Unstructured Data Mining

8기, 연세대학교 응용통계학과 김현중

=Text Mining

8기, 연세대학교 문헌정보학과 김현중

마이닝 기법 개요

1. 데이터 종류

1) Structured: 정형 데이터

- 정해진 구조에 따라 데이터가 배열되어 있음
- Ex) 데이터베이스

2) Semi-Structured: 반정형 데이터

- 완벽한 구조는 아니지만, 어느 정도 모양이 갖춰진 데이터
- 태그를 사용하는 웹 페이지에서 많이 보이는 형태
- Ex) HTML, XML

3) Unstructured: 비정형 데이터

- 형식이 없는 자유 데이터
- 빅 데이터의 대부분을 차지

Ex) 텍스트, 이미지, 동영상, 기타 등등

마이닝 기법 개요

2. 데이터 탐색 방법

1) Search: 검색

- 기존 데이터를 필요에 맞게 가공하는 것
- Ex) SQL Query, 웹 검색

2) Discovery: 발견

- 데이터가 갖고 있는 숨겨진 의미를 탐색하는 것
- Ex) 데이터 마이닝, 텍스트 마이닝

마이닝 기법 개요

3. 데이터 종류 x 데이터 탐색법

	Structured	Semi-Structured	Unstructured
Search	데이터베이스 (Database)	정보 검색 (Information Retrieval)	
Discovery	데이터 마이닝 (Data Mining)	웹 마이닝 (Web Mining)	텍스트 마이닝 (Text Mining)

비정형 데이터 마이닝

1. 비정형 마이닝 종류

1) 텍스트 마이닝

- 거대한 텍스트 데이터 셋에서 흥미로운 규칙을 찾는 것
- 흥미로운 규칙: 평범하지 않거나, 숨겨져 있거나, 앞으로 유용한 것
- Ex) 문장 의미 추출, 감성 분석, 초록 추출 등

