

---



# *Python Text Mining & Topic Modeling*

## IV. 텍스트 마이닝

---

1. 비정형데이터
2. 텍스트 처리 과정
3. 텍스트마이닝
4. 한글 처리 및 실습

# 1. 비정형데이터

- 왜 Unstructured data? Text ?
  - 대략 전세계 데이터의 80%는 unstructured formats
  - 텍스트 데이터! :Web, PDF, etc...



Web contents



PDF contents

## 바이오

### 바이오, 보편적 복지와 의료비용의 절충지대

한국 바이오 투자 10년, 이익회수가 진입 기업에 관심

2000년대 초반 한국의 바이오 기업들은 기초과학 연구 중심의 벤처기업들이 대부분이었으나 정부의 산업육성 정책들이 발표되고 '벤처 대박' 신화에 투기성 자금이 몰리면서 바이오 기업들의 연구 결과들이 나오기 이전의 기대감은 결국 산업의 거품을 형성하였다. 여 년이 지난 현재 다행스러운 점은 인구증가, 고령화 이슈, 건강에 대한 관심 등의 산업 경이 긍정적으로 변화하였고, 가시적 이익을 내는 회사들이 나타나기 시작하였다는 것이다.

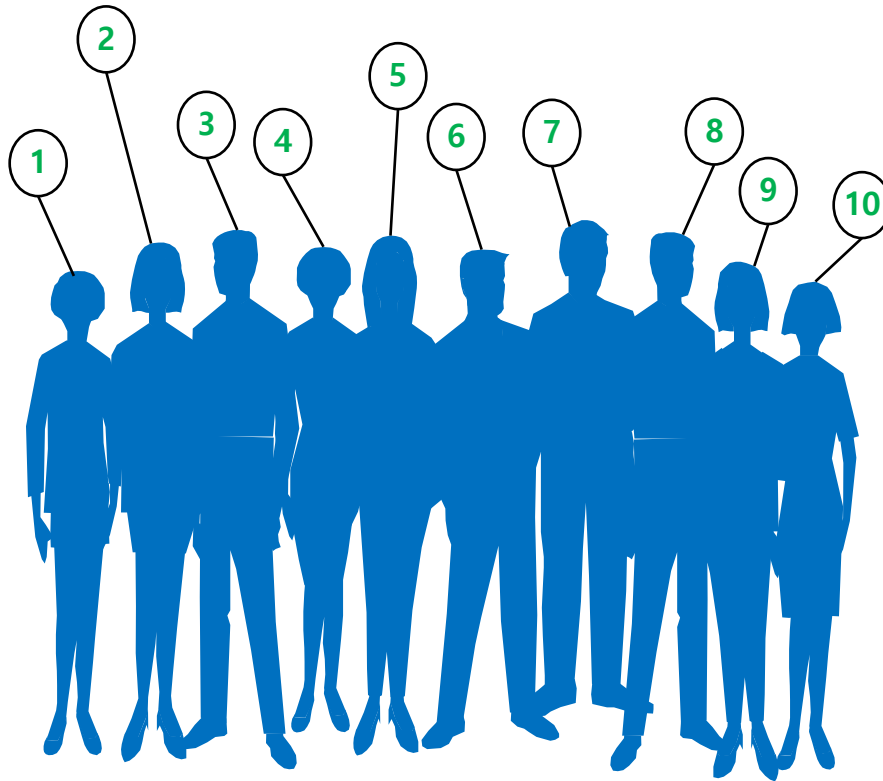
의료비 절감과 고령화 이슈를 한꺼번에 해결하는 바이오 산업

의료산업의 발전은 인간의 오랜 꿈인 생명연장을 가져왔으나 그 대가로 고령인구와 만성환의 급속 증가라는 숙제도 동시에 안겨주었다. 건강보험 재정측면에서는 의약품 소비 증가와 고가의 신규 치료법 적용으로 감당하기 어려운 수준의 의료비 상승을 가져왔고 이는 각종 재정 위협 요인이 되고 있다. 그러나 정부는 보편적 복지를 통해 보장범위를 확대하고 있다. 해결은 약가 인하와 의료수가 조정, 치료보다는 진단을 통한 예방산업 활성화 예산을 확보해야 가능해진다. 이를 가능케 하는 사업영역을 가지고 있는 회사들의 대다수가 바이오 기업들로 그 중 개인별 맞춤형 치료를 실현하는 조기진단과 예방, 재생의학이 사회적 비용 증가 고령화 시대 진입에 따른 건보재정의 부담을 감당시키는 방법이 될 것이다.

Ton Pinks: 미국리제 이지바이오

# 1. 비정형데이터

- 데이터
  - 10명의 사람들의 데이터



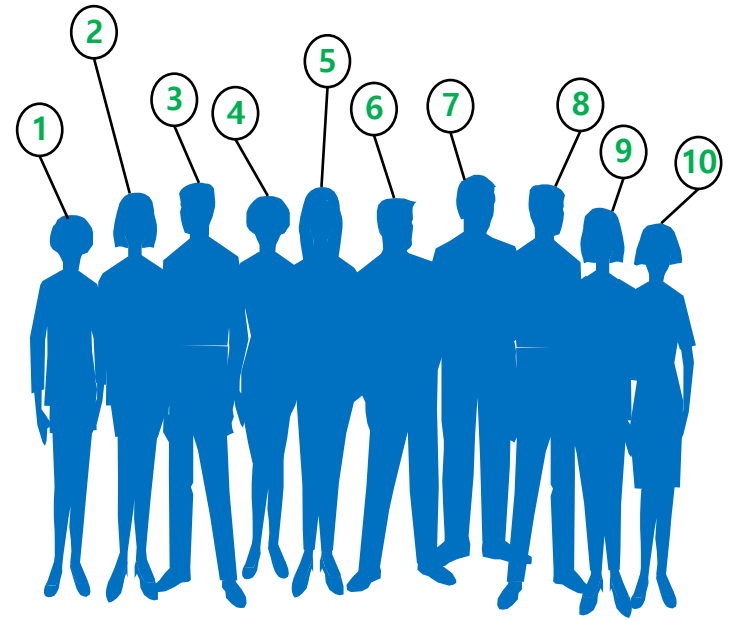
10명의 사람들을  
이해한다면?

# 1. 비정형데이터

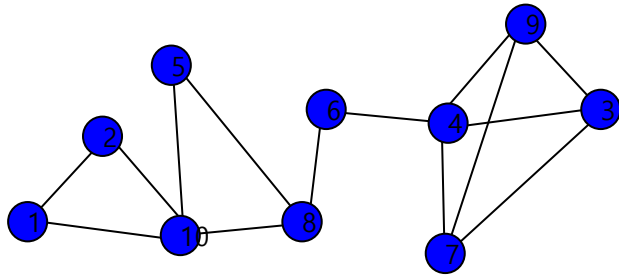
---

- 데이터의 유형

- 이름/성별/나이/거주지/직업
- 월별 요금/데이터사용량
- 휴대폰 교체주기/휴대폰 선호도
- 서로 전화하는 관계 여부
- 이용자의 VOC /서비스 선호도
- 주로 방문하는 인터넷 사이트



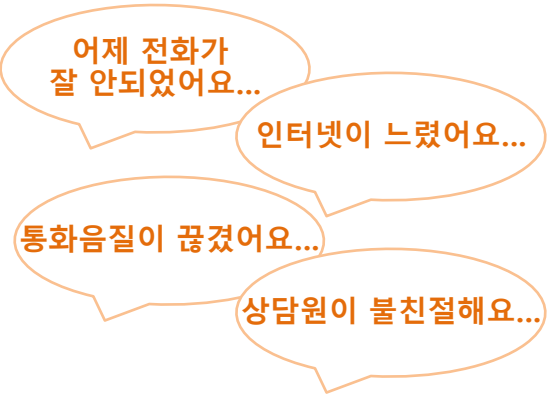
# 1. 비정형데이터



반정형데이터



비정형데이터



정형데이터

이름	성별	나이	거주지	직업	요금	데이터 사용량	휴대폰 선호도	서비스선 호도
AAA	F	20	서울	회사원	55000	3GB	LG	5
BBB	F	19	인천	자영업	45000	9GB	삼성	4
CCC	M	25	김포	회사원	35000	1GB	샤오미	3
DDD	F	42	대전	회사원	75000	4GB	LG	5
EEE	F	27	서울	자영업	65000	2GB	소니	4
FFF	M	20	서울	회사원	55000	3GB	LG	5
GGG	M	43	서울	자영업	45000	9GB	삼성	4
HHH	M	25	대전	회사원	95000	11GB	샤오미	3
III	F	42	김포	회사원	45000	3GB	LG	5
JJJ	F	27	인천	자영업	40000	4GB	소니	4

## 2. 텍스트 처리 과정

- Examples
  - 네이버 부동산 뉴스



## 2. 텍스트 처리 과정

- **Examples**

- 네이버 부동산 뉴스 수집

"450조 전세보증금, 월세화에 가계부채 '뇌관'되나"  
전세의 월세화가 진행되면서 450조원에 달하는 전세보증금이 가계부채 문제의 뇌관이 될 수 있다는 지적이 나왔다. 세입자들은 월세에 부담을 느껴 주택 매입에 나서고, 집주인들은 보증금을...

