0	1	...	30
문서 2	0	7	0	...	12
문서 3	2	0	0	...	8

Document Representation

- Word Weighting: Inverse Document Frequency
- 전체적으로 희소한 단어가 특정 문서의 중요한 단어일 수 있다는 가정



문서 A

	단어 1	단어 2	...	단어 N
문서 1	12	0	...	5



문서 B

	단어 1	단어 2	...	단어 N
문서 1	2	0	...	0



문서 C

	단어 1	단어 2	...	단어 N
문서 1	2	0	...	0

전체 문서 집합을 고려했을 때 문서 A의 대표 단어는 N번째 단어

Document Representation

- 단어 Frequency-Inverse Document Frequency (TF-IDF)
- Term Frequency + Inverse Document Frequency

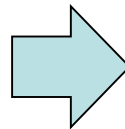
$$TF - IDF(w) = \underbrace{tf(w)}_{\text{Term Frequency}} \times \log \left(\frac{N}{\underbrace{df(w)}}_{\text{Inverse Document Frequency}} \right)$$

w 가 해당 문서에서 많이 나타나면 증가

w 가 다른 문서에 덜 나타나면 증가

Term Frequency matrix

	text	can	mining	...	is
문서 1	2	0	1	...	0
문서 2	0	7	0	...	12
문서 3	2	0	0	...	8



TF-IDF calculated

	text	can	mining	...	is
문서 1	5.25	1.54	3.18	...	0
문서 2	8.2	7	3.1	...	1.32
문서 3	6.1	0	4.15	...	1.9

Document Representation

- 단어 Frequency-Inverse Document Frequency (TF-IDF)
- Term Frequency + Inverse Document Frequency

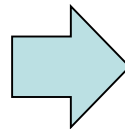
$$TF - IDF(w) = \underbrace{tf(w)}_{\text{Term Frequency}} \times \log \left(\frac{N}{\underbrace{df(w)}}_{\text{Inverse Document Frequency}} \right)$$

w 가 해당 문서에서 많이 나타나면 증가

w 가 다른 문서에 덜 나타나면 증가

Term Frequency matrix

	text	can	mining	...	is
문서 1	2	0	1	...	0
문서 2	0	7	0	...	12
문서 3	2	0	0	...	8



TF-IDF calculated

	text	can	mining	...	is
문서 1	5.25	1.54	3.18	...	0
문서 2	8.2	7	3.1	...	1.32
문서 3	6.1	0	4.15	...	1.9

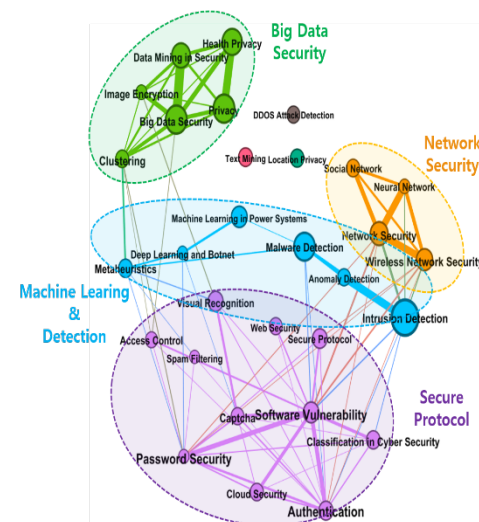
Document Representation

- Clustering과 Summarization 활용

TF-IDF Calculated

	text	can	mining	...	is
문서 1	5.25	1.54	3.18	...	0
문서 2	8.2	7	3.1	...	1.32
문서 3	6.1	0	4.15	...	1.9

$$\Rightarrow F(x) \Rightarrow$$

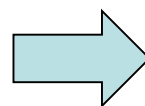


Document Representation

- Clustering과 Summarization 활용

TF-IDF Calculated

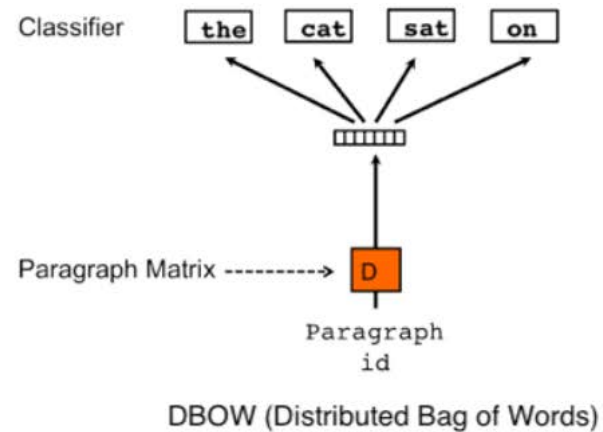
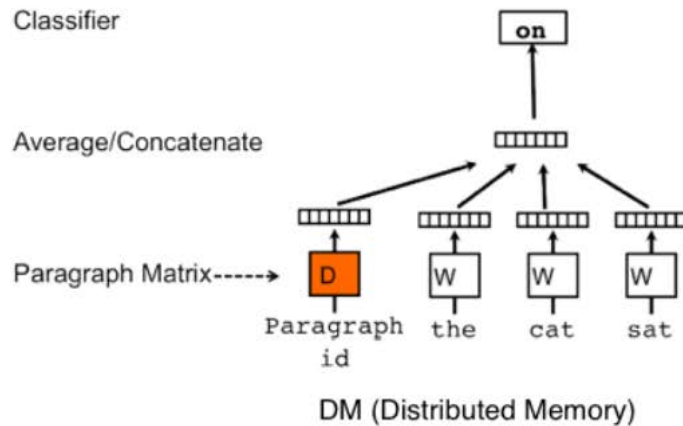
	text	can	mining	...	is
문서 1	5.25	1.54	3.18	...	0
문서 2	8.2	7	3.1	...	1.32
문서 3	6.1	0	4.15	...	1.9
빈도 수	19.55	8.54	10.43	0	3.22



“Text”