2) 이미지, 동영상 마이닝

- 텍스트 마이닝의 부분집합
- Ex) 이미지, 동영상의 설명을 가져와 텍스트 마이닝
- Ex) 딥 러닝 기법을 이용한 이미지 인식

비정형 데이터 마이닝

1. 비정형 마이닝 종류

3) 음악 마이닝

- 악보를 텍스트로 변환해 마이닝

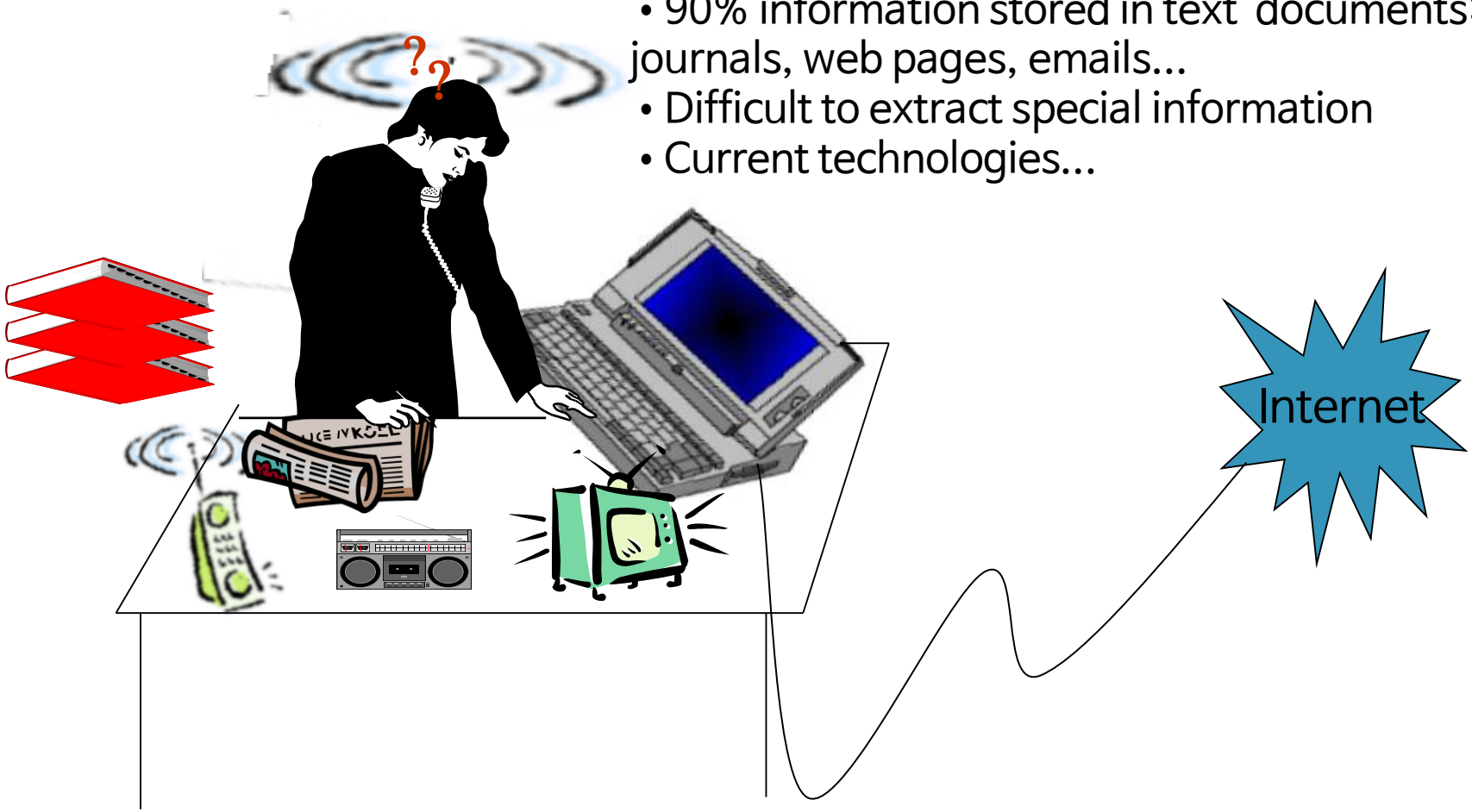
4) 웹 마이닝

- HTML, XML 등 웹 페이지 마이닝
- Semi-Structured Data지만 넓게 봤을 때 텍스트 마이닝에 포함

비정형 데이터 마이닝

2. 텍스트 마이닝 개요

- Information explosive
- 90% information stored in text documents: journals, web pages, emails…
- Difficult to extract special information
- Current technologies…



비정형 데이터 마이닝

2. 텍스트 마이닝 개요

It is necessary to automatically analyze, organize,summarize…

비정형 데이터 마이닝

3. 텍스트 마이닝 정의

- **Text Mining**

the procedure of **synthesizing the information**
by analyzing the relations, the patterns, and the rules among
textual data (semi-structured or unstructured text)

- **기법 (Techniques)**

- 데이터 마이닝, 머신 러닝, 정보 검색, 통계학, NLP 등등
- 기존의 정형 데이터 마이닝 기법 활용

- **기능**

- 문서 요약
- 문서 분류/군집
- 특성 추출

비정형 데이터 마이닝

3. 텍스트 마이닝 정의

- 왜 새로운 마이닝 분야가 생겨났을까요?
- 텍스트 데이터 분석이 어려운 이유?

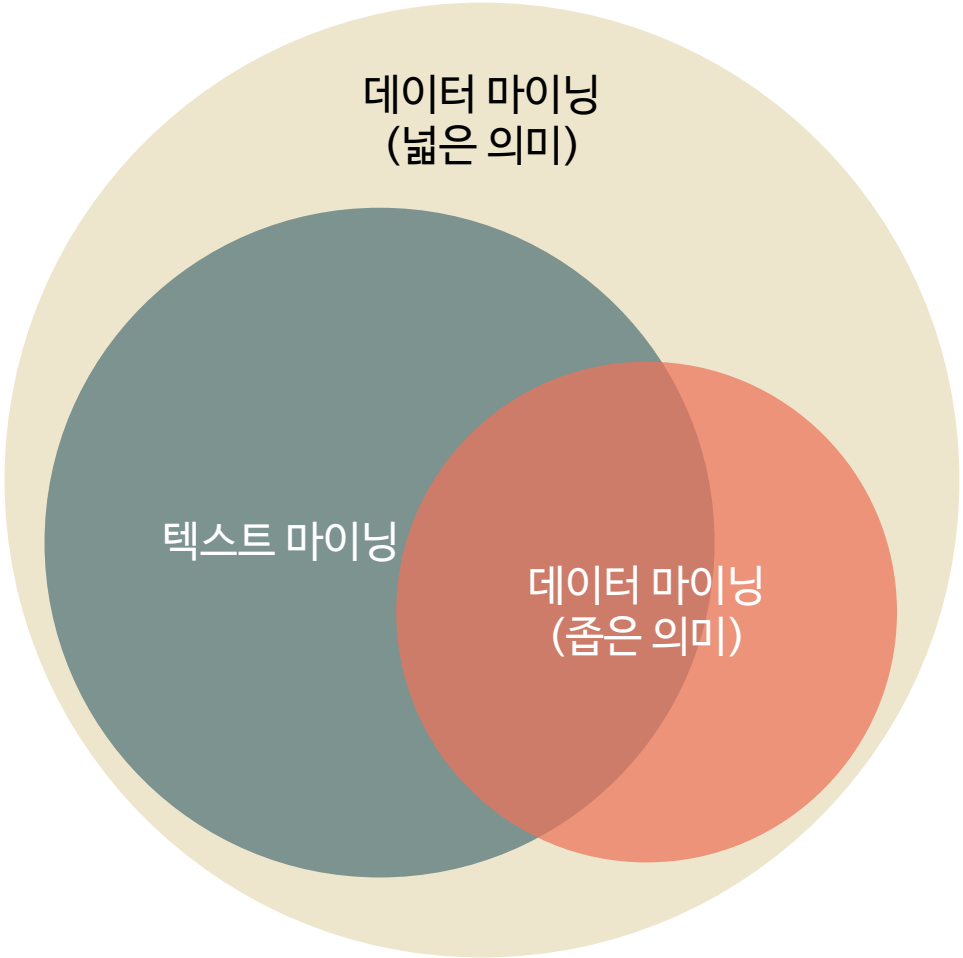
비정형 데이터 마이닝

4. 데이터 마이닝 vs 텍스트 마이닝

	Data Mining	Text Mining
목적 데이터	수치 & 명목 데이터	텍스트 데이터
데이터 형식	정형	반정형/비정형
데이터 표현	직관적	복잡함
차원(Attribute)	수천 개 이내	측정 불가
사용 도구	통계학, 머신 러닝 등	데이터 마이닝 + 정보 검색, NLP, …
역사	1994년 출발	2000년대 초반
시장 크기	대기업/중기업 대상	기업 및 개인 대상