다시 늘어난 아파트 미분양...부동산 열기 냉각?  
뜨겁게 달아오르던 부동산 시장이 다시 냉각되는 조짐입니다. 아파트 미분양이 최근 다시 크게 늘어난 것인데요. 앞으로 부동산시장에 찬물을 끼얹을 것으로 우려됩니다. 김대도 기자가 취재...

노후주택 많은 지역 '신규분양' 주목  
최근 몇 년간 신규 공급이 없었던 경기 안산·의정부·포천·오산 등지에서 하반기 신규 분양이 이뤄질 예정이어서 눈길을 끌고 있다. 공급 가뭄 지역은 기존 주택의 노후화로 새 아파트로 갈아타...

"가계부채 관리-부동산 경기부양" 두마리 토끼 잡을 수 있을까?  
정부가 작년 8월부터 시행중인 주택담보대출비율(LTV)과 총부채상환비율(DTI) 규제 완화 조치를 내년까지 1년 연장기로 확정했다. 이미 행정예고 등의 절차를 거쳐 기정사실화된 사안이지만 ...



## 2. 텍스트 처리 과정

- Examples

- 네이버 부동산 뉴스

- 각 기사의 의미는 단어에 의해 전달
    - 각 기사별 단어를 파악하는 것이 중요

"450조 전세보증금, 월세화에 가계부채 '뇌관'되나"

전세의 월세화가 진행되면서 450조원에 달하는 전세보증금이 가계부채 문제의 뇌관이 될 수 있다는 지적이 나왔다. 세입자들은 월세에 부담을 느껴 주택 매입에 나서고, 집주인들은 보증금을...

다시 늘어난 아파트 미분양...부동산 열기 냉각?

뜨겁게 달아오르던 부동산 시장이 다시 냉각되는 조짐입니다. 아파트 미분양이 최근 다시 크게 늘어난 것인데요. 앞으로 부동산시장에 찬물을 끼얹을 것으로 우려됩니다. 김대도 기자가 취재...

노후주택 많은 지역 '신규분양' 주목

최근 몇 년간 신규 공급이 없었던 경기 안산·의정부·포천·오산 등지에서 하반기 신규 분양이 이뤄질 예정이어서 눈길을 끌고 있다. 공급 가뭄 지역은 기존 주택의 노후화로 새 아파트로 갈아타...

"가계부채 관리-부동산 경기부양" 두마리 토끼 잡을 수 있을까?

정부가 작년 8월부터 시행중인 주택담보대출비율(LTV)과 총부채상환비율(DTI) 규제 완화 조치를 내년까지 1년 연장키로 확정했다. 이미 행정예고 등의 절차를 거쳐 기정사실화된 사안이지만 ...

## 2. 텍스트 처리 과정

---

- **Examples**
  - 네이버 부동산 뉴스
    - 주요 단어를 추출: 예를 들어 명사만 선택

전세보증금 월세화 가계부채 뇌관 전세 월세화 진행 전세보증금 가계부채 문제 뇌관 지적 세입자 월세 부담 주택 매입 집주인 보증금

아파트 미분양 부동산 냉각 부동산 시장 냉각 조짐 아파트 미분양 부동산시장 찬물 우려

노후주택 지역 신규분양 주목 신규 공급 경기 안산 의정부 포천 오산 하반기 신규 분양 예정 눈길 공급 가뭄 지역 기존 주택 노후화 아파트

가계부채 관리 부동산 경기부양 토끼 정부 주택담보대출비율 총부채상환비율 규제 완화 조치 연장 확정 행정예고 절차 기정사실화 사안

## 2. 텍스트 처리 과정

---

- **Examples**
  - 네이버 부동산 뉴스
    - 주요 단어를 추출: 예를 들어 명사만 선택

기사1

전세보증금 월세화 가계부채 뇌관 전세 월세화 진행 전세보증금 가계부채 문제 뇌관 지적 세입자 월세 부담 주택 매입 집주인 보증금

기사2

아파트 미분양 부동산 냉각 부동산 시장 냉각 조짐 아파트 미분양 부동산시장 찬물 우려

기사3

노후주택 지역 신규분양 주목 신규 공급 경기 안산 의정부 포천 오산 하반기 신규 분양 예정 눈길 공급 가뭄 지역 기존 주택 노후화 아파트

기사4

가계부채 관리 부동산 경기부양 토끼 정부 주택담보대출비율 총부채상환비율 규제 완화 조치 연장 확정 행정예고 절차 기정사실화 사안

## 2. 텍스트 처리 과정

---

- **Examples**
  - 네이버 부동산 뉴스
    - 주요 단어를 추출: 예를 들어 명사만 선택
    - 단어를 문서별로 정리하기

	전세보증금	월세화	가계부채	미분양	경기부양	...
기사1	2	2	1	0	0	
기사2	0	0	0	1	0	
기사3	0	0	0	0	0	
기사4	0	0	1	0	1	

## 2. 텍스트 처리 과정

- **Examples**

- 네이버 부동산 뉴스

- 정리된 결과를 다음과 같이 표현...

변수처럼 사용

Observation처럼 사용

	전세보증금	월세화	가계부채	미분양	경기부양	...
기사1	2	2	1	0	0	
기사2	0	0	0	1	0	
기사3	0	0	0	0	0	
기사4	0	0	1	0	1	

### 3. 텍스트 마이닝

---

- **Text Mining**

- Text-based (digitized) documents을 대상으로 함
  - e-mails, corporate Web pages, customer surveys, résumés, medical records, DNA sequences, technical papers, incident reports, news stories
- 모든 대상 문서로 부터, 지식이나 요약 정보를 얻고자 함

- **Text Mining Application**

- ① 정보 추출/요약 및 시각화
- ② 문서의 군집화 및 주제 발견
- ③ 문서 분류
- ④ 추천
- ⑤ 정형데이터와 같이 사용

### 3. 텍스트 마이닝

- **Term / document matrix**

- Most common form of representation in text mining
- Can be large: 크기가 매우 큰 행렬 -> Sparse matrix
- Can be binary, or use counts: 행렬의 값은 문서 내 단어의 빈도 또는 1/0으로 표시

Example: 10 documents: 6 terms

	Regression	Classification	Clustering	Exploration	Process	Open source
D1	24	21	9	0	0	3
D2	32	10	5	0	3	0
D3	12	16	5	0	0	0
D4	6	7	2	0	0	0
D5	43	31	20	0	3	0
D6	2	0	0	18	7	6
D7	0	0	1	32	12	0
D8	3	0	0	22	4	4
D9	1	0	0	34	27	25
D10	6	0	0	17	4	23

$$D_1 = (d_{i1}, d_{i2}, \dots, d_{it})$$

- 각 문서는 단어들로 이뤄진 Vector->Vector Space Model

### 3. 텍스트 마이닝

---

- **Stop words and Stemming**
  - Text mining을 통해 발견된 단어의 리스트를 모두 사용?
  - **Stop Words**
    - 텍스트마이닝 및 정보검색에서는 큰 의미가 없는 가장 빈번하게 사용되는 단어
      - 예: the, of, and, to, ....
      - 대략 400-500개 단어임(영어)
    - Stop words 처리를 통해
      - 데이터 사이즈 축소: 일반적으로 전체 단어의 20-30%정도를 차지
      - 효율성 제고!: 정보검색이나 텍스트 마이닝에 큰 의미가 없는 단어들 제거
  - **Stemming**
    - 단어의 어근(root/stem)을 찾는 기법
      - 예
        - user / users /used /using -> use
        - Engineering/engineered -> engineer
    - 효용성
      - 정보검색 및 텍스트 마이닝 성능 향상
      - 뜻은 같지만 형태가 다른 단어들의 매치
      - 데이터 사이즈 축소
        - » 일반적으로 40-50%정도의 크기를 줄이는 것으로 알려짐(영어)



### 3. 텍스트 마이닝

---


- **Weighting in Term Document Matrix**

- 모든 단어가 똑같지 않음!
  - 예: "햄버거"는 "와퍼"보다 덜 중요
  - 많은 문서에서 많이 출현하는 단어는 그만큼 특정 문서에 대한 discriminatory power가 적음
- 가중치 사용을 통해 보완: inverse-document frequency

$$\text{IDF} = \log(N/n_j)$$

- Term importance = Term Frequency (TF) x IDF
  - $n_j$  = # of docs containing the term
  - $N$  = total # of docs
  - 해석: TF가 높으면서 IDF가 높은 단어가 중요
  - **TF x IDF**: 단어 중요도에 대한 일반적 measure