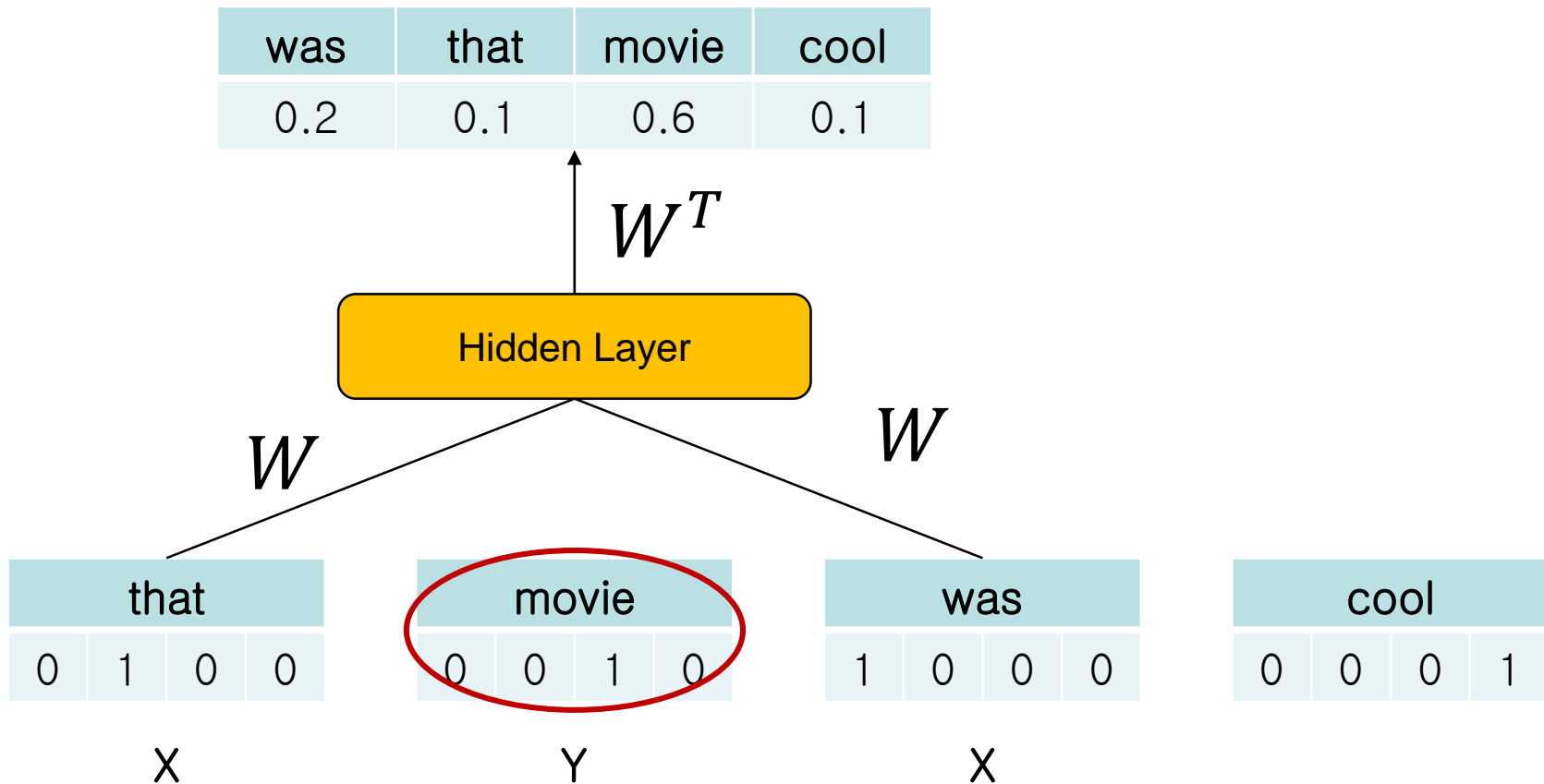
Document Representation

- Document Embedding: Doc2vec
- 문서 벡터를 단어처럼 학습시키고자 하는 목표
- 같은 단어가 쓰여도 각 문서의 의도 및 내용이 다르면 의미 파악 불가
- Word2vec과 비슷한 학습 방식을 가지고 있음
- DBOW 가 DM 보다 보통 좋은 성능을 냄