비정형 데이터 마이닝

4. 데이터 마이닝 vs 텍스트 마이닝



비정형 데이터 마이닝

4. 데이터 마이닝 vs 텍스트 마이닝

- **텍스트 마이닝 to 데이터 마이닝**

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire, is finding itself hard pressed to cope with the crisis…

Relation Extraction

Disease Outbreaks relation

Date	*Disease Name*	*Location*
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.
May 1995	Ebola	Zaire

„We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex.“

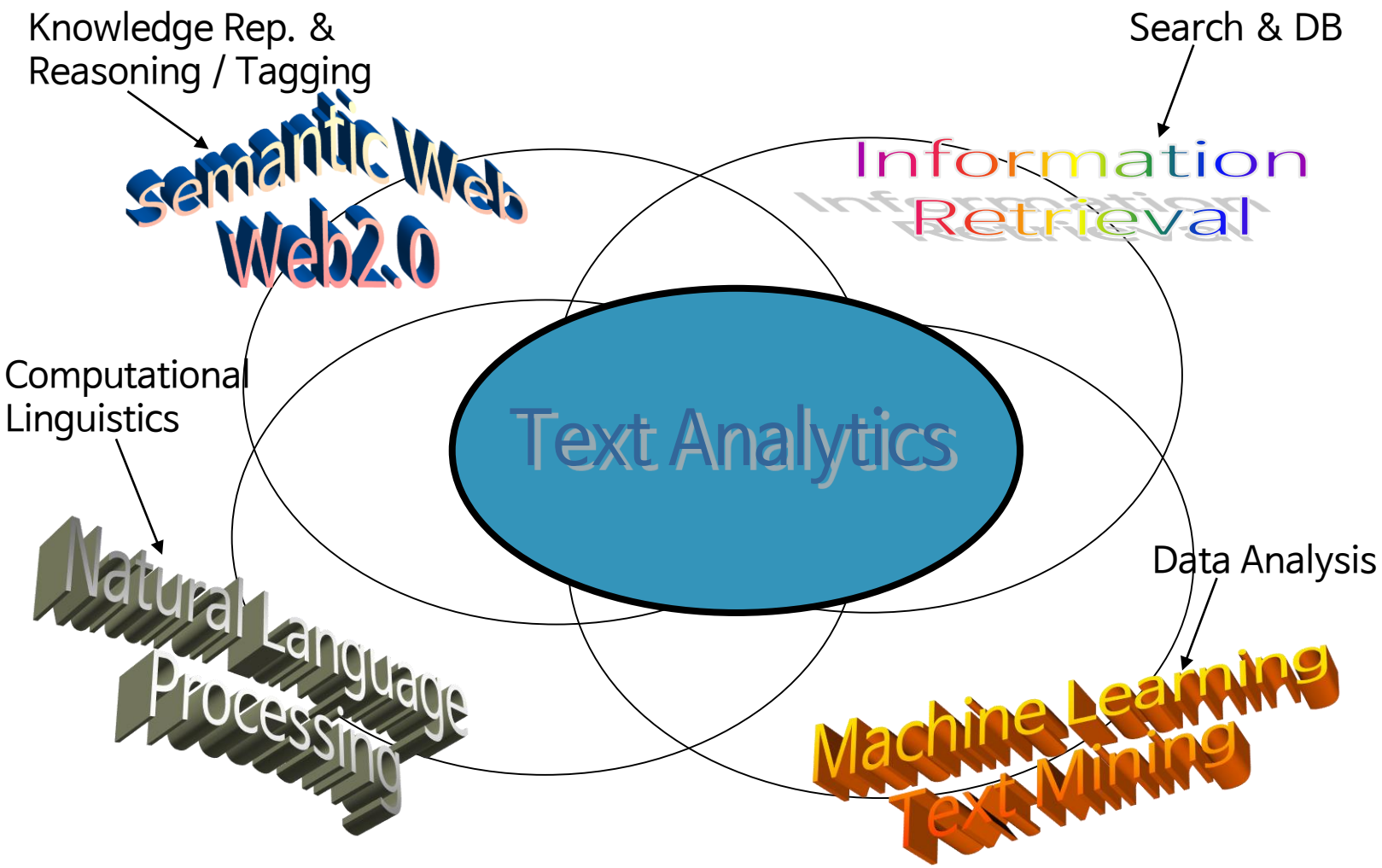
CBF-A <-- interact / complex --> CBF-C

CBF-B --- associates ---> CBF-A-CBF-C complex

[From AliBaba]