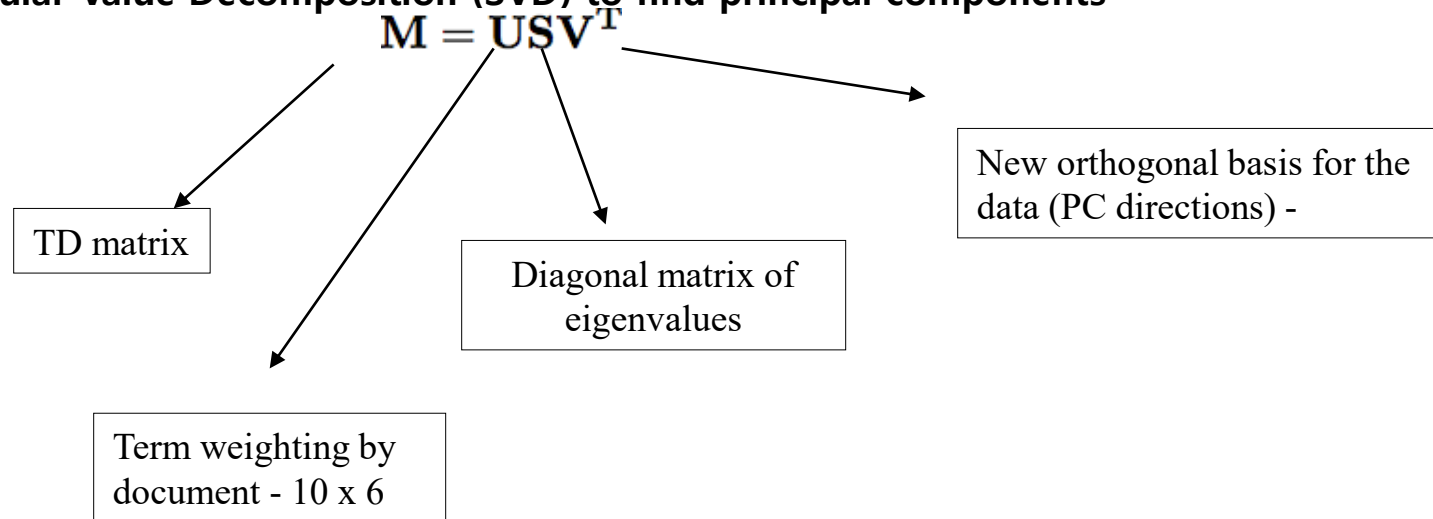
### 3. 텍스트 마이닝

	Regression	Classification	Clustering	Exploration	Process	Opensource							
D1	24	21	9	0	0	3							
D2	32	10	5	0	3	0							
D3	12	16	5	0	0	0							
D4	6	7	2	0	0	0							
D5	43	31	20	0	3	0							
D6	2	0	0	18	7	6							
D7	0	0	1	32	12	0		Regression	Classification	Clustering	Exploration	Process	Opensource
D8	3	0	0	22			D1	2.53	14.6	4.6	0	0	2.1
D9	1	0	0	34			D2	3.3	6.7	2.6	0	1.0	0
D10	6	0	0	17			D3	1.3	11.1	2.6	0	0	0
							D4	0.7	4.9	1.0	0	0	0
							D5	4.5	21.5	10.2	0	1.0	0
							D6	0.2	0	0	12.5	2.5	11.1
							D7	0	0	0.5	22.2	4.3	0
							D8	0.3	0	0	15.2	1.4	1.4
							D9	0.1	0	0	23.56	9.6	17.3
							D10	0.6	0	0	11.8	1.4	16.0

### 3. 텍스트 마이닝

---

- 이슈: 이음 동의어? 유사한 의미의 단어들?
  - Data mining and knowledge discovery
  - Car and automobile
  - Beet and beetroot
- **Latent Semantic Indexing**
  - *의미적으로*, 위의 단어들은 동일하며, 이런 단어들을 같이 갖는 문서들도 관련성이 높다고 고려되어야 함
  - 이음동의어 리스트를 통해 처리하는 방법이 있지만 계산이 많고 어려움
  - 그러므로 LSI를 문서들에 적용하여 hidden semantic structure를 찾는 것이 필요
- **Singular Value Decomposition (SVD) to find principal components**



### 3. 텍스트 마이닝

---

- **Text mining Tool**
  - SAS Text-miner VS R or Python

SAS	Python
<ul style="list-style-type: none"><li>• Commercial</li><li>• 영어/한글 지원</li><li>• 사전 우수(ontology도 지원)</li><li>• 보통의 확장성</li><li>• 유지보수!</li></ul>	<ul style="list-style-type: none"><li>• Non-commercial</li><li>• 영어/한글 지원</li><li>• Publicly available dictionary(사전 교체 가능!)</li><li>• 우수한 확장성</li><li>• 유지보수 없음</li><li>• 모듈: nltk, KoNLPy 등</li></ul>

## 4. 한글 처리 실습

- **KoNLPy**
  - 코엔엘피와이
  - 한글 처리를 위한 모듈



KoNLPy

```
>>> from konlpy.tag import Kkma
>>> from konlpy.utils import pprint
>>> kkma = Kkma()
>>> pprint(kkma.sentences(u'네, 안녕하세요. 반갑습니다.'))
[네, 안녕하세요.,
반갑습니다.]
>>> pprint(kkma.nouns(u'질문이나 건의사항은 깃헙 이슈 트래커에 남겨주세요.'))
[질문,
건의,
건의사항,
사항,
깃헙,
이슈,
트래커]
>>> pprint(kkma.pos(u'오류 보고는 실행 환경, 에러메세지와함께 설명을 최대한상세히!^^'))
[(오류, NNG),
(보고, NNG),
(는, JX),
(실행, NNG),
(환경, NNG),
(, SP),
(에러, NNG),
(메세지, NNG),
...]
```

- **설치 과정**
  - Java 1.7 이상
  - JAVA\_HOME 설정
  - JPyep1 (0.5.7 이상)을 설치
  - Pip를 이용한 설치

## 4. 한글 처리 실습

---

- **KoNLPy 설치**
  - pip install --upgrade pip
  - pip install JPytype1-0.5.7-cp27-none-win\_amd64.whl
  - pip install konlpy
- **코퍼스와 사전**
  - 코퍼스:
    - 연세 말뭉치
    - 고려대 한국어 말뭉치
    - HANBTEC 2.0
    - HKIB-40075
    - KAIST
    - Sejong
  - 사전:
    - Hannanum
    - Kkma
    - Mecab

## 4. 한글 처리 실습

Twitter Korean Text (ntags=19)	Komoran (ntags=42)	Mecab-ko (ntags=43)	Kkma(ntags=56)	Hannanum (ntags=22)
명사 (Nouns, Pronouns, Company Names, Proper Noun, Person Names, Numerals, Standalone, Depend ent)	일반 명사	일반 명사	보통명사	보통명사
동사	고유 명사	고유 명사	고유명사	고유명사
형용사	의존 명사	의존 명사	일반 의존 명사	의존명사
관형사 (ex: 새, 헌, 참, 첫, 이, 그, 저)	수사	단위를 나타내는 명사	단위 의존 명사	수사
부사 (ex: 잘, 매우, 빨리, 반드시, 과연)	대명사	수사	수사	대명사
접속사	동사	대명사	대명사	동사
감탄사 (ex: 헐, 어머니, 얼씨구)	형용사	동사	동사	형용사
조사 (ex: 의, 에, 에서)	보조 용언	형용사	형용사	보조 용언
선어말어미 (ex: 었)	긍정 지정사	보조 용언	보조 동사	관형사
어미 (ex: 다, 요, 여, 하댕ㅋㅋ)	부정 지정사	긍정 지정사	보조 형용사	부사
접미사	관형사	부정 지정사	긍정 지정사, 서술격 조사 '이다'	감탄사
구두점	일반 부사	관형사	부정 지정사, 형용사 '아 니다'	격조사
외국어, 한자 및 기타기호	접속 부사	일반 부사	수 관형사	보조사
알파벳	감탄사	접속 부사	일반 관형사	서술격 조사
숫자	주격 조사	감탄사	일반 부사	선어말어미
미등록어 (ex: ㅋㅋ)	보격 조사	주격 조사	접속 부사	종결 어미
트위터 해쉬태그 (ex: #히히)	관형격 조사	보격 조사	감탄사	연결 어미
트위터 아이디 (ex: @echojuliett)	목적격 조사	관형격 조사	주격 조사	전성 어미
이메일 주소	부사격 조사	목적격 조사	보격 조사	접두사
웹주소	호격 조사	부사격 조사	관형격 조사	접미사
	인용격 조사	호격 조사	목적격 조사	기호
	접속 조사	인용격 조사	부사격 조사	외국어
	보조사	접속 조사	호격 조사	
	선어말어미	보조사	인용격 조사	
	종결 어미	선어말어미	접속 조사	
	연결 어미	종결 어미	보조사	
	명사형 전성 어미	연결 어미	존칭 선어말 어미	
	관형형 전성 어미	명사형 전성 어미	시제 선어말 어미	
	체언 접두사	관형형 전성 어미	공손 선어말 어미	
	명사파생 접미사	체언 접두사	평서형 종결 어미	

## 4. 한글 처리 실습

Twitter Korean Text (ntags=19)	Komoran (ntags=42)	Mecab-ko (ntags=43)	Kkma(ntags=56)	Hannanum (ntags=22)
	형용사 파생 접미사	동사 파생 접미사	명령형 종결 어미	
	어근	형용사 파생 접미사	청유형 종결 어미	
	마침표, 물음표, 느낌표	어근	감탄형 종결 어미	
	줄임표	마침표, 물음표, 느낌표	존칭형 종결 어미	
	따옴표,괄호표,줄표	줄임표 ...	대등 연결 어미	
	쉼표,가운뎃점,콜론,빗금	여는 괄호 (, [	보조적 연결 어미	
	붙임표(물결,숨김,빠짐)	닫는 괄호 ), ]	의존적 연결 어미	
	기타기호 (논리수학기호, 화폐기호)	구분자 , · / :	명사형 전성 어미	
	한자	기타 기호	관형형 전성 어미	
	외국어	한자	체언 접두사	
	숫자	외국어	용언 접두사	
	명사추정범주	숫자	명사파생 접미사	
	용언추정범주		동사 파생 접미사	
	분석불능범주		형용사 파생 접미사	
			어근	
			마침표, 물음표, 느낌표	
			줄임표	
			따옴표,괄호표,줄표	
			쉼표,가운뎃점,콜론,빗금	
			붙임표(물결,숨김,빠짐)	
			기타기호 (논리수학기호, 화폐기호)	
			한자	
			외국어	
			숫자	
			명사추정범주	



## V. 머신러닝 응용

---

1. 데이터 분석과 머신러닝
2. 온라인비정형데이터의 수집
3. 토픽모델링
4. 인공신경망과 네트워크
5. Word2vec을 통한 유사도 계산

# 1.데이터 분석과 머신러닝

## ➤ Deep Learning?

- 데이터로부터 귀납적으로 추론을 하는 머신러닝 기법 중
- 다층 인공신경망에 적용되는 기법 의미
- 다양한 오픈소스 가용: nnet, neuralnet, deepnet, h2o, mxnet, ....

