텍스트 마이닝

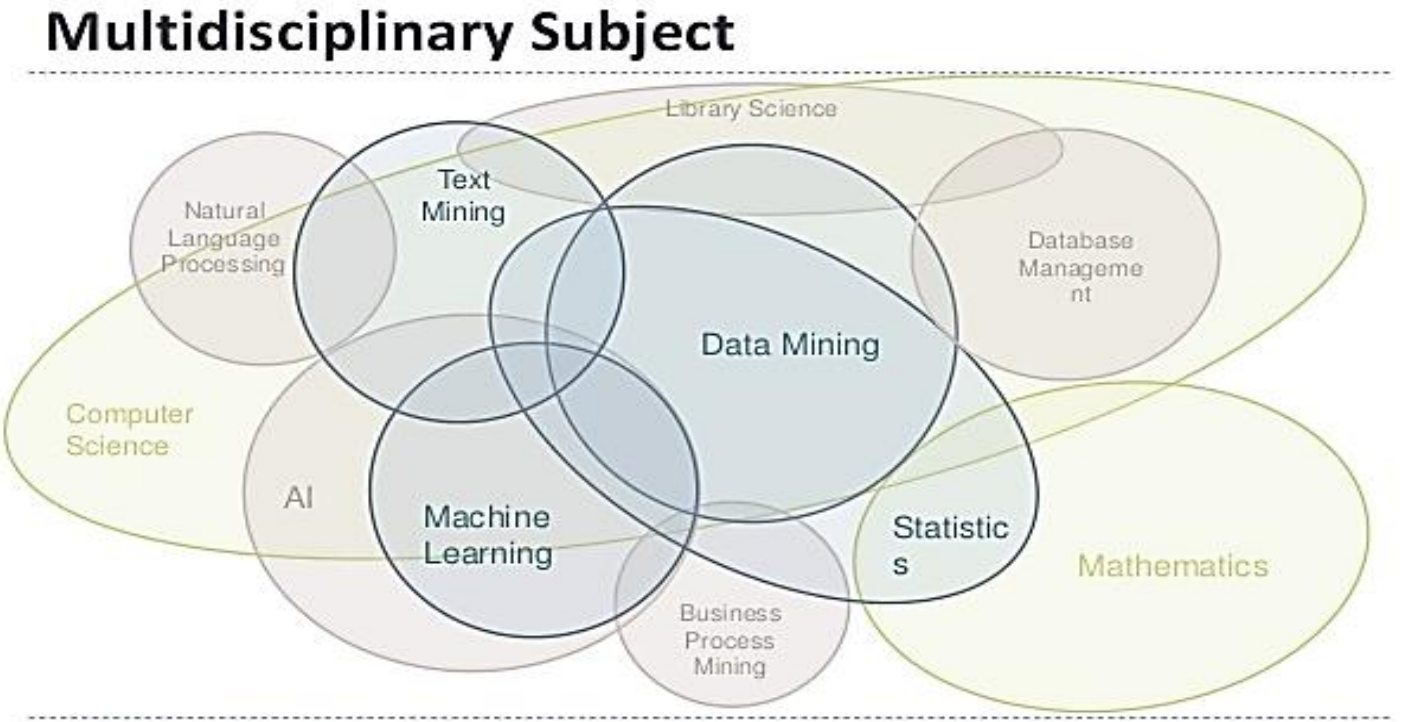
SNU Data Mining Center

신 훈 식

hunsik@dm.snu.ac.kr

Introduction

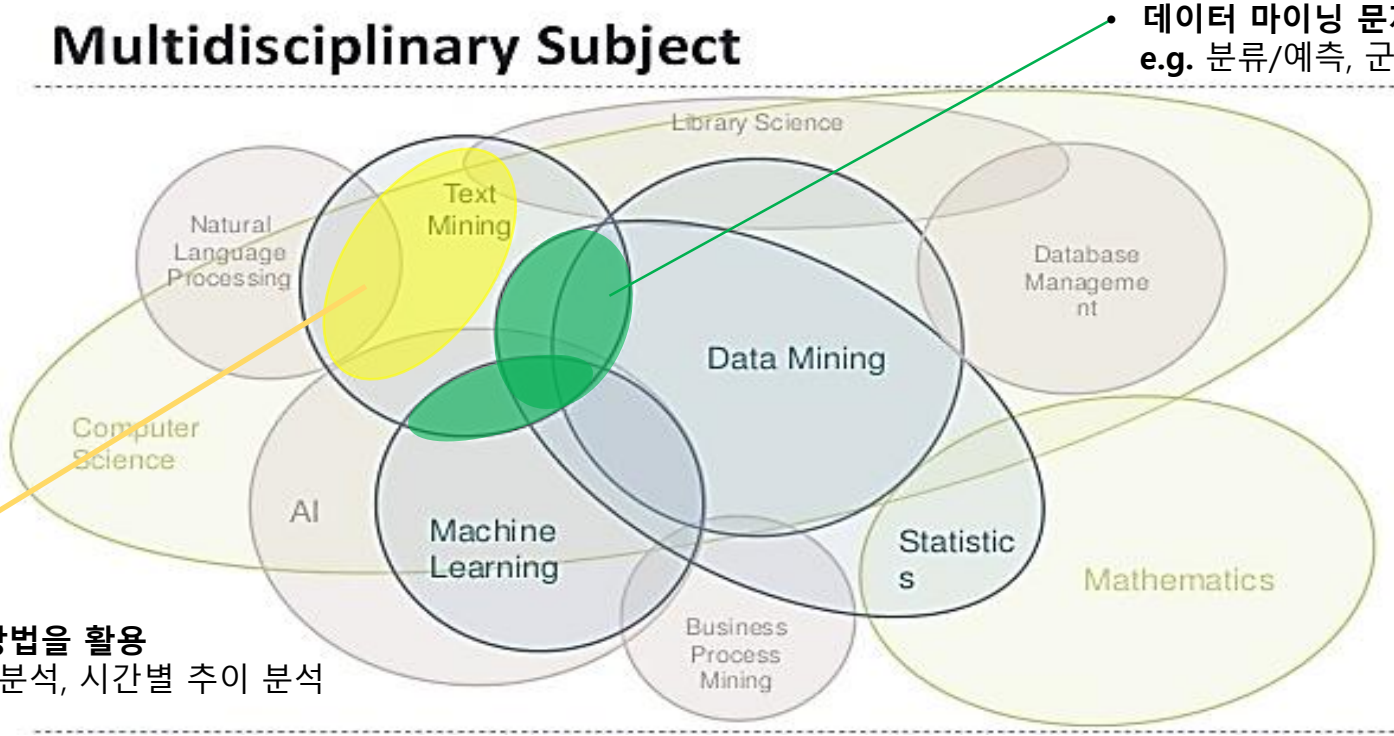
- **텍스트 마이닝이란?**
  - ✓ 자연어 처리 기술(Natural Language Processing)을 기반으로 (대규모) 텍스트 데이터로 부터 의미 있는 **정보**와 **지식**을 추출 하는 기술
  - ✓ 텍스트를 컴퓨터가 인지할 수 있는 **수치 데이터**로 변환 하여 알고리즘을 적용하는 것!