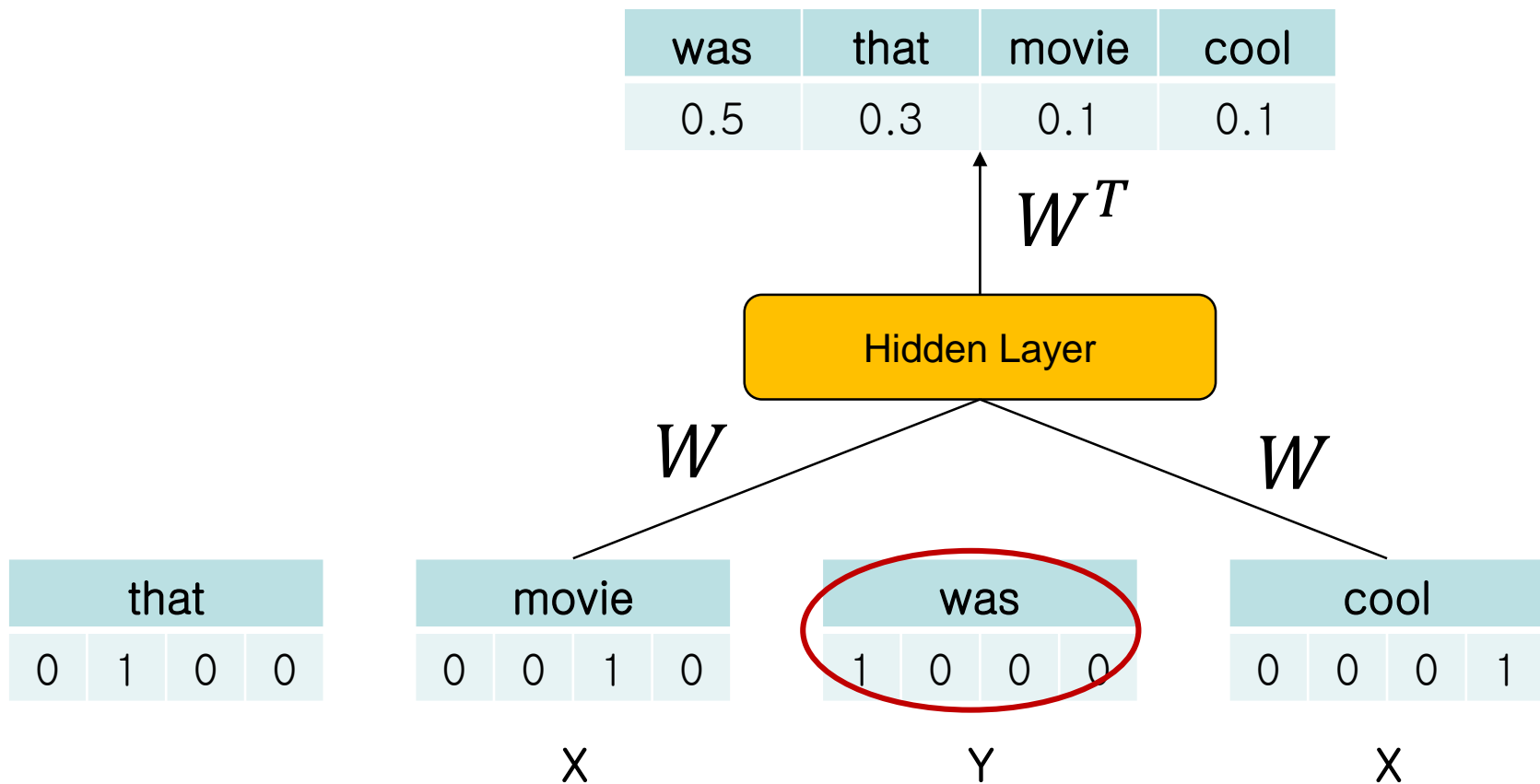
Word2Vec: CBOW 구조

- 임의의 단어를 주변 단어로 예측하는 신경망 모델 구성



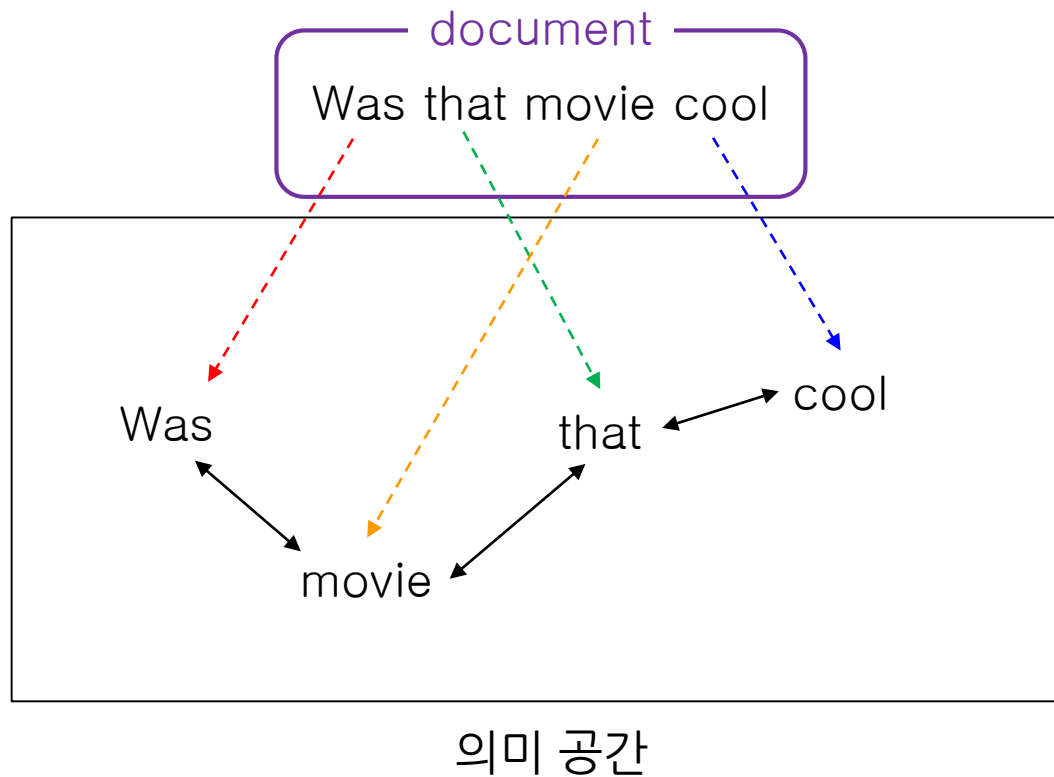
Word2Vec: CBOW 구조

- 임의의 단어를 주변 단어로 예측하는 신경망 모델 구성
- Hidden Layer의 unit를 단어를 표현하는 저차원의 새로운 변수로 해석



Word2Vec: CBOW 구조

- 문장 내 각 단어를 같은 의미 공간에서 좌표값을 갖게 만듦



Doc2Vec: PV-DM 구조

- 임의의 단어를 주변 단어로 예측하는 신경망 모델 구성

Paragraph ID	was	that	movie	cool
0.5	0.8	0.2	0.3	0.1

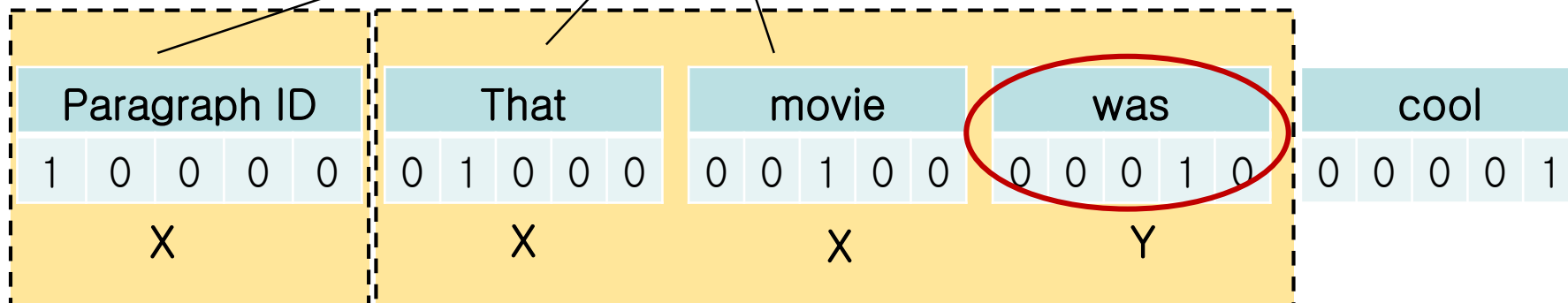
W^T

Hidden Layer

W

W

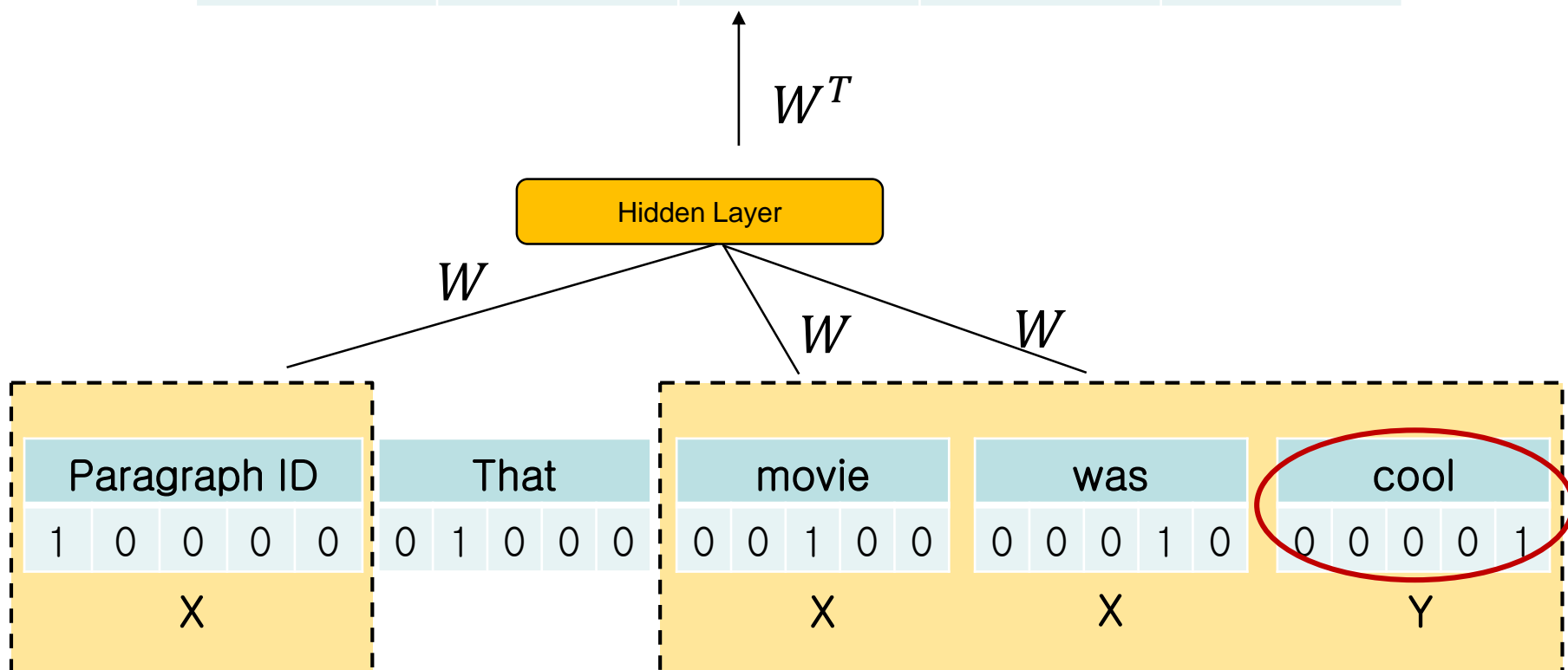
W



Doc2Vec: PV-DM 구조

- 임의의 단어를 주변 단어로 예측하는 신경망 모델 구성

Paragraph ID	was	that	movie	cool
0.3	0.2	0.1	0.1	0.9