### Data Analytics

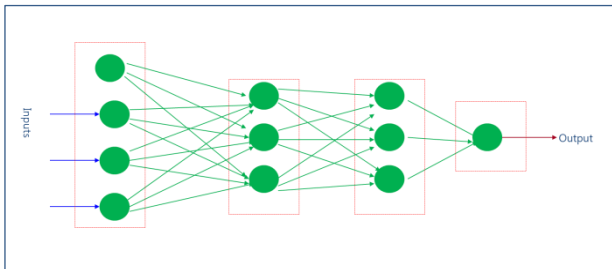
#### Machine Learning

SVM

KNN

Tree

#### ANN > Deep Learning



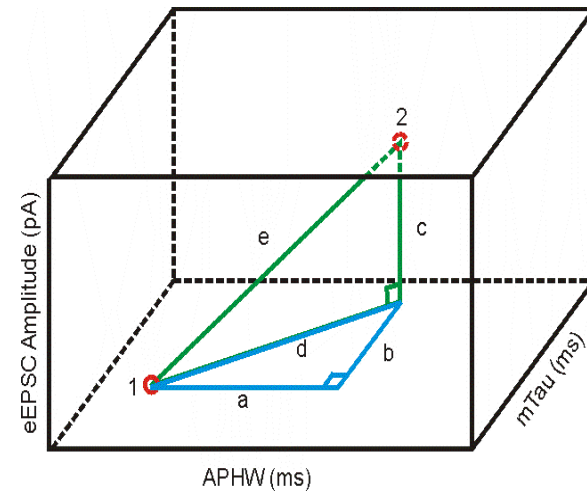
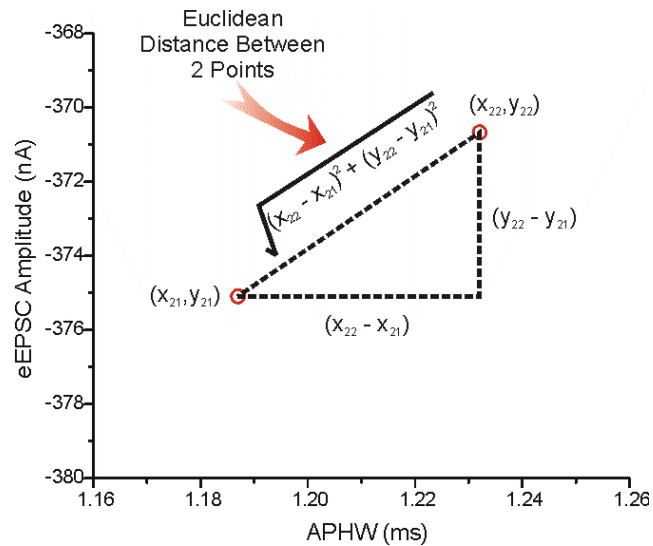
Graph

AR

# 1.데이터 분석과 머신러닝

## ➤ Clustering?

- Cluster의 개수나 구조에 관한 특별한 사전 가정없이, 개체들 사이의 유사성/거리에 근거해 cluster를 찾고 다음 단계의 분석을 하게 하는 기법
- 유사한 개체들을 cluster로 그룹화하여 각 집단의 성격을 파악

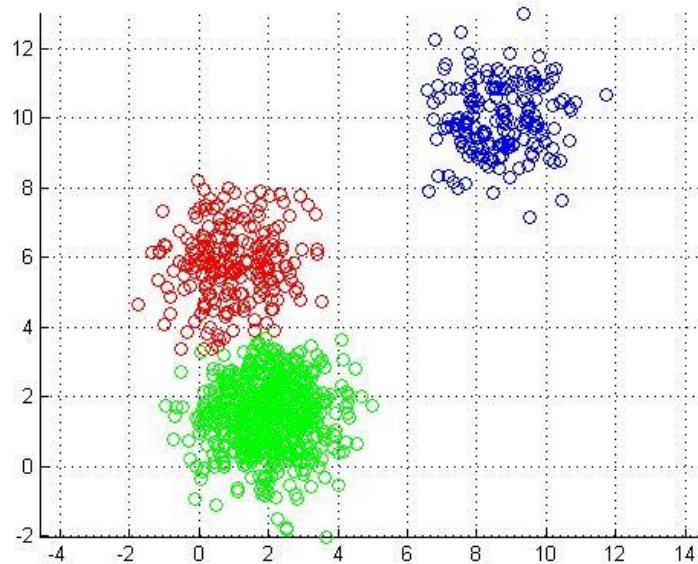


# 1.데이터 분석과 머신러닝

## ➤ Clustering

### – K-means clustering

- K개: 주어진 cluster의 개수
- 절차: k개 만큼의 cluster seed 선택(지정/임의로 선택 등)
- 각 개체들에 대해 cluster seed와 거리 계산하여, 가장 가까운 seed에 개체 할당
- 각 개체가 seed에 할당될 때마다, 군집의 중심이 그 군집에 속하는 개체들의 평균 벡터로 다시 계산
- 개체들을 가장 가까운 Cluster seed에 재할당하고, 이 과정을 cluster center의 변화가 일정 수준 이하가 되도록 반복
- 초기 군집수 결정이 중요: EDA를 통해 파악(시각화), 주성분 분석을 통해 2,3차원 그래프로 분석



# 1.데이터 분석과 머신러닝

---

## ➤ Decision Tree

- An empirical tree represents a segmentation of the data that is created by applying a series of simple rules
- Target 변수가 범주형이 아닌 수치형 데이터에도 사용 가능: Regression Tree
- Type of decision tree
  - C4.5 / C5.0 : information theory, entropy, Quinlan (1983)
  - CART(Classification and regression Tree): Gini index, Breiman et al. (1984)
  - CHAID(Chi-squared Automatic Interaction Detector): Chi-square test 이용, Kass(1980)
- 장점:
  - 해석의 용이성 / 상호작용 효과의 해석: / 비모수적 모형(선형성, 정규성, 등분산성의 가정 불필요)
- 단점:
  - 비연속성 / 선형성 또는 주효과 결여 / 불안정성(분석용 자료에만 의존하므로, 새로운 자료의 예측에서는 불안정할 수 있음. Test data에 의한 교차타당성 평가 등이 필요)

## 2. 온라인비정형데이터의 수집

---

- **Understanding Web data**

- Location

- URL 파악 및 출현 횟수 (한 페이지 내 표? 또는 여러 페이지의 여러 표?)
    - 여러 웹사이트의 다양한 정보인지 파악

- Accessibility

- **Free direct immediate access 확인**
    - **Web Form 또는 Web API 필요한지 확인**
    - 인증이 필요한지 확인
    - 특정 protocol을 사용하는지 확인

- Format

- **plain text 또는 tabular (spreadsheet-like) form?**
    - **HTML? XML? JSON format?**
    - Other formats: binary, images, maps, etc?