출처 : http://www.slideshare.net/AdrianCuyugan/text-mining-association-rules-and-decision-tree-learning-48455111

Introduction

- **텍스트 마이닝이란?**
  - ✓ 자연어 처리 기술(Natural Language Processing)을 기반으로 (대규모) 텍스트 데이터로 부터 의미 있는 **정보**와 **지식**을 추출 하는 기술
  - ✓ 텍스트를 컴퓨터가 인지할 수 있는 **수치 데이터**로 변환 하여 알고리즘을 적용하는 것!



- **데이터 마이닝 문제를 활용하여 해결!**
  e.g. 분류/예측, 군집화, 이상치 탐색…

- **기존 분석 방법을 활용**
  e.g. 빈도수 분석, 시간별 추이 분석

출처 : http://www.slideshare.net/AdrianCuyugan/text-mining-association-rules-and-decision-tree-learning-48455111

적용 분야

1. 문서 분류 - 스팸 메일 분류하기
   - ✓ 계정을 통해 전달되는 메일의 내용을 사용자가 직접 확인하지 않고 메일 서버에서 자동으로 스팸 메일인지 아닌지를 구분
   - ✓ 유해 정보나 바이러스 노출로부터 사용자 보호

데이터 마이닝 문제 – 분류 알고리즘을 활용!

<전달된 메일>

수치 데이터

$$\begin{bmatrix} 0 \\ 3 \\ 4 \\ 7 \\ \vdots \\ 1 \end{bmatrix}$$

<변환된 수치 데이터>

Classification

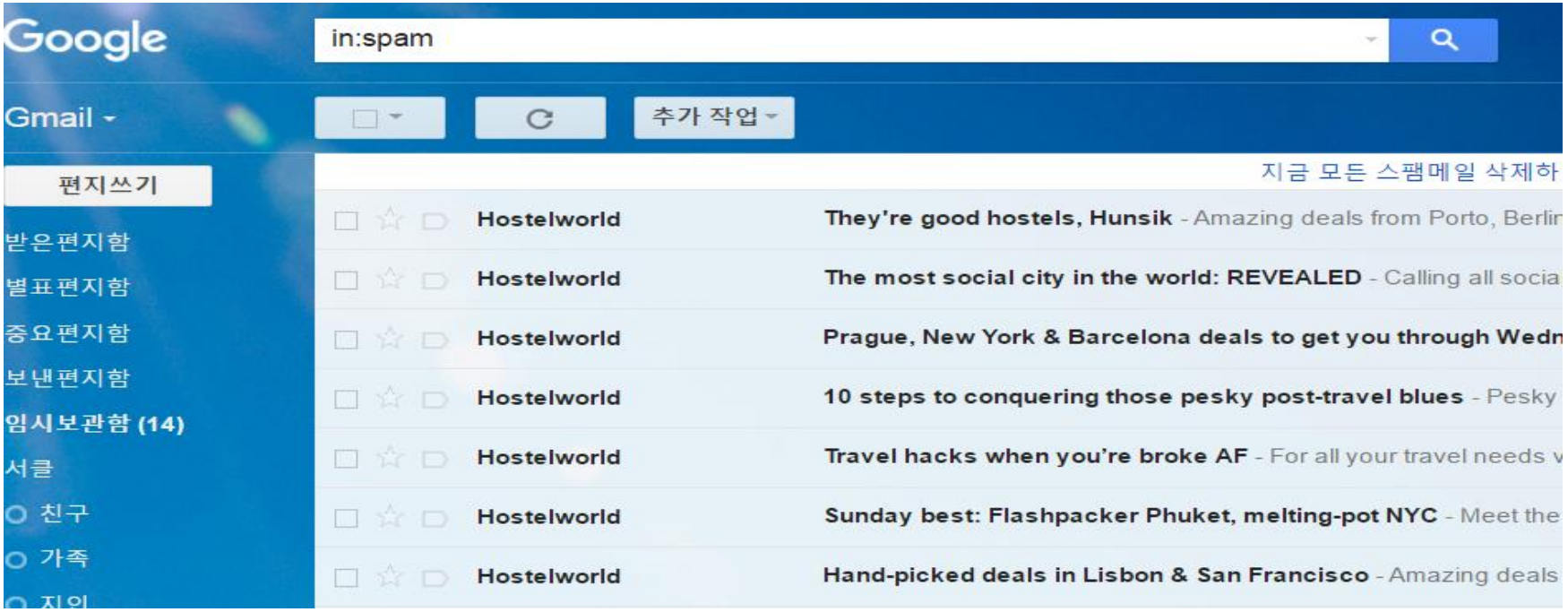
- Linear regression
- SVM
- Random Forest

Or

적용 분야

1. 문서 분류 - 스팸 메일 분류하기
   - ✓ 계정을 통해 전달되는 메일의 내용을 사용자가 직접 확인하지 않고 메일 서버에서 자동으로 스팸 메일인지 아닌지를 구분
   - ✓ 유해 정보나 바이러스 노출로부터 사용자 보호

스팸 메일 분류 예시 – Gmail



적용 분야

1. 문서 분류 - 스팸 메일 분류하기

- ✓ 계정을 통해 전달되는 메일의 내용을 사용자가 직접 확인하지 않고 메일 서버에서 자동으로 스팸 메일인지 아닌지를 구분
- ✓ 유해 정보나 바이러스 노출로부터 사용자 보호