비정형 데이터 마이닝

5. 텍스트 마이닝 활용 분야



비정형 데이터 마이닝

5. 텍스트 마이닝 활용 분야

- **잠재적으로 적용할 수 있는 부분은 셀 수 없음**

- 고객 관리
- 트렌드 분석
- 사건 추적
- 정보 필터링
- 뉴스 분류
- 웹 검색
- 기타 등등…

비정형 데이터 마이닝

6. 텍스트 마이닝 과정(1): 자질 추출

- **자질 추출(Feature Extraction)?**

- Attribute(Word) Extraction
- 텍스트 데이터에서 중요한 단어를 선별하는 과정(**데이터 전처리**)
- 추출한 단어로 Corpus 구성(데이터 관리의 기본 구조)
- 자질 추출 방법에 따라 마이닝 성능이 달라짐
- Parsing(Tokenization, Stopword Removing), POS tagging 등을 포함
- 사전(Dictionary)을 이용해 자질을 부여하기도 함(Ex. SentiWordNet)
- 한글의 경우, 형태소 분석 필요

비정형 데이터 마이닝

6. 텍스트 마이닝 과정(1): 자질 추출

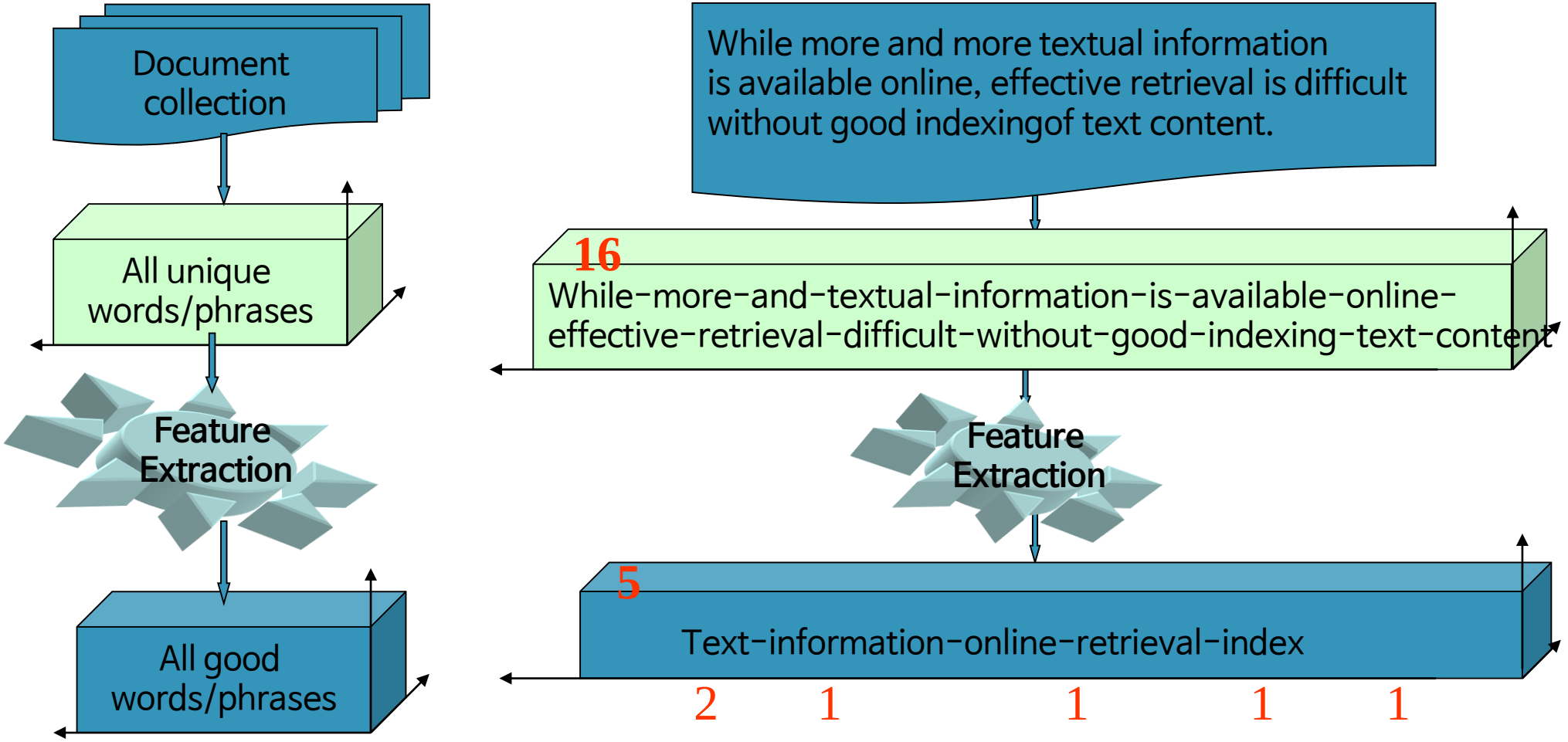
- **텍스트 표현 레벨**
  - Character (character n-grams and sequences)
  - Words (stop-words, stemming, lemmatization)
  - Phrases (word n-grams, proximity features)
  - Part-of-speech tags
  - Taxonomies / thesauri

  - Vector-space model
  - Language models
  - Full-parsing
  - Cross-modality

  - Collaborative tagging / Web2.0
  - Templates / Frames
  - Ontologies / First order theories