Doc2Vec: PV-DBOW 구조

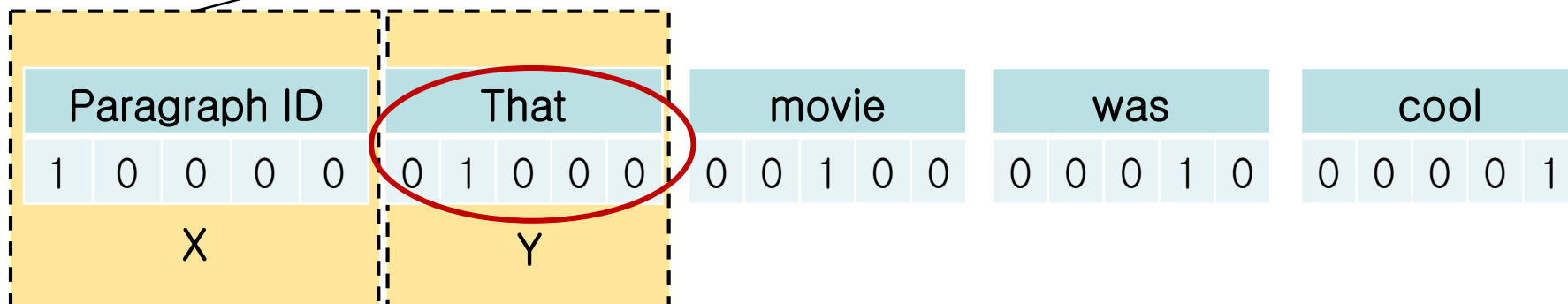
- 임의의 단어를 주변 단어로 예측하는 신경망 모델 구성

Paragraph ID	was	that	movie	cool
0.3	0.2	0.8	0.1	0.2

W^T

Hidden Layer

W



Doc2Vec: PV-DBOW 구조

- 임의의 단어를 주변 단어로 예측하는 신경망 모델 구성

Paragraph ID	was	that	movie	cool
0.3	0.2	0.1	0.9	0.1

W^T

Hidden Layer

W

Paragraph ID
1 0 0 0 0

X

That
0 1 0 0 0

movie
0 0 1 0 0

Y

was
0 0 0 1 0

cool
0 0 0 0 1

Doc2Vec: PV-DBOW 구조

- 임의의 단어를 주변 단어로 예측하는 신경망 모델 구성

Paragraph ID	was	that	movie	cool
0.3	0.7	0.1	0.1	0.3

W^T

Hidden Layer

W

Paragraph ID
1 0 0 0 0
X

That
0 1 0 0 0

movie
0 0 1 0 0

was
0 0 0 1 0
Y

cool
0 0 0 0 1