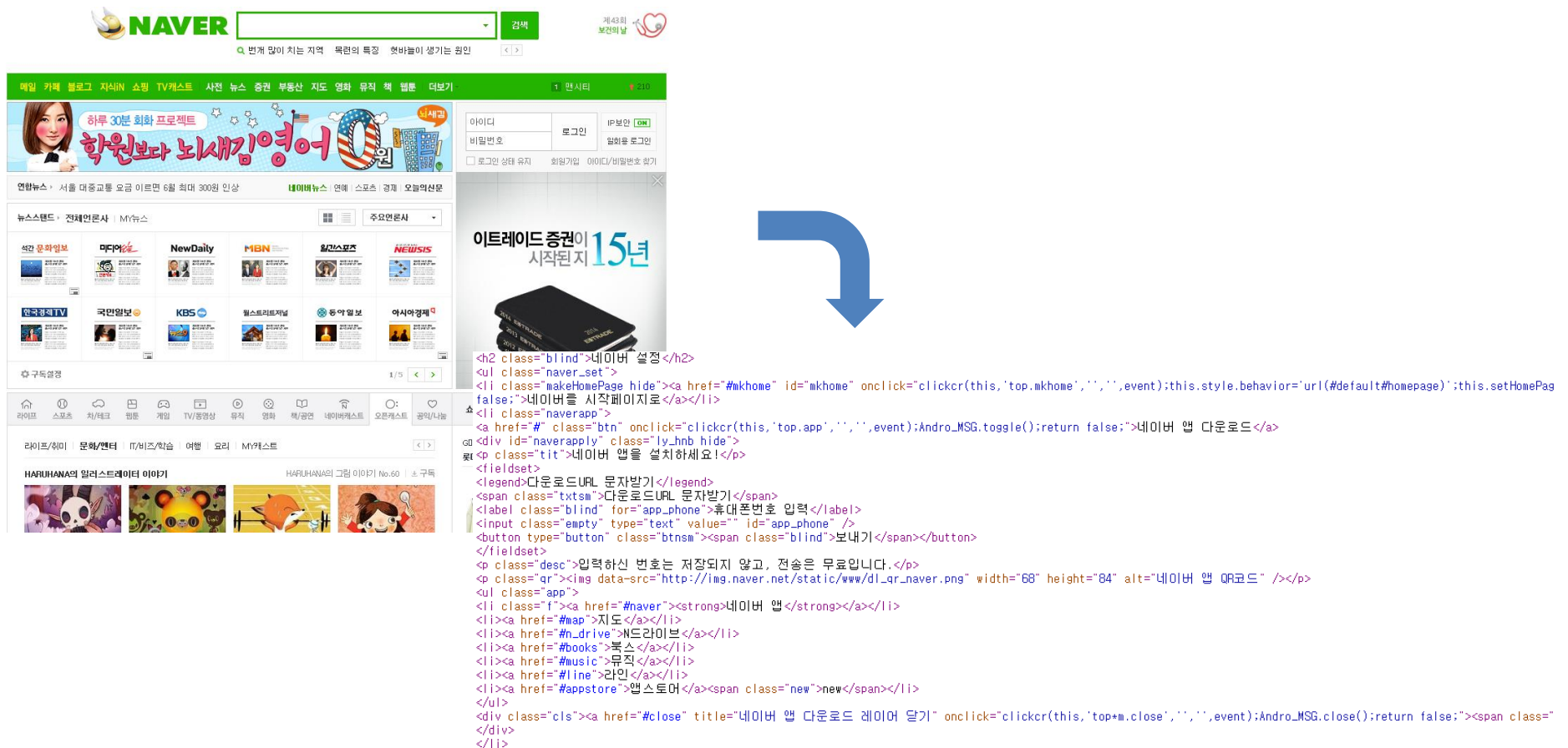
***WWW** World Wide Web*  
***W3C** World Wide Web Consortium*  
***URL** Uniform Resource Locator*  
***HTTP** HyperText Transfer Protocol*  
***XML** Extensible Markup Language*  
***HTML** HyperText Markup Language*  
***JSON** JavaScript Object Notation*

## 2. 온라인비정형데이터의 수집

- Understanding Web data

- 왜 XML / HTML?

- 인터넷에서는 많은 양의 정보와 자료가 HTML 및 XML 형태로 표현/저장/배포되고 있음
- 범용적으로 사용되며, 웹의 데이터를 처리한다는 것은 HTML 등을 처리하는 것을 의미



The image shows a screenshot of the Naver homepage. A large blue arrow points from the main content area to a snippet of HTML code. The code is a JavaScript function for the Naver app download, showing various HTML elements and JavaScript logic.

```
<h2 class="blind">네이버 설정</h2>
<ul class="naver_set">
<li class="makeHomePage hide"><a href="#mkhome" id="mkhome" onclick="clickcr(this,'top.mkhome','',event);this.style.behavior=url(#default#homepage);this.setHomePag
false;">네이버를 시작페이지로</a></li>
<li class="naverapp">
<a href="#" class="btn" onclick="clickcr(this,'top.app','',event);Andro_MSG.toggle();return false;">네이버 앱 다운로드</a>
<div id="naverapply" class="ly_hnb hide">
<p class="tit">네이버 앱을 설치하세요!</p>
<fieldset>
<legend>다운로드 URL 문자발기</legend>
<span class="txtsm">다운로드URL 문자발기</span>
<label class="blind" for="app_phone">휴대폰번호 입력</label>
<input class="empty" type="text" value="" id="app_phone" />
<button type="button" class="btnsm"><span class="blind">보내기</span></button>
</fieldset>
<p class="desc">입력하신 번호는 저장되지 않고, 전송은 무료입니다.</p>
<p class="qr"></p>
<ul class="app">
<li class="f"><a href="#naver"><strong>네이버 앱</strong></a></li>
<li><a href="#map">지도</a></li>
<li><a href="#n_drive">N드라이브</a></li>
<li><a href="#books">북스</a></li>
<li><a href="#music">뮤직</a></li>
<li><a href="#line">라인</a></li>
<li><a href="#appstore">앱스토어</a><span class="new">new</span></li>
</ul>
<div class="cls"><a href="#close" title="네이버 앱 다운로드 레이더 닫기" onclick="clickcr(this,'top#close','',event);Andro_MSG.close();return false;"><span class="
</div>
</li>
```

## 2. 온라인비정형데이터의 수집

---

- **Understanding Web data**

- HTML: Hypertext Markup Language
  - 웹 문서를 만들기 위하여 사용하는 기본적인 프로그래밍 언어의 한 종류, markup language
- XML: eXtensible Markup Language
  - human-readable 그리고 machine-readable한 형태의 문서 구성 규칙을 정의하는 markup language
- XML Tree Structure
  - 각 노드는 Name/attributes/optional content/nested elements 등이 포함될 수 있음
- 예:
  - `<movie mins="126" lang="en">`
    - » `<title>Good Will Hunting</title>`
    - » `<director>Gus Van Sant</director>`
    - » `<year>1998</year>`
    - » `<genre>drama</genre>`
  - `</movie>`
- 참고
  - `<![CDATA[ ]]>` CDATA Character Data
    - » CDATA내의 정보는 Parsing 되지 않음



## 2. 온라인비정형데이터의 수집

---

- **Understanding Web data**

- JSON

- JSON stands for JavaScript Object Notation and it is a format for representing data
    - JSON can be used directly in JavaScript code for Web pages

- JSON Data Types: null / true / false / number / string

- JSON Data Containers: square brackets [ ] / curly brackets { }

- []: ordered unnamed arrays, [ 1, 2, 3, ... ]
    - {}: named arrays, { "dollars" : 5, "euros" : 20, ... }

```
[  
  { "name": "X",  
    "grams": 300,  
    "qty": 4,  
    "new": true },  
  { "name": "Y",  
    "grams": 200,  
    "qty": 5,  
    "new": false },  
  { "name": "Z",  
    "grams": 500,  
    "qty": null,  
    "new": true }  
]
```

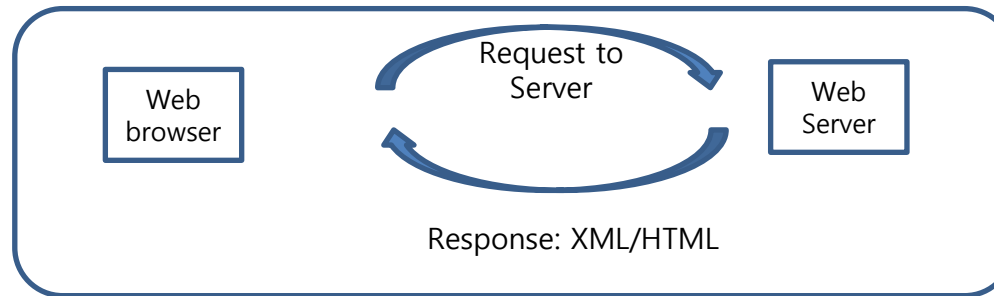
**JSON sample**

## 2. 온라인비정형데이터의 수집

---

- **Understanding Web data**
  - HTTP

Protocol of communications:  
HTTP(HyperText Transfer Protocol)



- **HTTP Method**
  - Standardized method for transferring data or documents over the Web
  - GET retrieves whatever information is identified by the Request-URI
  - POST request with data enclosed in the request body