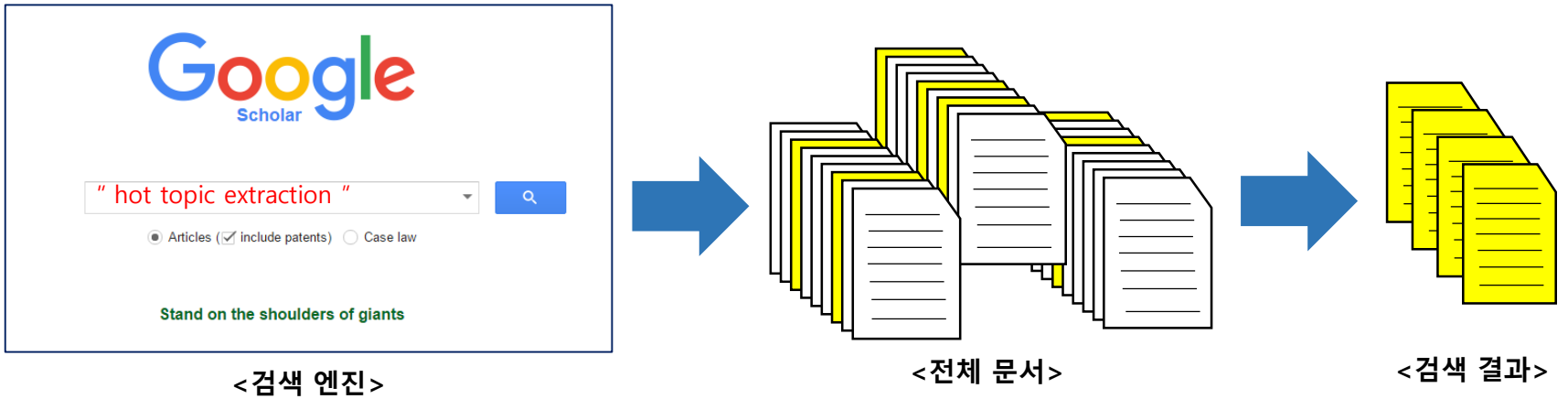
데이터 마이닝 문제 – 분류 알고리즘을 활용!

적용 분야

2. 문서 검색 – 주어진 단어(문장)와 가장 비슷한 문서 검색

- ✓ 전체 문서로부터 주어진 단어와 문장과의 유사도(가까운 정도)를 계산하여 원하는 문서를 검색(e.g. 구글링)
- ✓ 유사문서 검색 및 유사 문서 군집화 가능