Doc2Vec: PV-DBOW 구조

- 임의의 단어를 주변 단어로 예측하는 신경망 모델 구성

Paragraph ID	was	that	movie	cool
0.3	0.2	0.1	0.1	0.9

W^T

Hidden Layer

W

Paragraph ID
1 0 0 0 0
X

That
0 1 0 0 0

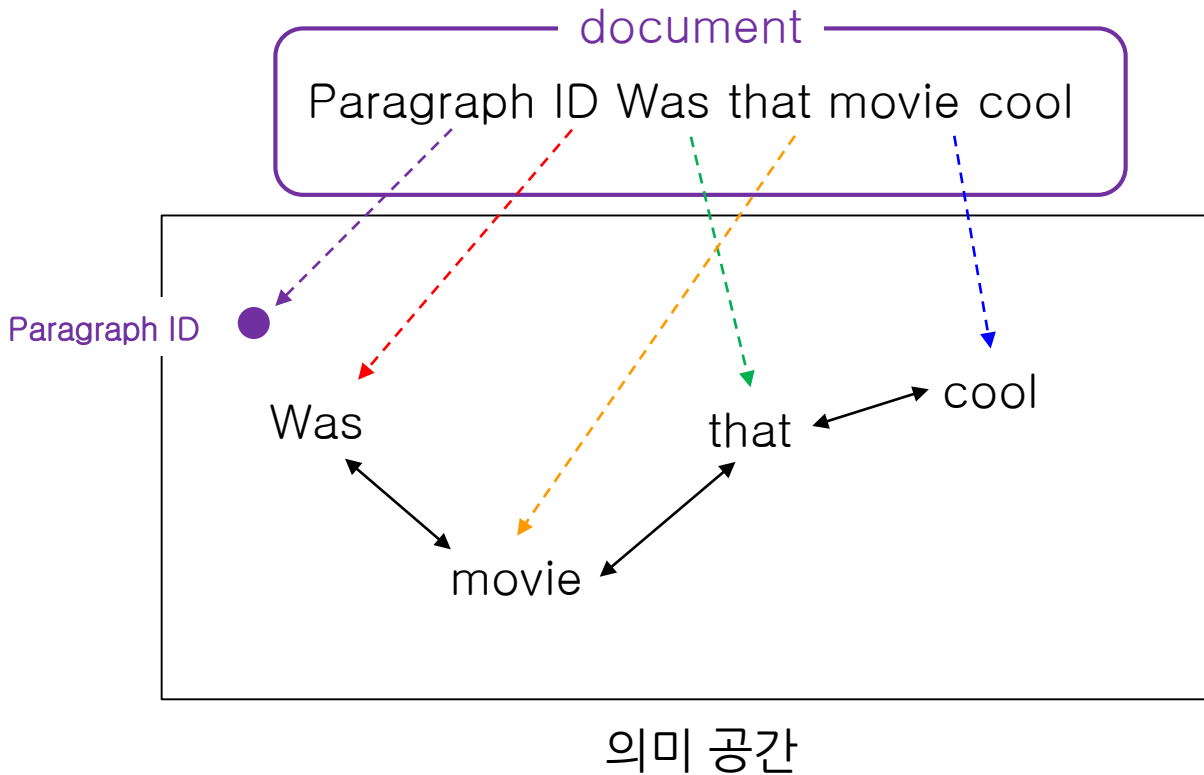
movie
0 0 1 0 0

was
0 0 0 1 0

cool
0 0 0 0 1
Y

Doc2Vec: PV-DBOW 구조

- Doc2vec: 각 단어와 Paragraph ID 를 동시에 학습
- Paragraph ID 도 단어들과 같은 의미 공간에 좌표값을 지님



Document Representation

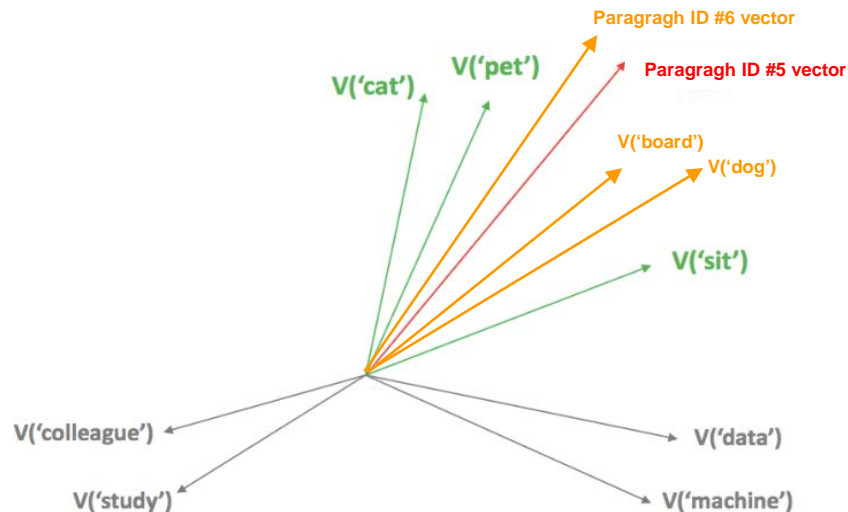
- 단어가 달라도 의미 공간 상의 단어 벡터들이 유사함
- 따라서, Paragraph ID 의 벡터도 유사하게 됨

문서 #5 Sentence 1 : A little cat sit on the table.

↕ 유사

↕ 유사

문서 #6 Sentence 2 : A little dog sit on the board.