### 3. 토픽모델링

---

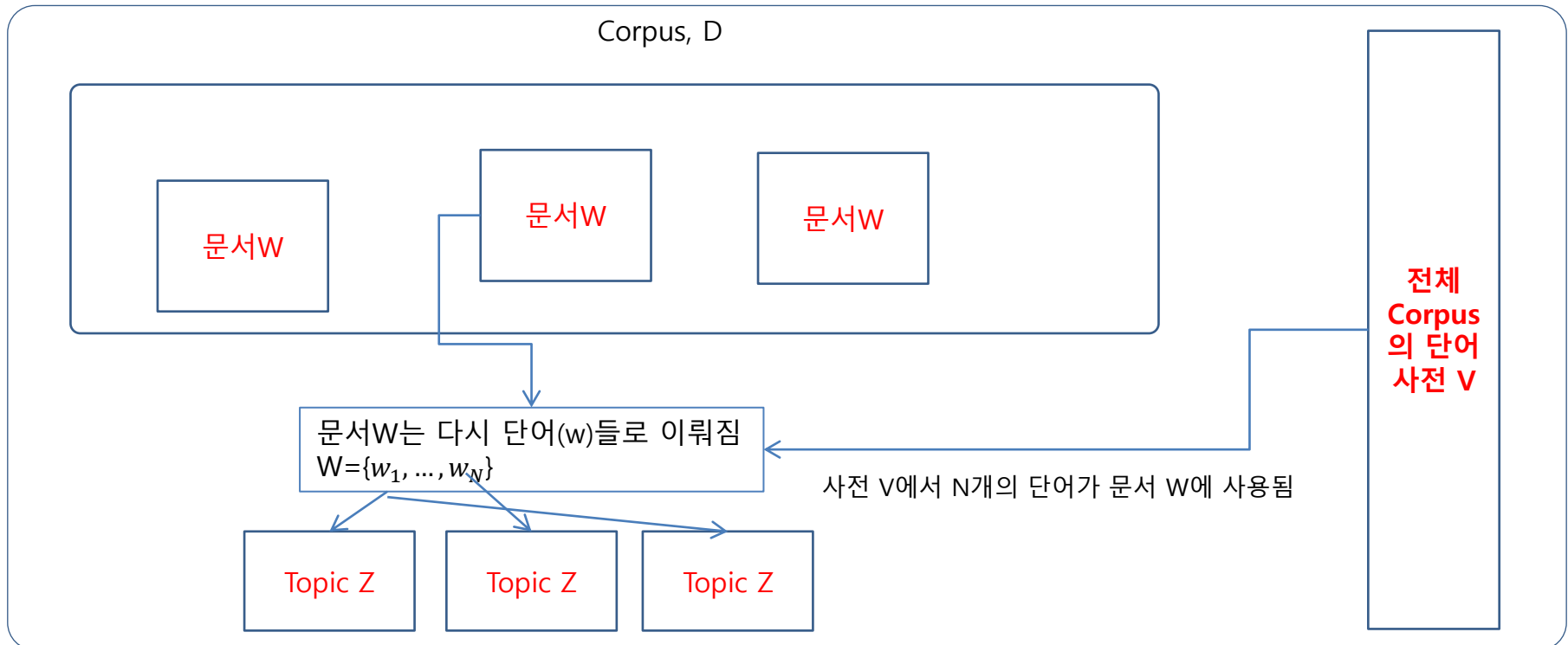
- **Topic modeling**

- Topic Model
  - Generative probabilistic models for the term frequency of documents in a given corpus
  - 여러 방법들(LSI, pLSI, LDA, etc)이 있으며, LDA도 그 중의 하나
- LDA
  - Latent Dirichlet Allocation
  - 배경
    - 예를 들어, 문서1과 문서2가 주제는 유사해도 각 문서에 등장하는 단어의 종류나 빈도는 다를 수 있는데, 단순한 키워드 기반의 모델로는 유사도를 계산하거나 주제 분류를 하는 데에는 한계가 발생
    - 많은 텍스트에 기초에  $\alpha$ 와  $\beta$ 를 찾고, 개별 문서의  $\theta$ 를 계산할 수 있으면, 이  $\theta$ 를 가지고 유사도 계산이나 분류 작업을 할 수 있음
  - 특징
    - Bayesian mixture model for discrete data (Topics are assumed to be uncorrelated)
    - Mixture Membership Model (문서는 하나의 토픽에만 속하는 것이 아니라 여러 다른 토픽들의 혼합(Mixture)으로 정의할 수 있음)

### 3. 토픽모델링

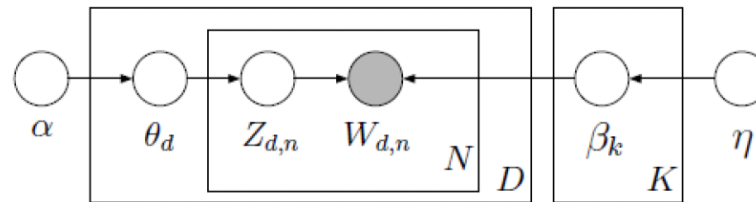
- LDA(Latent Dirichlet Allocation)

- Topics의 수  $K$ 는 사전에 결정되어야 함
- 문서 집합(Corpus)에 대한 Generative Probabilistic Model이며, 1) 문서의 주제 분포와 2) 주제별 단어의 분포를 알고 있다면, 특정 문서가 만들어질 확률을 알 수 있음



### 3. 토픽모델링

- **LDA Step**
  - Step 1
    - 각 주제( $k=1,\dots,K$ )에 대해서 Draw Vector of Term Proportion from  $\beta_k \sim \text{Dirichlet}_V(\eta)$
  - Step 2
    - 각 문서( $d=1,\dots,D$ )에 대해서, Draw Topic proportions  $\theta$  from  $\theta_d \sim \text{Dirichlet}_K(\alpha)$
  - Step 3
    - For each of the  $N$  words ( $n=1,\dots,N$ )  $w_{d,n}$ 
      - Choose a topic  $z_{d,n} \sim \text{Multinomial}_K(1, \theta_d)$
      - Choose a word  $w_{d,n}$  from a multinomial probability distribution conditioned on the topic  $z_{d,n}$ ,  $\text{Multinomial}_V(1, \beta_{z_{d,n}})$



### 3. 토픽모델링

---

- More on LDA Steps...

- 위의 과정을 다시 표현하면 아래와 같음

- $w_{d,n} | z_{d,n}, \theta_d, \beta \sim \text{Multinomial}_V(1, \beta_{z_{d,n}})$
    - $z_{d,n} | \theta_d \sim \text{Multinomial}_K(1, \theta_d)$
    - $\theta_d \sim \text{Dirichlet}_K(\alpha)$
    - $(\beta_1, \dots, \beta_K) \sim \prod_{k=1}^K \text{Dirichlet}_V(\eta)$

- $\alpha$ 와  $\eta$ 가 주어질때,  $(\theta, z, \beta, w)$ 의 Joint Distribution은  $p(\theta, z, \beta, w; \alpha, \eta)$ 이며 아래와 같이 표시됨

$$p(\theta, z, \beta, w; \alpha, \eta) = [\prod_{k=1}^K p(\beta_k | \eta)] \prod_{d=1}^D [p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | z_{d,n}, \theta_d, \beta)]$$

- $\theta, z, \beta$ 는 parameter이고  $w$ 는 관측된 데이터여서,  $(\theta, z, \beta)$ 의 Posterior Distribution은 아래와 같음

$$= \frac{(\theta, z, \beta, w) \text{의 Joint Distribution}}{w \text{의 Marginal Distribution}} = \frac{p(\theta, z, \beta, w | \alpha, \eta)}{\int \int \sum_z p(\theta, z, \beta, w | \alpha, \eta) d\theta d\beta}$$

- LDA를 통해 아래의 값을 구하고자 함

- Topic probability of term  $\hat{\beta}_{K,V} = E_{\pi}[\beta_{K,V} | w]$
    - Per-Document topic proportion  $\hat{\theta}_{d,k} = E_{\pi}[\theta_{d,k} | w]$
    - Per-word topic proportion  $\hat{z}_{d,n,k} = Pr_{\pi}(z_{d,n} = k | w)$

### 3. 토픽모델링

- 참고: Multinomial & Dirichlet Distribution

- Multinomial Distribution: 여러 개의 값을 가질 수 있는 독립 확률변수들에 대한 확률분포로, 여러 번의 독립적 시행에서 각각의 값이 특정 횟수가 나타날 확률을 정의

$$p(x_1, x_2, \dots, x_n; n, p_1, \dots, p_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

- Dirichlet Distribution: k차원의 실수 벡터 중 벡터의 요소가 양수이며 모든 요소를 더한 값이 1인 경우에 대해 확률값이 정의되는 분포

$$f(x_1, \dots, x_k; \alpha_1, \dots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i - 1} \quad B(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}$$

- Conjugate Prior
  - posterior와 prior가 동일한 분포를 따르면, prior를 likelihood의 conjugate prior

Likelihood	Conjugate Prior	Posterior
Binomial(N, $\theta$ )	Beta(r,s)	Beta(r+n, s+N-n)
Multinomial( $\theta_1, \dots, \theta_K$ )	Dirichlet( $(\alpha_1, \dots, \alpha_K)$ )	Dirichlet( $\alpha_1 + n_1, \dots, \alpha_K + n_K$ )