주어진 단어(문장)과 문서의 유사도 계산



<검색 엔진>

<전체 문서>

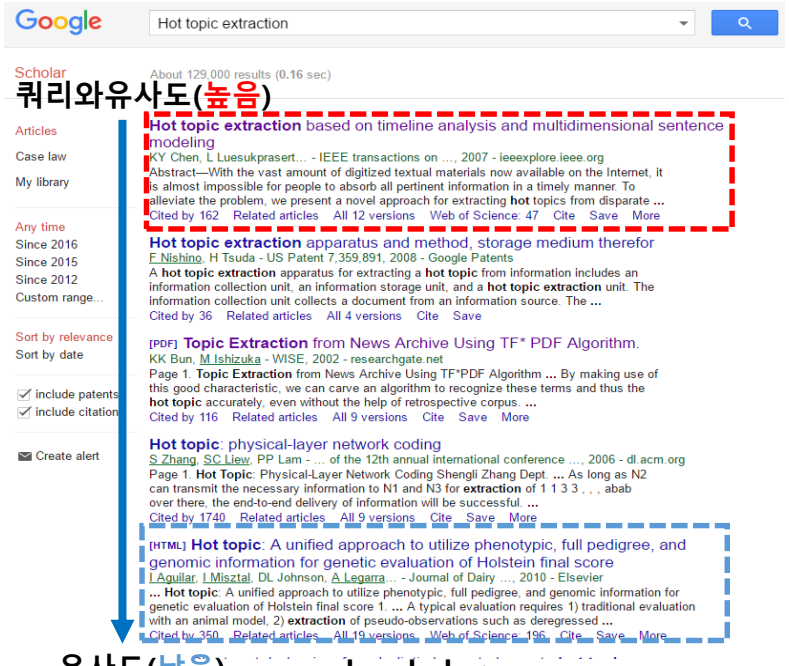
<검색 결과>

적용 분야

2. 문서 검색 – 주어진 단어(문장)와 가장 비슷한 문서 검색

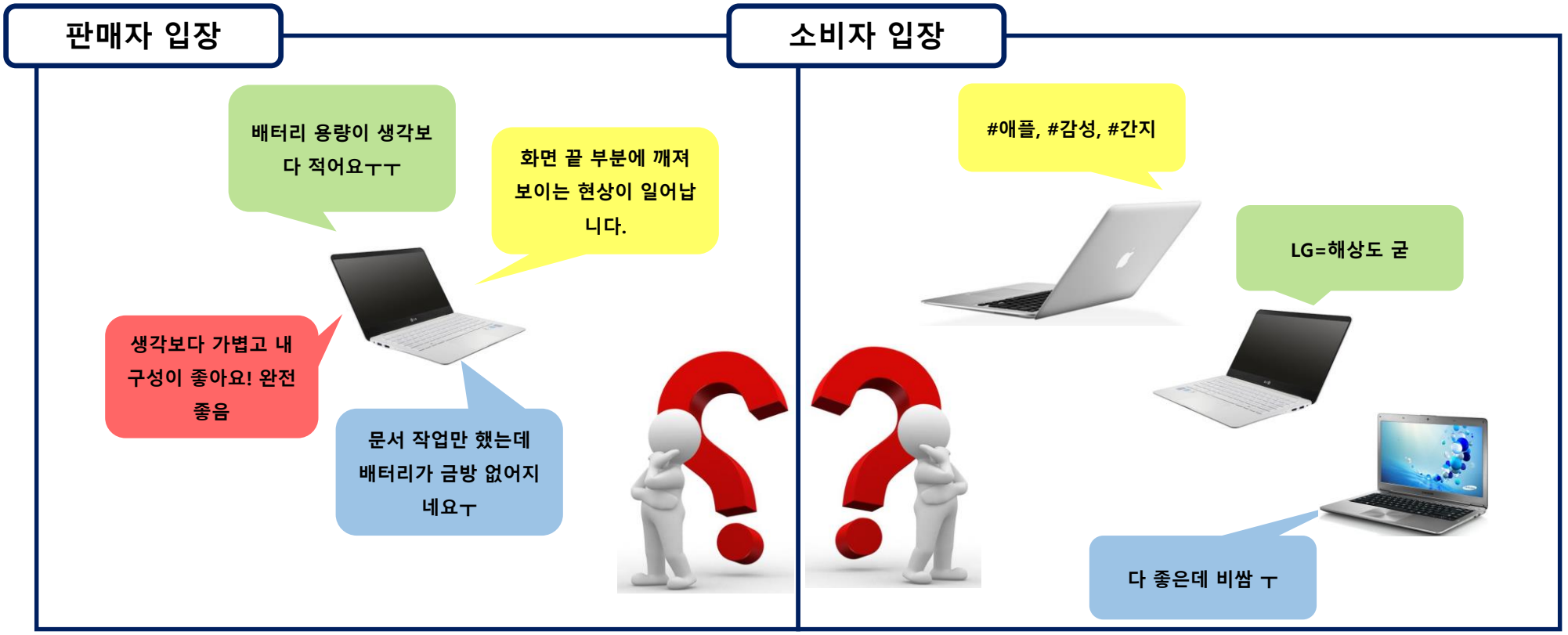
- ✓ 전체 문서로부터 주어진 단어와 문장과의 유사도(가까운 정도)를 계산하여 원하는 문서를 검색(e.g. 구글링)
- ✓ 유사문서 검색 및 유사 문서 군집화 가능

주어진 단어(문장)과 문서의 유사도 계산



실제 사례 (1)

- **제품 리뷰 데이터 분석**
  - ✓ 특정 제품에 대해 소비자가 어떻게 생각하는지 파악하는 것은 판매자 입장에서 매우 중요하다.
  - ✓ 소비자 역시 제품을 구매하기 전, 제품에 대한 평가 또는 사용 리뷰에 대해 정보를 얻는 것은 현명한 소비를 위해 중요.
  - ✓ 판매자, 소비자 모두 제품에 대한 리뷰는 생산, 구매 입장에서 중요한 정보



실제 사례 (1)

- **제품 리뷰 데이터 분석**
  - ✓ 특정 제품에 대해 소비자가 어떻게 생각하는지 파악하는 것은 판매자 입장에서 매우 중요하다.
  - ✓ 소비자 역시 제품을 구매하기 전, 제품에 대한 평가 또는 사용 리뷰에 대해 정보를 얻는 것은 현명한 소비를 위해 중요.
  - ✓ 판매자, 소비자 모두 제품에 대한 리뷰는 생산, 구매 입장에서 중요한 정보

<연관어 네트워크 분석>

2009년

No.	브랜드	빈도	비율
1	좋어좋다	1710	13.0%
2	위축되다	1591	12.1%
3	깊다	752	5.7%
4	어렵다	670	5.1%
5	심각하다	524	4.0%
6	필요하다	448	3.4%
7	힘들다	365	2.8%
8	감회하다	343	2.6%
9	선호하다	316	2.4%
10	가능하다	276	2.1%
11	뜨다	249	1.9%
12	다르다	242	1.8%
13	유려하다	229	1.7%
14	바꾸다	218	1.7%
15	강하다	184	1.4%
16	비슷하다	167	1.3%
17	따오르다	167	1.3%
18	꺼리다	146	1.1%
19	저렴하다	141	1.1%
20	좋다	138	1.1%
21	부족하다	133	1.0%
22	불투명하다	114	0.9%
23	각광받다	113	0.9%
24	익숙하다	110	0.8%
25	각광받다	108	0.8%

2010년

No.	브랜드	빈도	비율
1	위축되다	487	11.2%
2	좋어좋다	434	10.0%
3	유려하다	204	4.7%
4	어렵다	190	4.4%
5	심각하다	166	3.8%
6	힘들다	139	3.2%
7	불투하다	104	2.4%
8	더하다	102	2.3%
9	깊다	91	2.1%
10	바꾸다	90	2.1%
11	적다	88	2.0%
12	심하다	88	2.0%
13	뜨다	88	2.0%
14	가능하다	87	2.0%
15	필요하다	87	2.0%
16	따오르다	84	1.9%
17	확인하다	82	1.9%
18	선호하다	79	1.8%
19	빠르다	69	1.6%
20	부족하다	69	1.6%
21	싸다	67	1.5%
22	비싸다	61	1.4%
23	익숙하다	59	1.4%
24	사롭다	57	1.3%
25	감회하다	55	1.3%

2011년

No.	브랜드	빈도	비율
1	위축되다	511	14.6%
2	어렵다	357	10.2%
3	좋어좋다	300	8.6%
4	유려하다	184	5.3%
5	심각하다	152	4.3%
6	깊다	88	2.5%
7	자랑하다	84	2.4%
8	불안하다	76	2.2%
9	다르다	71	2.0%
10	뜨다	71	2.0%
11	필요하다	62	1.8%
12	아쉽다	60	1.7%
13	힘들다	55	1.6%
14	유명하다	54	1.5%
15	바꾸다	47	1.3%
16	빠다	45	1.3%
17	선호하다	43	1.2%
18	강하다	42	1.2%
19	행복하다	42	1.2%
20	꺼리다	41	1.2%
21	익숙하다	40	1.1%
22	중요하다	40	1.1%
23	부족하다	38	1.1%
24	어둡다	37	1.1%
25	끌리다	35	1.0%