Document Representation

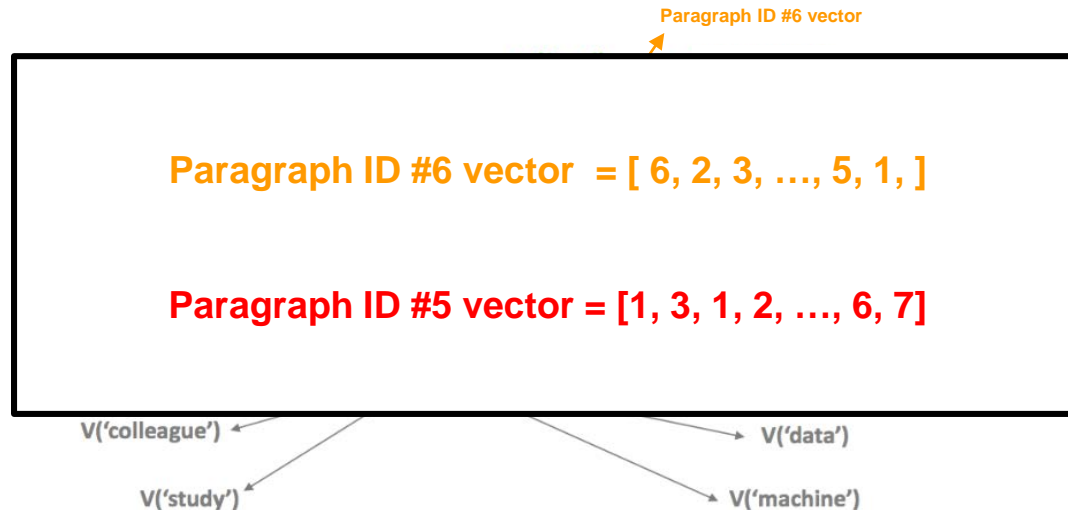
- 단어가 달라도 의미 공간 상의 단어 벡터들이 유사함
- 따라서, Paragraph ID 의 벡터도 유사하게 됨

문서 #5 Sentence 1 : A little cat sit on the table.

↑ 유사

↑ 유사

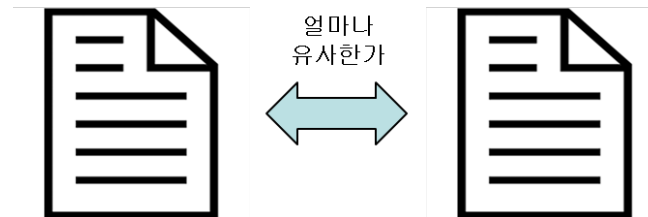
문서 #6 Sentence 2 : A little dog sit on the board.



Document Similarity

- 가정

1. 두 문서 간 유사성은 공유되는 feature가 많을수록 증가
2. 개별 feature는 서로 독립
3. 각 feature가 포함되어 있는 개념 영역이 비슷해야 함
4. 유사도가 높다면, 문서의 의미도 비슷함



단어 frequency

	단어 1	단어 2	단어 3	...	단어 N
문서 1	2	0	1	...	0
문서 2	0	7	0	...	12
문서 3	2	0	0	...	8

Binary

	단어 1	단어 2	단어 3	...	단어 N
문서 1	1	0	1	...	0
문서 2	0	1	0	...	1
문서 3	1	0	0	...	1

Document Similarity

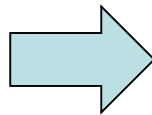
- Common features model
- 두 문서에 동시에 등장한 단어 수를 전체 단어 수로 나누어 구함

$$S_{doc1,doc2} = \frac{a}{a + b + c + d}$$

Binary

	Term 1	Term 2	Term 3	...	Term N
Doc 1	1	0	1	...	0
Doc 2	0	1	0	...	1
Doc 3	1	0	0	...	1

Doc1 Doc2	Y	N
Y	a	b
N	c	d



Common feature	문서 1	문서 2	문서 3
문서 1		0.2	0.2
문서 2			0.4
문서 3			

Document Similarity

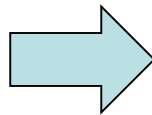
- Ratio model
- 두 문서에 모두 나타나지 않은 단어는 제외하고 계산

$$S_{doc1,doc2} = \frac{a}{a + b + c}$$

Binary

	Term 1	Term 2	Term 3	...	Term N
Doc 1	1	0	1	...	0
Doc 2	0	1	0	...	1
Doc 3	1	0	0	...	1

Doc1 Doc2	Y	N
Y	a	b
N	c	d



Common feature	문서 1	문서 2	문서 3
문서 1		0.25	0.2
문서 2			0.4
문서 3			

Document Similarity

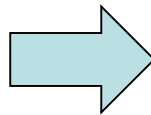
- Jaccard similarity
- 두 문서에 모두 나타나지 않은 단어는 제외하고 계산

$$S_{doc1, doc2} = \frac{\sum_k \min(x_{ik}, x_{jk})}{\sum_k \max(x_{ik}, x_{jk})}$$

Term frequency

	Term 1	Term 2	Term 3	...	Term N
Doc 1	2	0	1	...	0
Doc 2	0	7	0	...	12
Doc 3	2	0	0	...	8

Doc1 Doc2	Y	N
Y	a	b
N	c	d



Common feature	문서 1	문서 2	문서 3
문서 1		3/11	2/12
문서 2			2/9
문서 3			

Document Similarity

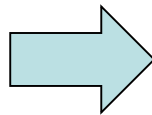
- Cosine similarity
- 의미 공간 안의 벡터 간 내적 (각도) 계산

$$S_{doc1,doc2} = \frac{\sum_k (x_{ik} \times x_{jk})}{\sqrt{(\sum_k x_{ik}^2)(\sum_k x_{jk}^2)}}$$

Term frequency

	Term 1	Term 2	Term 3	...	Term N
Doc 1	2	0	1	...	0
Doc 2	0	7	0	...	12
Doc 3	2	0	0	...	8