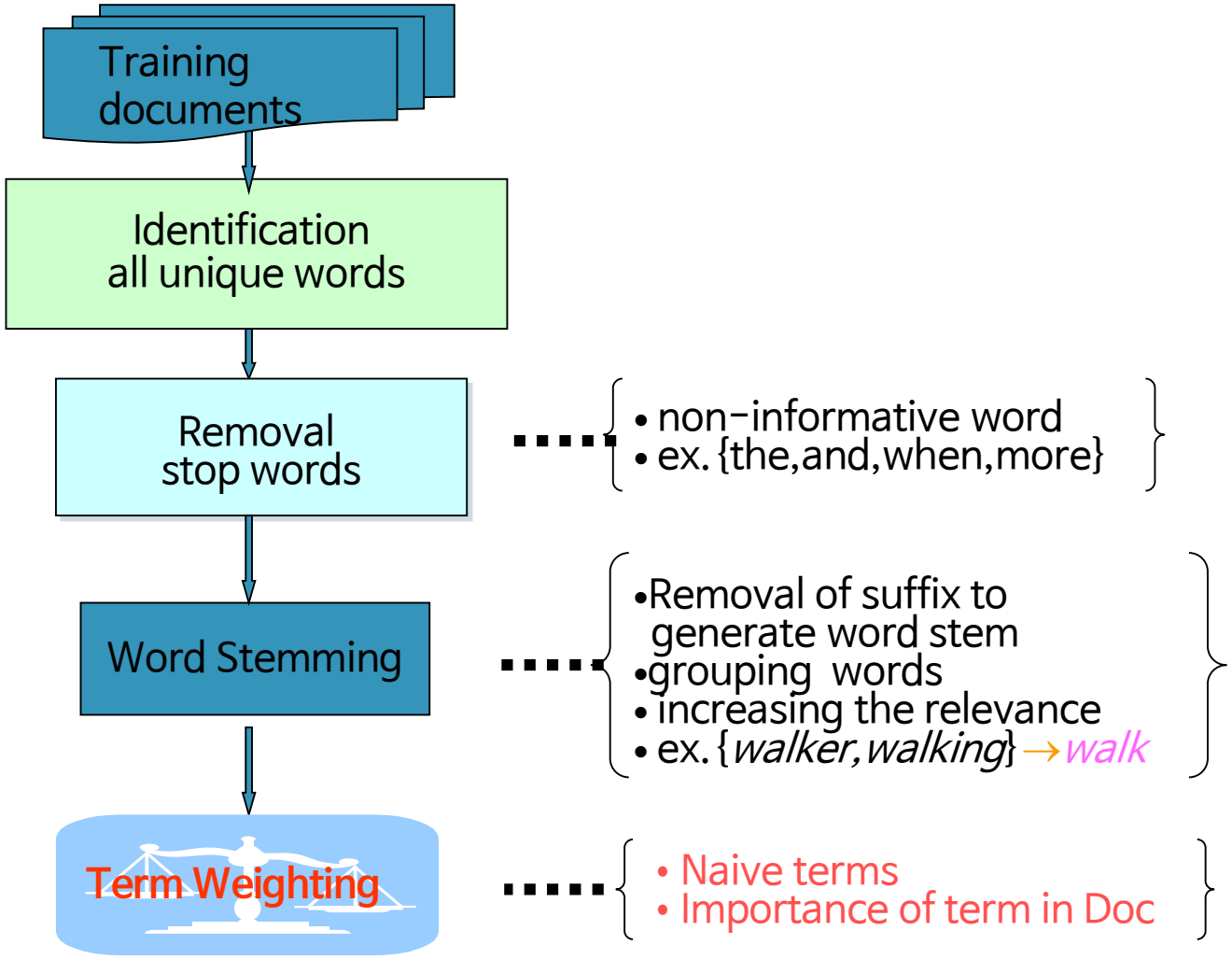
비정형 데이터 마이닝

6. 텍스트 마이닝 과정(1): 자질 추출



비정형 데이터 마이닝

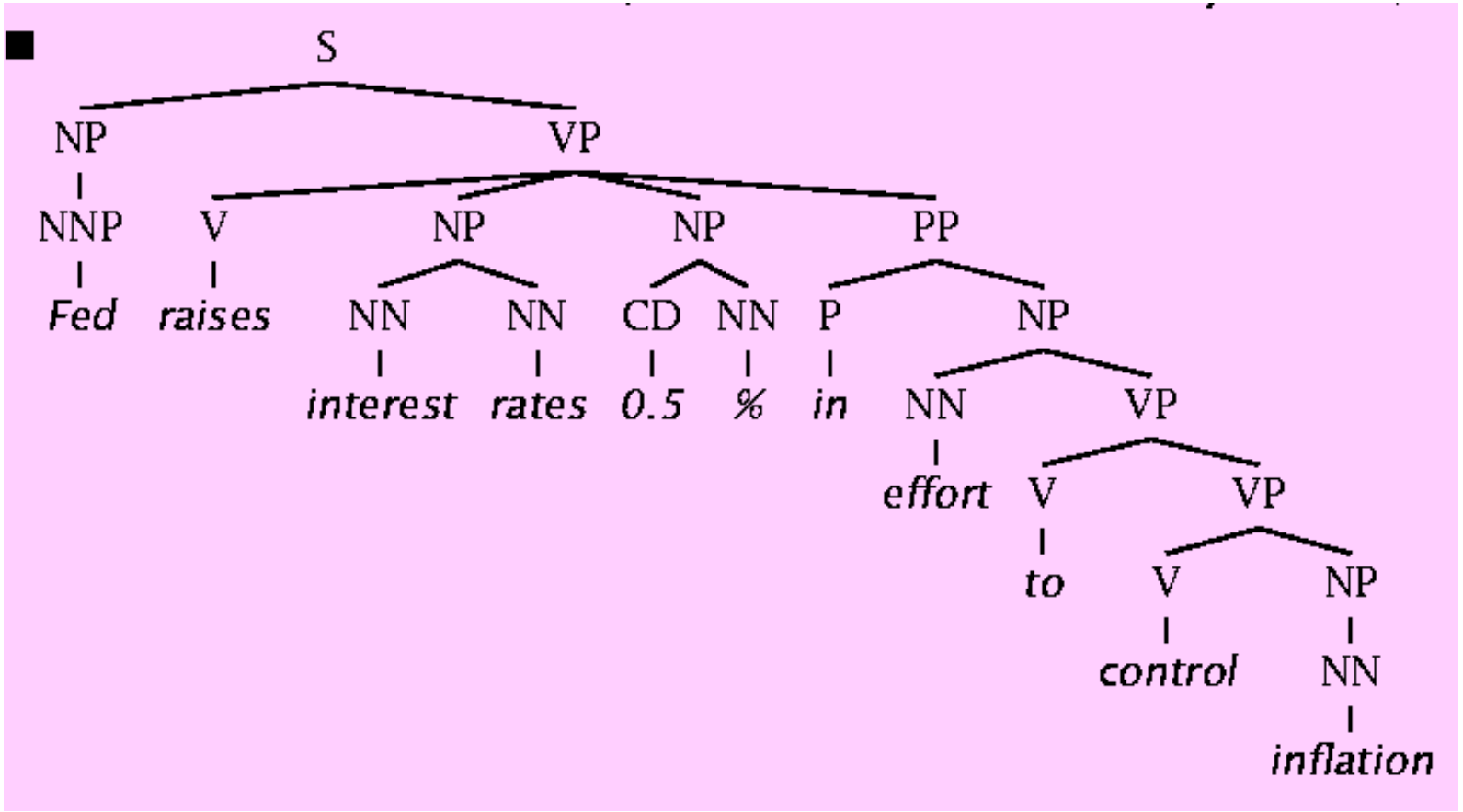
6. 텍스트 마이닝 과정(1): 자질 추출



비정형 데이터 마이닝

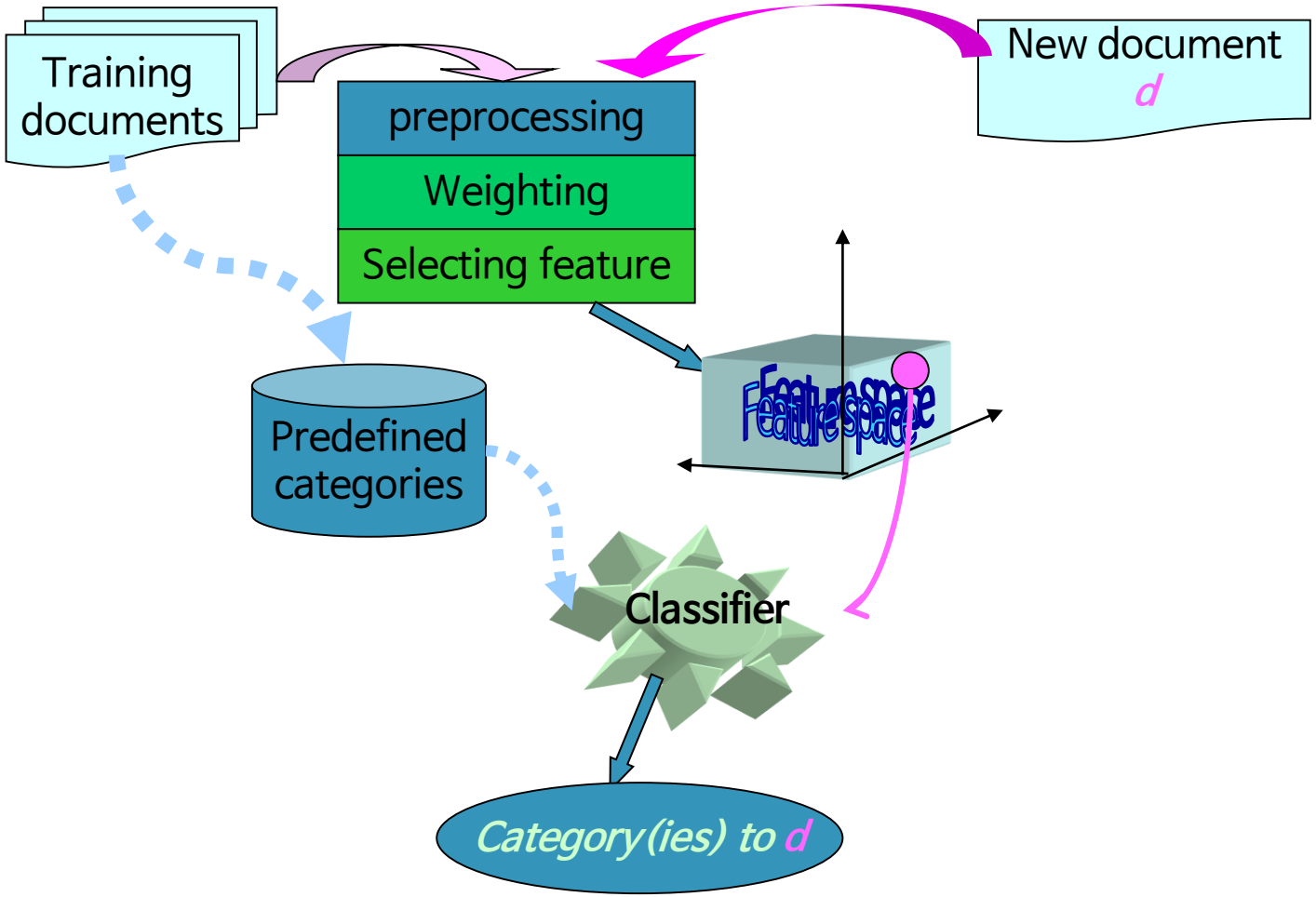
6. 텍스트 마이닝 과정(1): 자질 추출

- POS Tagging(Part Of Speech Tagging)



비정형 데이터 마이닝

6. 텍스트 마이닝 과정(2): 문서 분류



비정형 데이터 마이닝

6. 텍스트 마이닝 과정(2): 문서 분류

- **Term-Document Matrix**
  - Co-occurance Matrix
  - 문서/단어 빈도 수 행렬
  - 단어(문서) 유사도의 기준이 됨

class	아이유	소녀시대	씨스타	달샤벳
아이유	31725	217	28	9