2012년

No.	브랜드	빈도	비율
1	위축되다	622	13.7%
2	깊다	292	6.4%
3	좋어좋다	276	6.1%
4	어렵다	262	5.8%
5	유려하다	243	5.3%
6	심각하다	160	3.5%
7	선호하다	154	3.4%
8	충분하다	149	3.3%
9	다르다	125	2.7%
10	개혁하다	96	2.1%
11	심하다	91	2.0%
12	끌리다	81	1.8%
13	싸다	66	1.4%
14	나빠지다	65	1.4%
15	필요하다	64	1.4%
16	힘들다	59	1.3%
17	지나치다	53	1.2%
18	익숙하다	53	1.2%
19	불투명하다	52	1.1%
20	꺼리다	52	1.1%
21	저렴하다	52	1.1%
22	개선하다	51	1.1%
23	뜨다	49	1.1%
24	감회하다	48	1.1%
25	사롭다	44	1.0%

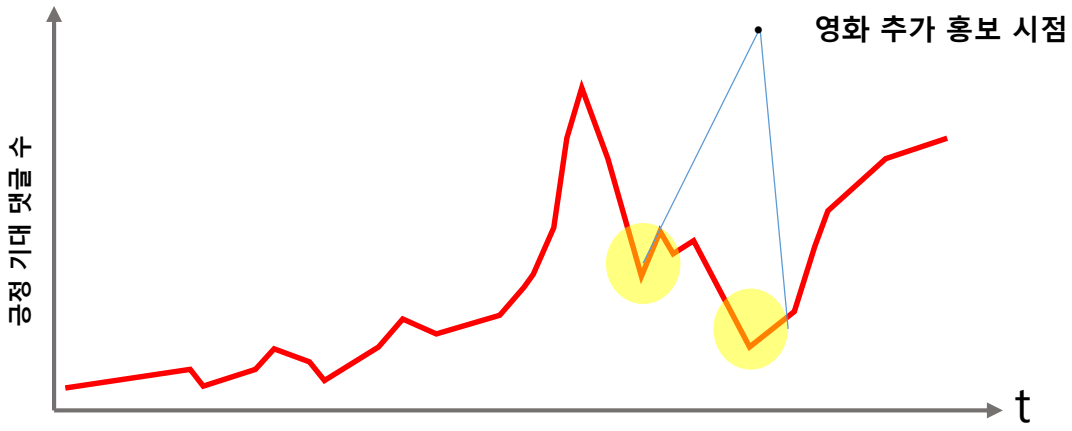
<시기별 키워드 변화 추이 분석>

실제 사례 (2)

- **영화 리뷰 데이터 분석**
  - ✓ 영화 개봉 이후 관람객의 리뷰 텍스트 분석하여 긍정과 부정 댓글 차이를 비교
  - ✓ 개봉 전 영화 기대에 대한 댓글 추이를 분석하여 영화 홍보 시점 탐색

<네이버 긍정 리뷰 워드클라우드>

<네이버 부정 리뷰 워드클라우드>



<네이버 영화 평점 및 리뷰 >

실제 사례 (3)

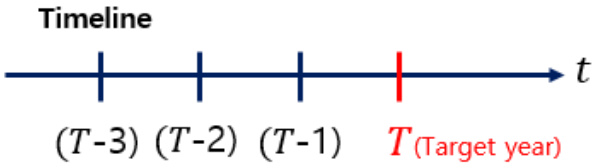
- **소셜 미디어 분석**
  - ✓ 트위터와 같은 소셜 미디어에서 특정 사건에 대한 사용자들의 생각과 의견을 분석

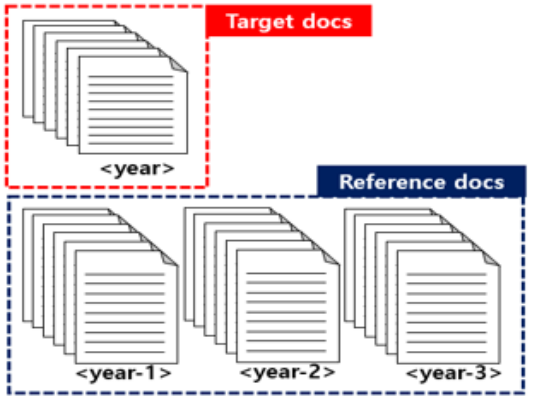
<다음소프트 – 국내 트위터, 블로그 키워드 분석>

진행 중인 프로젝트 (1)

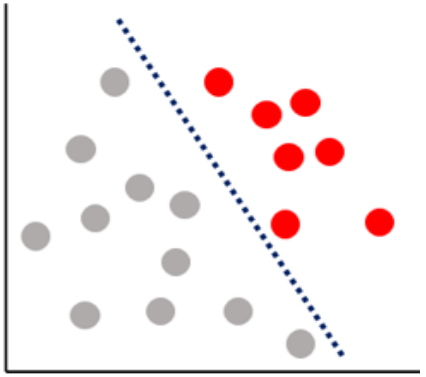
- **세계 중앙 은행장 연설문 키워드 분석**
  - ✓ 매년 세계 중앙 은행장은 세계 금융 변화 및 정책에 대한 연설문을 발표
  - ✓ 연도별 은행장 연설문의 키워드를 추출하여 미래의 금융 시장 변동을 예측

- **Process**





< Data representation>



<Linear classification>

2008-word	2008-coeff
energy/nn food/nn	0.534844862
market/nn turmoil/nn	0.517357994
food/nn prices/nns	0.443155044
citizenry/nn	0.393837544
daiwa/nnp	0.381499054
turmoil/nn	0.37646525
bank/nnp boston/nnp	0.34368244
banking/nn operations/nns	0.324523395
documents/nns us/nnp	0.297546025
credit/nn default/nn	0.29299273
materials/nns prices/nns	0.288758651
cushions/nns	0.288582679

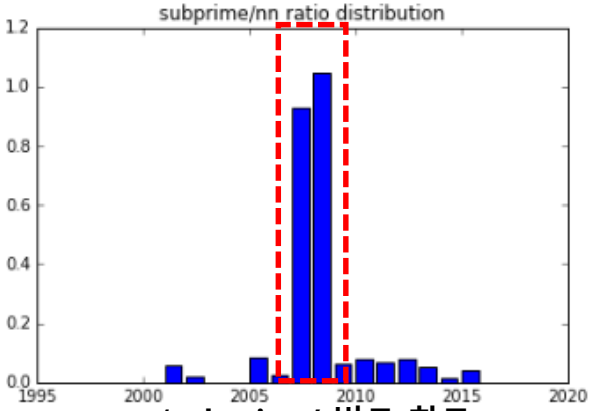
<Features with β != 0>

진행 중인 프로젝트 (1)