Doc1 Doc2	Y	N
Y	a	b
N	c	d



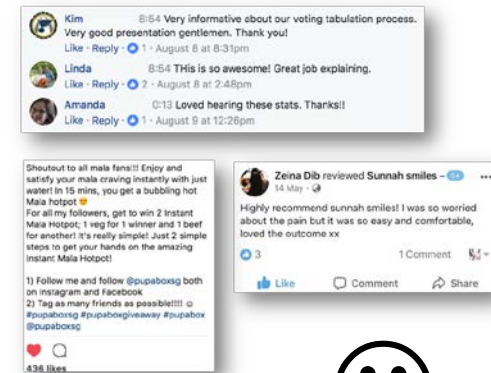
Common feature	문서 1	문서 2	문서 3
문서 1		0.68	0.3254
문서 2			0.3380
문서 3			

Document Classification

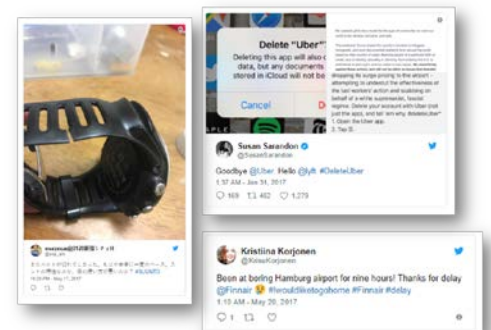
Sentimental Analysis

- 문서 감성 분석(Supervised Learning)
- 문제 상황: Movie Review 문서들을 보고 긍정인지 부정인지 파악

Positive



Negative



Document Representation


- N-grams
- 구 혹은 절 단위로 tokenize 하는 방법

Language Model

$$P(w_n | w_{n-1}, w_{n-2}, \dots, w_1) = \frac{P(w_n, w_{n-1}, w_{n-2}, \dots, w_1)}{P(w_{n-1}, w_{n-2}, \dots, w_1)}$$

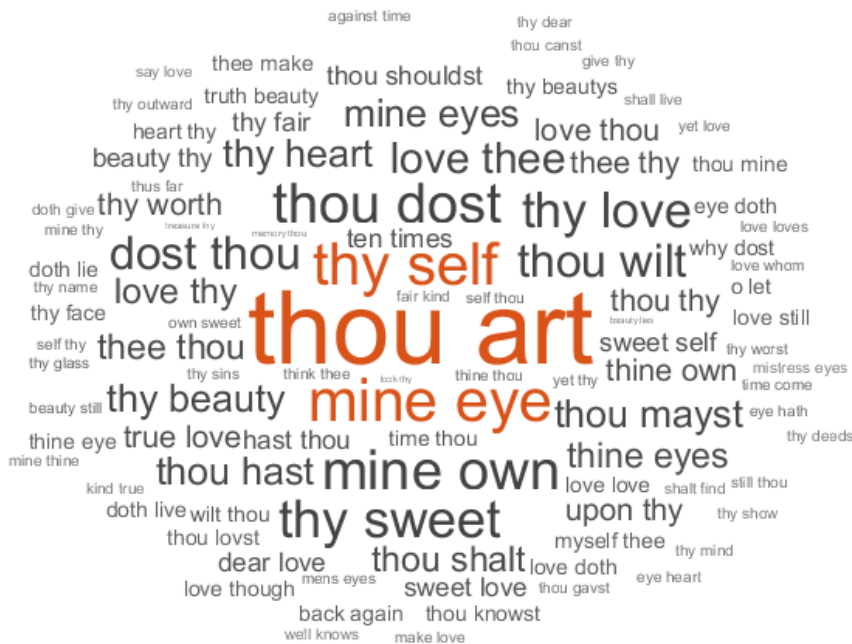
Q) One of the hottest topics in artificial intelligence is deep _____

1. deep
2. learning
3. supply
4. chain
5. big
6. data

- 
1. deep learning
 2. 6 sigma
 3. supply chain
 4. big data

Document Representation

- N-gram은 도메인 정보를 반영하기 때문에 문서 표현에 용이
- 따라서, Document classification과 Document clustering에 효과적임

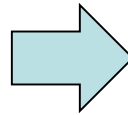


Document Representation

- N-gram은 도메인 정보를 반영하기 때문에 문서 표현에 용이
- 따라서, Document classification과 Document clustering에 효과적임

Term document matrix

	text	Mining	Is	...	fun
문서 1				...	
문서 2				...	
문서 3				...	



Term document matrix

	Text mining	is	...	fun
문서 1			...	
문서 2			...	
문서 3			...	

Document Classification

Sentimental Analysis

- 감성 레이블이 부여된 문서 활용
- 긍정: 1, 부정: -1

Document Reviews (train)	감성
films adapted from comic books have had plenty of ...	1
moviemaking is a lot like being the general manager of an ...	1
your first clue that something isn't gonna be quite right with the movie	-1
every now and then a movie comes along from a suspect ...	1
carry on at your convenience is all about the goings on in the factory of a toilet manufacturer ...	-1
you've got mail works a lot better than it deserves to ...	1
the title is taken from the writings of ralph waldo emerson ...	-1

Document Classification

Sentimental Analysis

- 감성 레이블이 부여된 문서 활용
- 긍정: 1, 부정: -1

Train

Test

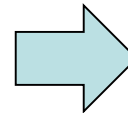
Document Reviews (train)	감성
films adapted from comic books have had plenty of ...	1
moviemaking is a lot like being the general manager of an ...	1
your first clue that something isn't gonna be quite right with the movie	-1
every now and then a movie comes along from a suspect ...	1
carry on at your convenience is all about the goings on in the factory of a toilet manufacturer ...	-1
you've got mail works a lot better than it deserves to ...	1
the title is taken from the writings of ralph waldo emerson ...	-1

Document Classification

Sentimental Analysis

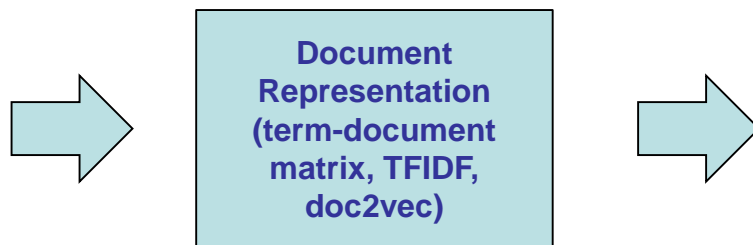
- 감성 레이블이 부여된 문서 활용
- 긍정: 1, 부정: -1

Document Reviews (train)	
films adapted from comic books have had plenty of ...	Train
moviemaking is a lot like being the general manager of an ...	
your first clue that something isn't gonna be quite right with the movie	
every now and then a movie comes along from a suspect ...	
carry on at your convenience is all about the goings on in the factory of a toilet manufacturer ...	
you've got mail works a lot better than it deserves to ...	Test
the title is taken from the writings of ralph waldo emerson ...	