소녀시대	169	33121	186	35
씨스타	106	411	31182	84
달샤벳	272	311	297	23323
나인뮤지스	170	369	66	102
2NE1	117	1102	264	6
걸스데이	244	276	272	193
레이디스코	61	53	87	56
레인보우	73	136	83	70
미스에이	237	278	71	17
브라운아이	257	279	78	22
시크릿	16	146	24	4
애프터스쿨	140	439	300	127
에일리	304	162	199	63
원더걸스	338	2004	73	13
주니엘	1655	127	59	54
카라	99	185	23	8
크레용팝	129	135	57	28
티아라	336	595	209	19
포미닛	162	400	477	33
피에스타	588	40	22	31

비정형 데이터 마이닝

7. 모델 평가: Precision / Recall

- **정확도(Precision)**

– 전체 분류 결과 중 옳은 결과의 비율

- **재현율 (Recall)**

– 실제 옳은 결과 중 옳은 결과로 분류된 비율

	Classified Positive	Classified Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

where

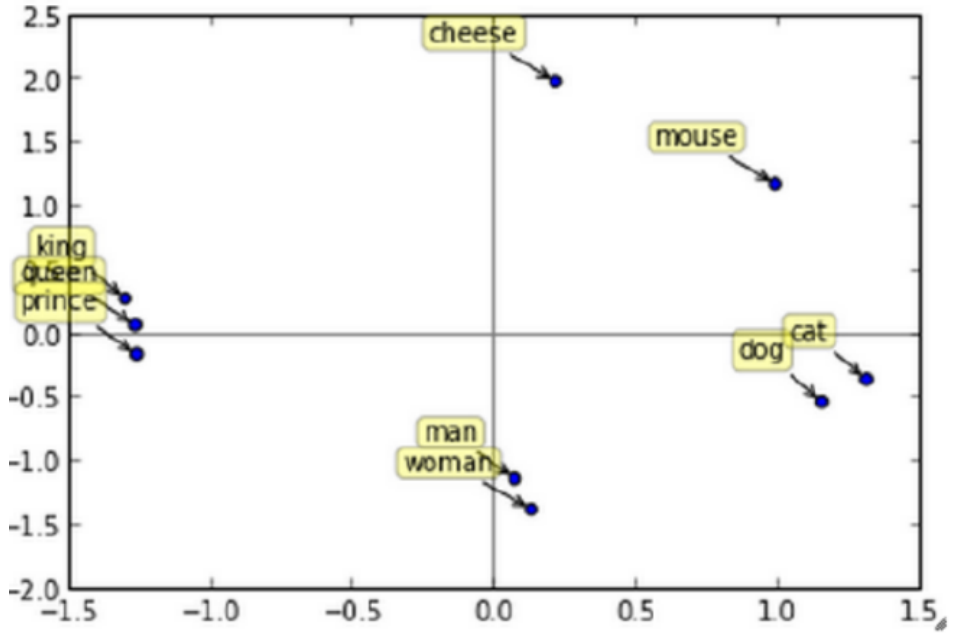
TP: the number of correct classifications of the positive examples (**true positive**),
FN: the number of incorrect classifications of positive examples (**false negative**),
FP: the number of incorrect classifications of negative examples (**false positive**), and
TN: the number of correct classifications of negative examples (**true negative**).