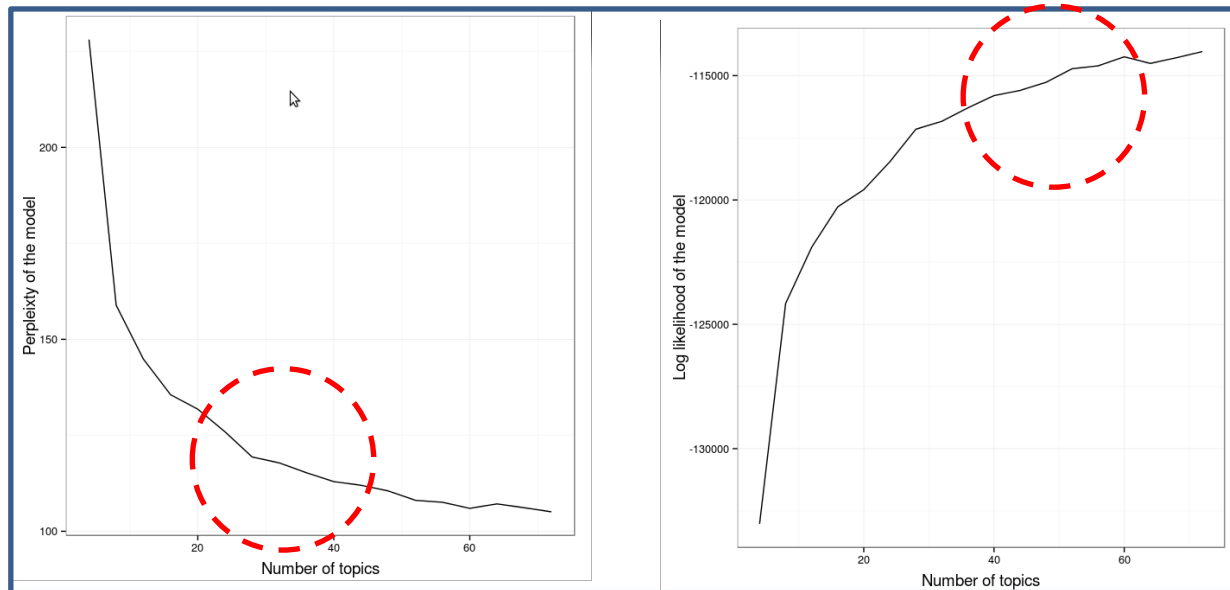
### 3. 토픽모델링

- **Parameter Estimation**

- 데이터의 log-likelihood 를 최대화하는 Parameter 추정
- 이번 분석에서는 Variational Expectation Maximization Algorithm을 사용

$$l(\alpha, \beta) = \log(p(w|\alpha, \beta)) = \log \int \left\{ \sum_z \left[ \prod_{i=1}^N p(w_i|z_i, \beta) p(z_i|\theta) \right] \right\} p(\theta|\alpha) d\theta$$

**Model Selection: Choose K**

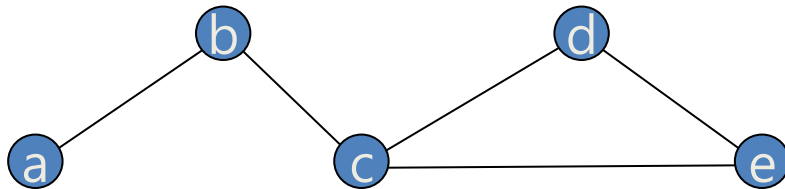




## 4. 인공지능망과 네트워크

- **Graph**

- Actors와 relations (또는 "nodes" and "edges")으로 구성
- Graph는 Node들의 연결에 대한 패턴에 대해 두 가지 방법으로 표현: graphs와 matrix
- 특히 graphics는 "socio-grams " 으로 불리기도 함
- 수학자들은 "directed graphs" "signed graphs" or simply "graphs " 등으로 지칭



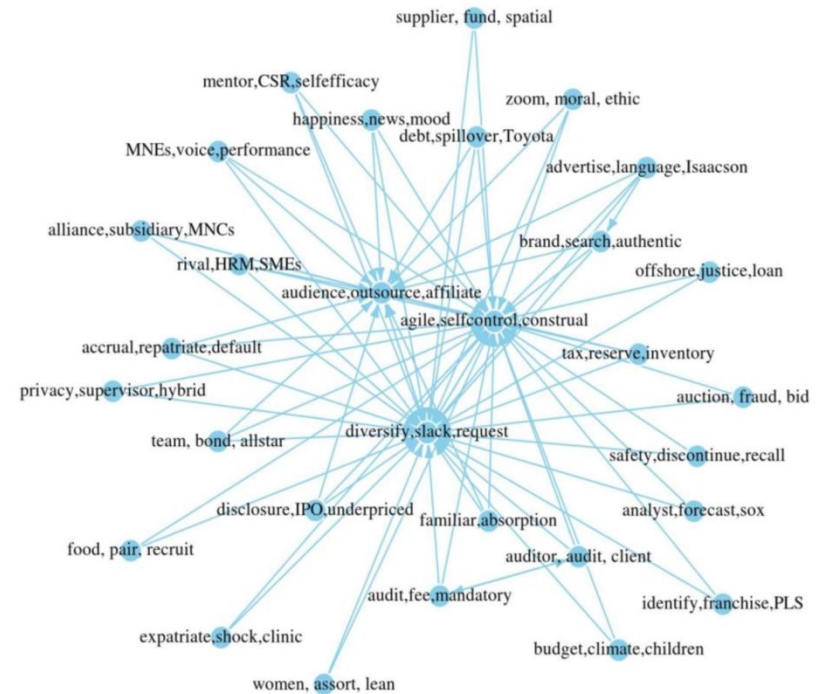
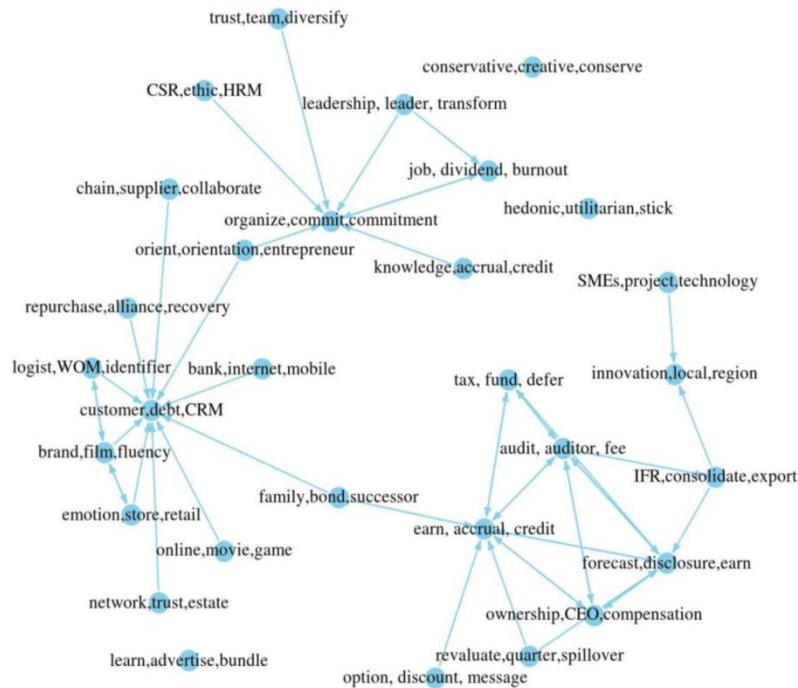
Undirected, binary

	a	b	c	d	e
a		1			
b	1		1		
c		1		1	1
d			1		1
e			1	1	

## 4. 인공지능망과 네트워크

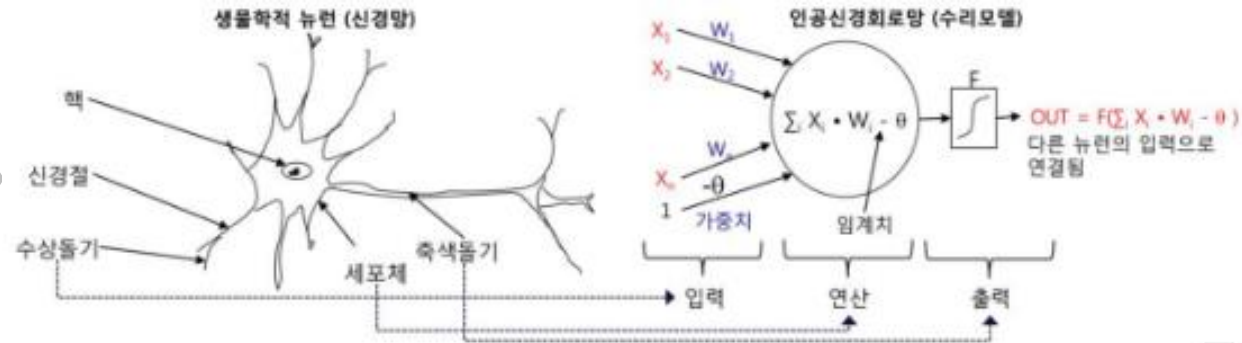
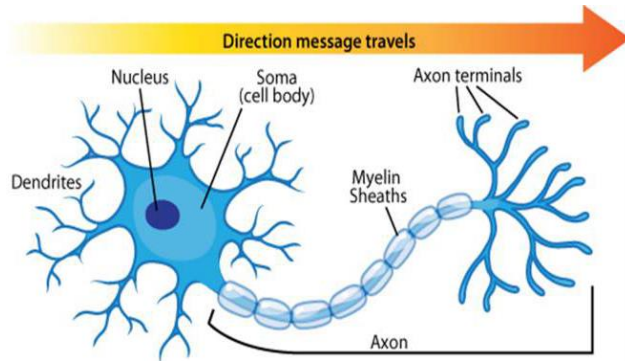
- Topic Network in Business/Economics

- Business 분야: 국내 저널 논문 약 3,600여건, 국외 저널 논문 약 7,700여건이 수집
- Finance/Economy 분야: 국내 저널 논문 약 2,800여건, 국외 저널 논문 약 7,000여건이 수집
- 국내 저널도 영문 초록을 분석대상으로 하여 해외 저널의 Trend와 직접적인 비교를 함