전처리

Term document matrix



	films	Adapt	Comic	...	emerson
문서 1				...	
⋮				...	
문서 N				...	

Train

Test

Document Classification

Sentimental Analysis

- Document reviews에서 감정을 나타내는 부사 및 형용사 추출
- 내용 기반 classification할 때에는 명사 추출

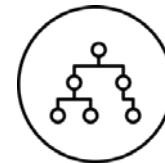
Document Reviews (train)	감성
films adapted from comic books have had plenty of ...	1
moviemaking is a lot like being the good manager of an ...	1
your first clue that something isn't gonna be quite right with the movie	-1
every now and then a movie better comes along from a suspect ...	1
carry on at your convenience is all about just the goings on in the factory of a toilet manufacturer ...	-1
you've got mail works a lot better so it deserves to ...	1
the title is mistakenly taken from writings of Ralph Waldo Emerson ...	-1

Document Classification

Sentimental Analysis

- Train 데이터로 분류 모델 학습

Train_x



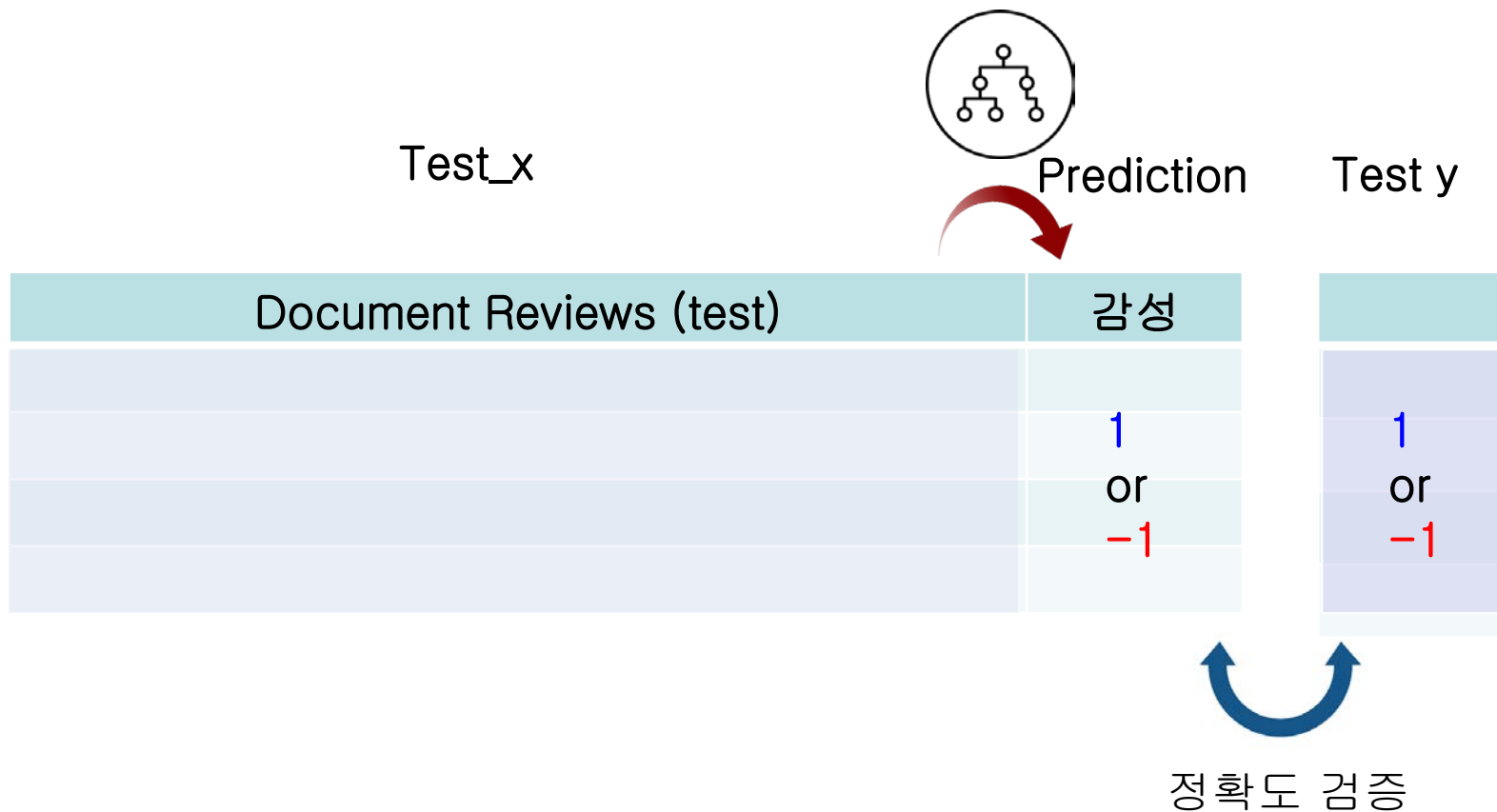
Train x

Document Reviews (train)		감성
		-1

Document Classification

Sentimental Analysis

- Test 데이터로 분류 모델 검증



실습

- 튜토리얼 (영어 감성 분석)

전처리 기법: pos tagging 명사 추출, 불필요한 text 제거

토큰 단위 기초 통계

English Movie review 데이터 감성 분석

학습된 doc2vec으로 유사 문서 찾기 (d2v_pretrain.py)

- 실습

한국어 영화 후기 데이터 감성 분석

- 형태소 분석기: 한나눔 형태소 분석기

- TF-IDF Term document matrix

- 분류 모델 성능 검증

EOD