마무리

1. 텍스트 마이닝 최근 트렌드

- word2vec

- 단어(혹은 단어 쌍)를 벡터공간에 표현
- 유사한 의미나 관계의 단어는 유사한 벡터로 표현됨
- Input text를 받아 학습하는 지도학습(Supervised Learning) 방식
- 기존 텍스트 마이닝에 적용되던 머신 러닝 기법에 비해 성능이 좋음

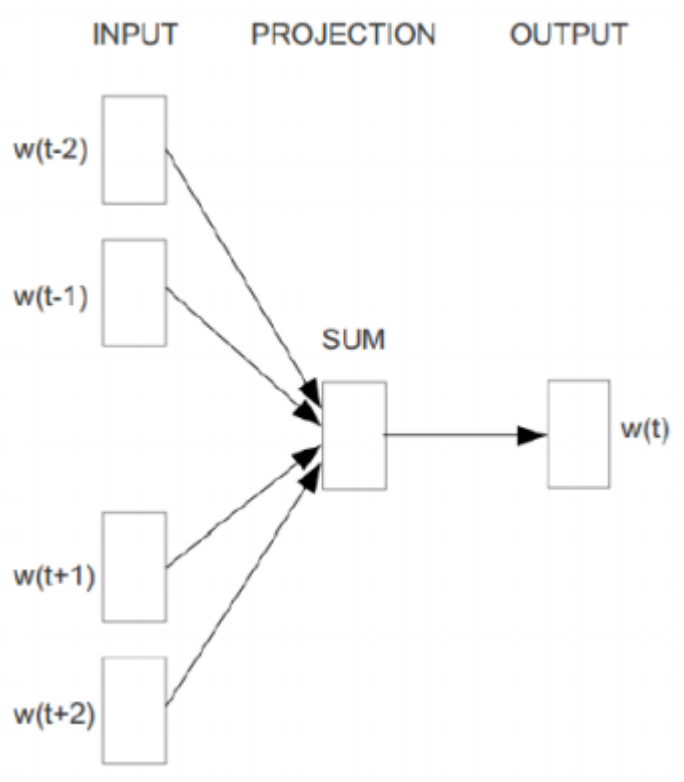


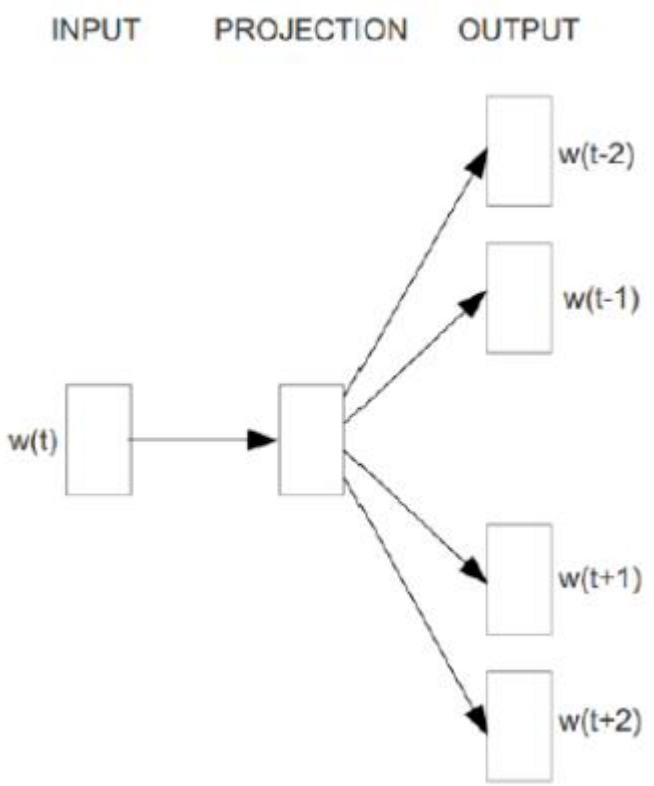
마무리

1. 텍스트 마이닝 최근 트렌드

- **word2vec**

- CBOW(Continuous Bag Of Words) vs Skip-gram





마무리

2. 텍스트 마이닝 프로젝트 예시

- 영화 '곡성' 리뷰 분석(☆정회빈☆님 제공): R로 진행
- 2013 아이돌 관계 및 감성분석: JAVA로 진행

마무리

3. 그래서 하고 싶은 말이 무엇인가요?

- **텍스트 마이닝을 위한 코딩 언어**

- 텍스트 처리 특화 언어인 '파이썬'이 R보다 절대적으로 유리(파이썬 짱!)
- 파이썬의 Numpy, Pandas 같은 정형 패키지는 이용이 어려움
- 무시무시한 데이터 양으로 병렬 처리 필수: 자바(혹은 C 언어)도 배우는 것이 좋아요

- **관심 있으신 분 저와 함께 공부해요! (프로젝트도…)**

- SNS Text Mining
- Social Network Analysis
- Sentiment Analysis
- Auto Document Classification

THANK YOU