## 4. 인공지능망과 네트워크

### ➤ Artificial Neural Network

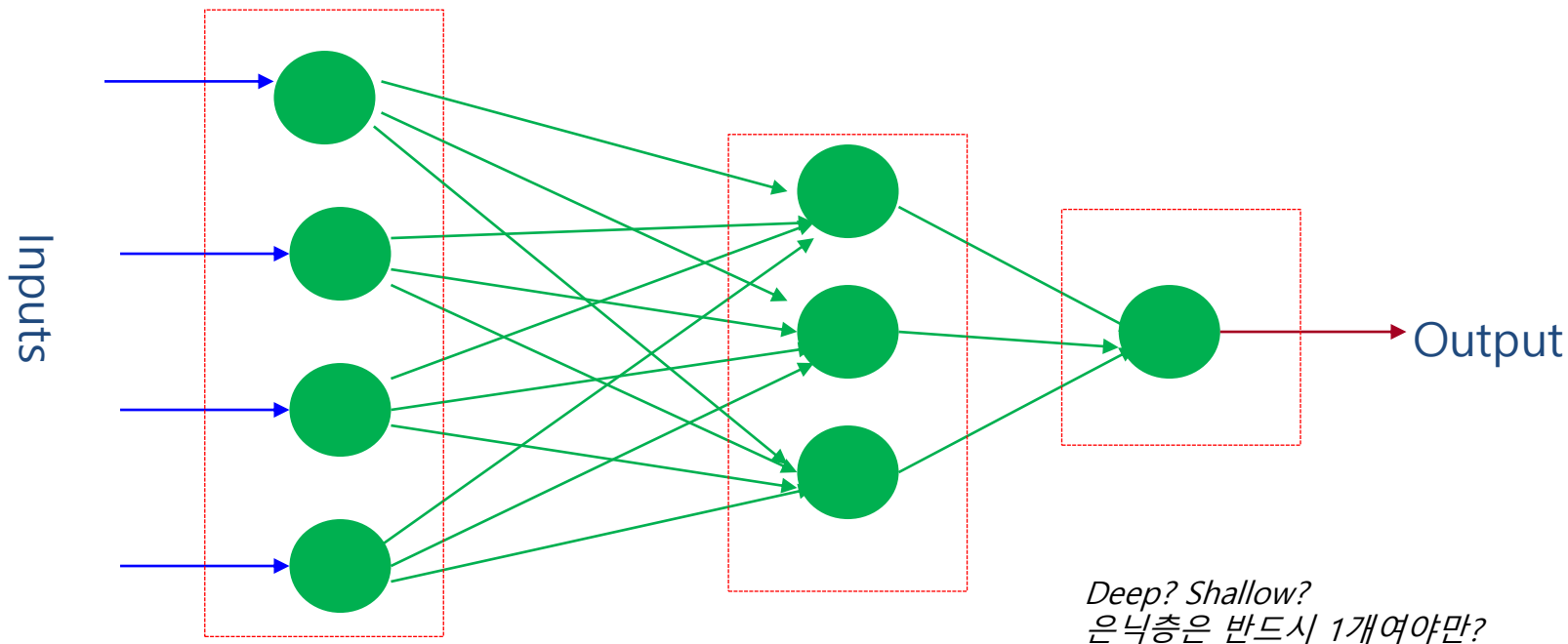


- 뉴런을 모방한 노드들이 각각 Input Layer, Hidden Layer, Output Layer로 구분
- ANN의 목적은 데이터를 입력받아 변환하여 원하는 결과로 출력하는 것
- 예측 성능이 우수하다고 알려진 반면, 모형을 직관적으로 이해하기가 어려우며 수정하기도 어려움

## 4. 인공지능망과 네트워크

### ➤ ANN

- 입력층에 자료가 주어지면 아래 화살표처럼 은닉층으로 전달
- 은닉층의 노드는 주어진 입력에 의해 활성화가 결정
- 은닉층은 다시 출력값을 계산하여 출력층에 값을 전달
- 그 결과에 따라 다시 출력층이 활성화되며, 활성화된 출력층 노드가 출력값을 계산



## 5. Word2vec을 통한 유사도 계산

- 단어의 표현
  - 프랑스+붕어빵-김치=?

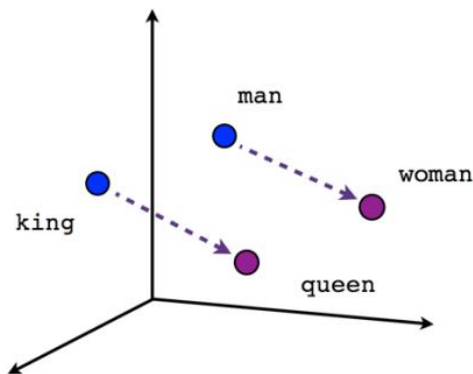
Bag of Words Model	Word Embeddings
<ul style="list-style-type: none"><li>• One hot encoding</li><li>• 문맥정보는 활용하지 않음</li><li>• 사전 내 단어가 큰 벡터의 위치에 1로 표시</li><li>• 예: 사전에서 표현하려는 단어 "ABCD"가 네번째에 위치한다면, 0 0 0 1 0 0 . . . . . 0 0 0 0 이런식으로 표현됨</li></ul>	<ul style="list-style-type: none"><li>• 각 단어를 공간 내 점으로 저장하고 표현: 일반적으로 300개 정도의 dimension</li><li>• 큰 말뭉치를 통해 구현, 비지도학습</li><li>• 예를 들어, "ABCD"는 다음과 같이 표현하는 dimension의 벡터로 표현 :</li><li>• [0.4, -0.11, 0.55, 0.3 . . . 0.1, 0.02]</li></ul>

## 5. Word2vec을 통한 유사도 계산

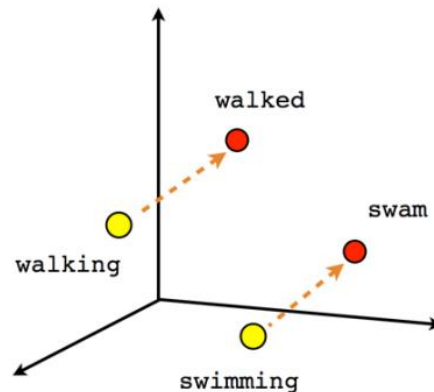
- Word2vec

- Example:

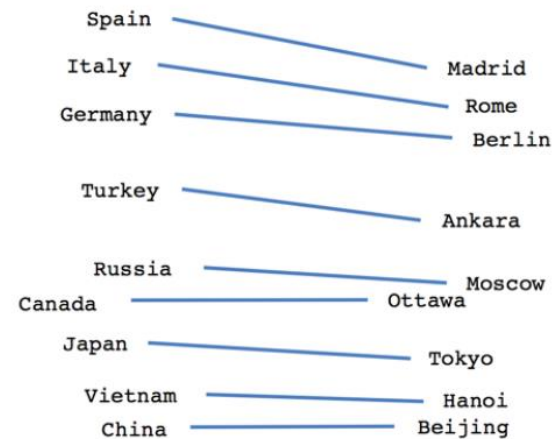
- $\text{vector}[\text{Queen}] = \text{vector}[\text{King}] - \text{vector}[\text{Man}] + \text{vector}[\text{Woman}]$



Male-Female



Verb tense

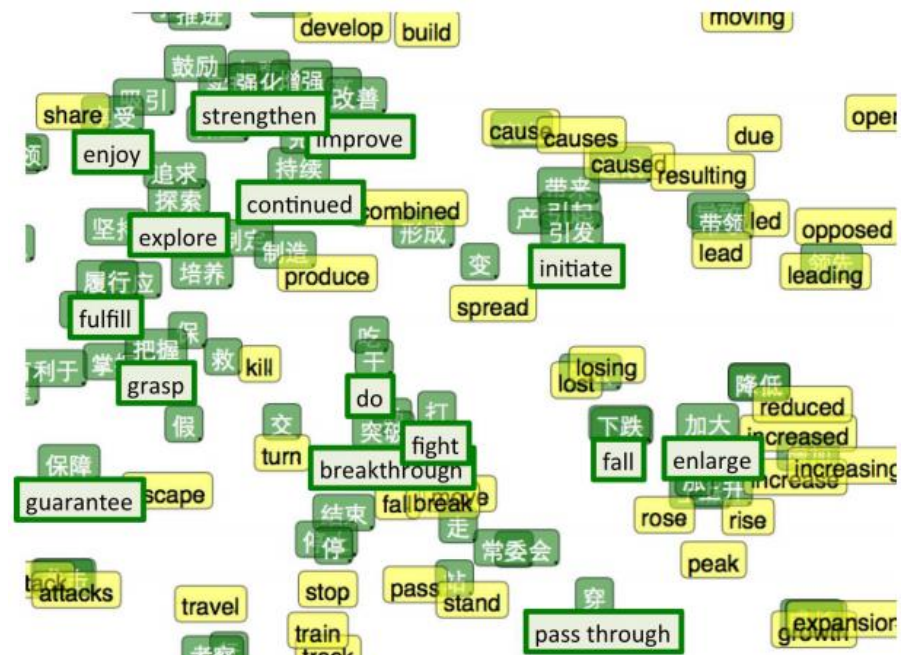


Country-Capital

## 5. Word2vec을 통한 유사도 계산

### • Word2vec Application

- 단어 유사도
- 기계번역
- 관계추출
- 거리/유사도를 사용하는 다양한 방법에 활용



Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

---

Q&A