- **세계 중앙 은행장 연설문 키워드 분석**
  - ✓ 매년 세계 중앙 은행장은 세계 금융 변화 및 정책에 대한 연설문을 발표
  - ✓ 연도별 은행장 연설문의 키워드를 추출하여 미래의 금융 시장 변동을 예측

2007-word	2007-coeff	2007-score	>score
randall/nnp	0.387109916	2.909908855	1
subprime/nn	0.289732813	4.659367052	1
chamber/nnp	0.260342591	2.056653275	1
development/nnp	0.257122509	0.812340462	1
expertise/nn	0.239707193	1.588144338	1
report/nn	0.238650112	0.334000945	0
convenience/nn	0.226272537	5.304532087	1
indicator/nn	0.225067039	1.170813289	1
group/nn	0.210731288	1.088755094	1
singapore/nnp	0.194199668	3.815705644	1
banking/nn system/nn	0.191568291	0.803184998	1
inflation/nn dynamics/nns	0.177899355	4.652373927	1
supervisory/nn	0.169899742	0.906977001	1
wages/nns	0.164835364	1.63474068	1
erm/nnp	0.163595515	3.534751172	1
turbulence/nn	0.162516606	4.625856061	1
case/nn	0.145262798	0.51782998	0
february/nnp	0.144545942	0.289446262	0
conditions/nns	0.141359015	0.95234702	1
inflation/nn process/nn	0.140898709	5.523920393	1

<2007 세계 중앙 은행장 연설문 키워드>



<'subprime' 빈도 차트>

<2008년 S&P 차트>

진행 중인 프로젝트 (2)

- **미래 사회 키워드 추출**
  - ✓ 미래 관련 빅데이터를 수집, 추출, 가공, 분석하여 미래 한국사회의 핵심 키워드 도출
  - ✓ 국내외 전문가 및 대중의 전망이나 예측을 포함한 보고서, 뉴스기사, SNS를 분석

토픽	토픽 키워드 리스트 일부
1	서비스, 필요, 클라우드, 가치, 최신, 관계, 고령화, 제어, 운용, 주도, 감시
2	달러, 모델, 방송, 이슈, 서버, 노력, 매출, 시리즈, 화면, 암호화, 중앙, 대화
3	의료, 창출, 확보, 작업, 헬스, 모니터링, 동작, 환자, 의무, 병원, 실험, 스포츠
4	인터넷, 빅데이터, 능력, 장비, 응답, 장기, 가구, 정치, 첨단, 주택, 에코
5	증가, 중심, 구축, 비중, 사물인터넷, 비교, 기록, 영역, 업무, 아이폰, 로열티
6	콘텐츠, 디지털, 애플, 가상현실, 침해, 전파, 디지털콘텐츠, 포털, 커뮤니티, 영화
7	스마트폰, 에너지, 플랫폼, 포커스, 고용, 설정, 인력, 상승, 스마트워치, 태양전지
8	개발, 데이터, 진흥, 원고, 반도체, 한계, 하둡, 프로토콜, 실리콘밸리, 자전거, 유전자
9	중국, 경제, 수출, 규모, 향후, 지역, 방향, 가격, 목표, 공공, 시사점, 점유, 벤처, 유발
10	사용, 프린터, 공식, 내장, 디스플레이, 육성, 스파크, 블루투스, 로그, 협회
11	소셜미디어, 보고서, 페이스북, 위험, 영상, 비즈니스, 스토리, 필자, 파급, 역기능
12	국가, 세계, 연결, 인간, 지식, 차량, 통합, 개념, 지능, 판단, 행위, 실현, 도시, 과학
13	자동차, 인프라, 부품, 교통, 감지, 협업, 착용, 아이패드, 운전자, 증강현실, 안경
14	보호, 개인정보, 센터, 활성, 프린팅, 로봇, 관심, 광고, 이점, 움직임, 우수, 법제도
15	지원, 투자, r&d, 동력, 문화, 결제, 효율, 구현, 창업, 아마존, 자동화, 재생, 투자자

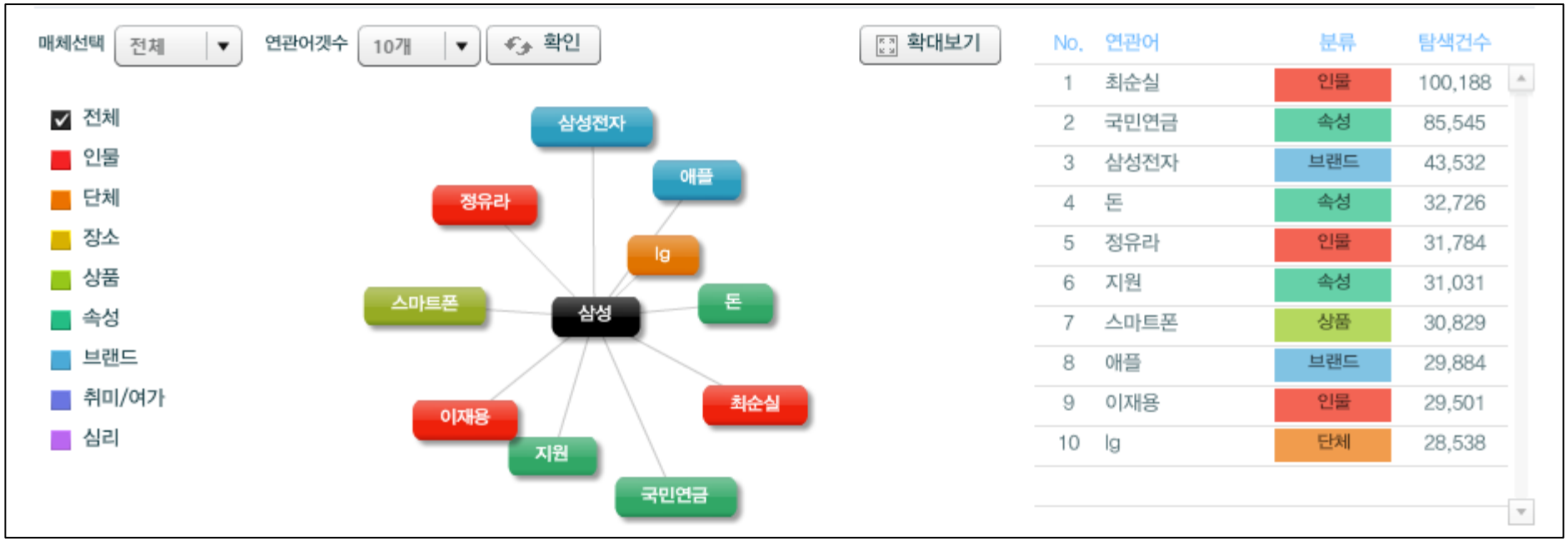
<추출된 토픽 키워드 리스트 일부>

활용 분야..?

- 미국 대선으로 본 소셜 네트워크의 선거캠페인 기능 분석
  - 지난 미국 대선 주자였던 힐러리, 트럼프의 소셜 네트워크 계정에 댓글을 분석하여 어떤 차이가 있는지 어떤 효과를 가졌는지 분석

활용 분야..?

- 기업 마케팅 관련 데이터 마이닝
  - 특정 기업이나 제품에 관련된 키워드를 추출 및 네트워크 분석



< Social metrics 예시 >

활용 분야..?

- 국내 여행객 vs 방한 외국인 여행 루트 비교 분석
  - 블로그 및 SNS, 검색어를 살펴보고 국내 여행객과 방한 외국인 여행 루트에는 어떤 차이가 있는지 살펴보고 유형화 하기

실습 범위

- **텍스트 표현법**
  - Discrete representation(Bag-of-words)
  - Distributed representation(word2vec, doc2vec)

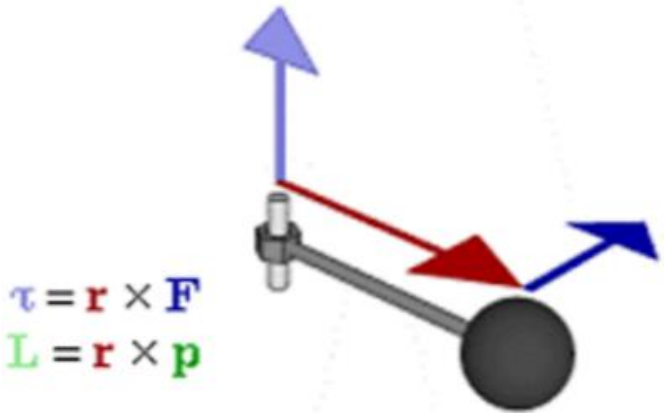
- **분석기법**
  - 문서 분류(+감성 분석)
  - 유사 문서 검색 및 군집화

텍스트 표현법

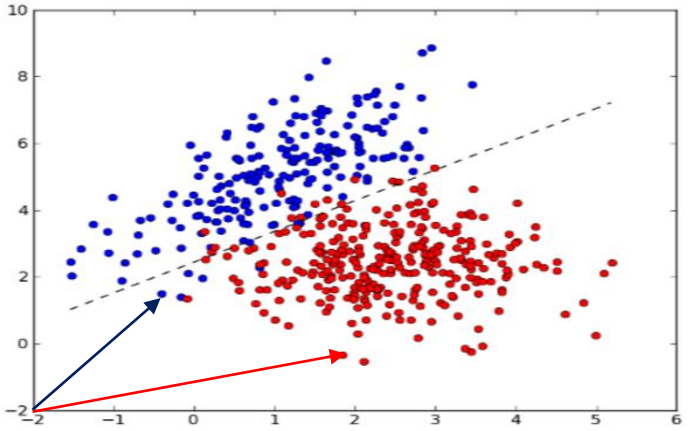
- 텍스트 데이터에 기존의 기계학습 알고리즘을 적용하거나 유사도와 같은 척도(measure)를 사용하기 위해서는 텍스트를 컴퓨터가 인식할 수 있는 수치로 표현해야 함

벡터(Vector) 표현법

- 물리 벡터 : 벡터란 크기와 방향을 갖는 물리량
- 위치벡터 : 유클리드 공간의 모든 벡터들을 평행이동하여 특정 시점(보통 원점)을 같게하여 표현



< 물리 벡터 - torque>



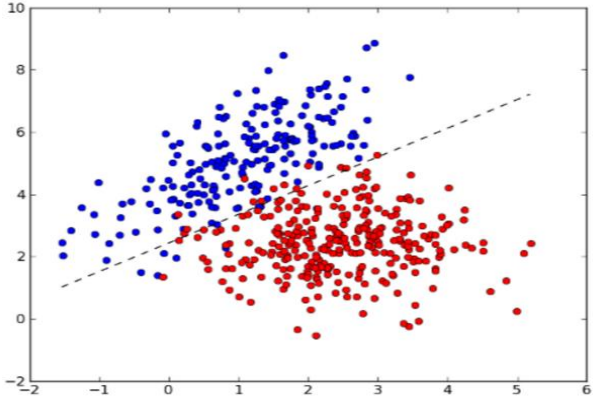
< 위치 벡터 >

텍스트 표현법

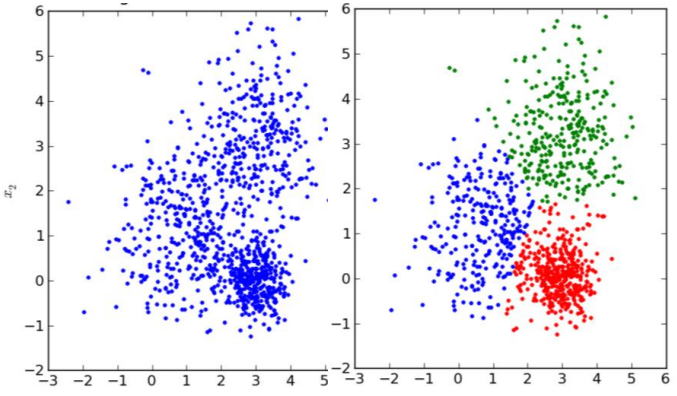
- 텍스트 데이터에 기존의 기계학습 알고리즘을 적용하거나 유사도와 같은 척도(measure)를 사용하기 위해서는 텍스트를 컴퓨터가 인식할 수 있는 수치로 표현해야 함

벡터(Vector) 표현법 - 위치벡터

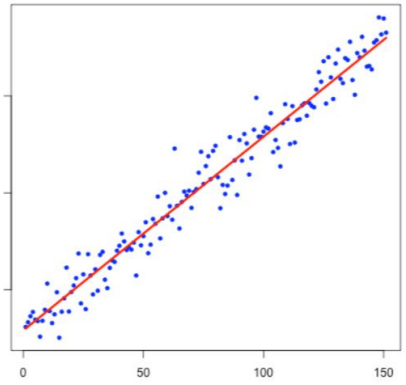
- 위치벡터 : 유클리드 공간의 모든 벡터들을 평행이동하여 특정 시점(보통 원점)을 같게하여 표현
  → 데이터를 수치 (위치)벡터로 표현하여 다양한 기계학습 알고리즘(분류, 군집화 및 회귀분석) 적용 가능
  → 텍스트를 수치 벡터로 표현 필요



< classification task >



< clustering task >



< regression task >

텍스트 표현법

- 텍스트 데이터에 기존의 기계학습 알고리즘을 적용하거나 유사도와 같은 척도(measure)를 사용하기 위해서는 텍스트를 컴퓨터가 인식할 수 있는 수치로 표현해야 함
- 최근에는 단순히 수치로 표현하는 것이 아니라 문서, 문장 및 단어 의미를 고려할 수 있음

Discrete representation	**Distributed representation**
- One-hot vector/ Bag-of-word vector 'dog' = $\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$ 'cat' = $\begin{bmatrix}0\\1\\0\\0\end{bmatrix}$ Doc1 = $\begin{bmatrix}12\\0\\0\\3\\1\\0\\5\end{bmatrix}$	- Word2vec, Doc2vec 'dog' = $\begin{bmatrix}0.5\\0.3\\-0.1\\1\end{bmatrix}$ 'cat' = $\begin{bmatrix}0.8\\-0.3\\-0.2\\0.6\end{bmatrix}$ Doc1 = $\begin{bmatrix}0.68\\0.23\\0.10\\-0.41\\0.90\\0.51\\-0.33\end{bmatrix}$
• Frequency 기반으로 표현하는 방법 • 구성 변수들을 직관적으로 이해 가능함 • 전처리 과정이 뚜렷하지 않으며 단어 빈도가 낮은 경우 중요하지 않게 판별됨	• Neural Network를 통해 continuous vector로 변환 가능 • 단어 별 유사도 계산가능 - 'king' - 'man' + 'woman' → closest('Queen')

실습 범위

- **텍스트 표현법**
  - Discrete representation(Bag-of-words)
  - Distributed representation(word2vec, doc2vec)

- **분석기법**
  - 문서 분류(+감성 분석)
  - 유사 문서 검색 및 군집화

텍스트 표현법 : Bag-of-words

- Bag-of-words : 문서를 문법과 단어의 순서를 고려하지 않고 단어들의 빈도수로 표현하는 방법

Bag-of-words model 예시

- **Doc1** : { "You are a very good boy." }
- **Doc2** : { "You are also a good girl too." }
- **Doc3** : { "We are the world" }

X1 "You"	X2 "are"	X3 "a"	X4 "good"	X5 "boy"	X6 "girl"	X7 "the"	X8 "world"
1	1	1	1	1	0	0	0
1	1	1	1	0	1	0	0
0	1	0	0	0	0	1	1

Bag-of-words model 장단점

- **장점**
  - Frequency 기반으로 표현하는 방법으로 구성 변수들을 직관적으로 이해 가능함

  Ex) Sports 관련 문서의 경우 'player'나 'referee' 단어의 빈도수가 다른 문서에 비해 높음

- **단점**
  - 단어의 수가 증가할 수록 차원이 엄청나게 증가함.(이는 분류, 군집화 등의 성능을 떨어뜨림)
  - 빈도가 낮은 단어의 경우 중요하지 않게 판별됨
  - 문맥을 고려하지 않음. 즉, '밤이 길다' 와 '밤이 맛있다.' 의 '밤'을 동일하게 여김

텍스트 표현법 : Bag-of-words

Documents Acquisition

- Transform from 'JSON(JavaScript Object Notation)' type to 'txt' type

2008-07-03-stocks-fall-in-europe-asia-uks-ftse-100-enters-bear-market.json


```
{"crawledAt":"2015-06-30 19:54","thumbnail":"http://media.gotraffic.net/images/iRlI.nGleFlE/v1/400x225.jpg",
    "writtenAt":"2008-07-03T07:20:57+00:00",
    "bodySnippest":"July 3 (Bloomberg) -- Stocks dropped in Europe and Asia as oil topped $145 a barrel, damping earnings prospects f
    "topics":["Canada","Oil","Bear Market","Stocks","Asia","Europe","London","Earnings","Dow Jones Industrial Average","Germany"],
    "headline":"Stocks Decline in Europe, Asia; U.S. Index Futures Advance",
    "articleUrl":"http://www.bloomberg.com/news/articles/2008-07-03/stocks-fall-in-europe-asia-uks-ftse-100-enters-bear-market",
    "content":"Share on Facebook Share on Twitter Share on LinkedIn Share on Reddit Share on Google+ E-mail July 3 (Bloomberg) -- Sto
    "crawlerVersion":"1.0"}
```


Json type

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

Share on Facebook Share on Twitter Share on LinkedIn Share on Reddit Share on Google+ E-mail July 3 (Bloomberg
arnings, said Mark Bon, a London-based fund manager at Canada Life, which oversees about $15 billion. ``Sentim
sses and the worst housing slump in 30 years dimmed the earnings outlook for banks and retailers. The index re
according to a survey by the Chartered Institute of Purchasing and Supply, adding to evidence that Britain is
elmaker, dropped 4.6 percent to 53.37 euros. BHP Billiton, the world's biggest mining company, lost 1.9 percen
n this story: Sarah Jones in London at sjones35@bloomberg.net To contact the editor responsible for this story

txt type

텍스트 표현법 : Bag-of-words

Tokenizing

- Separate each document into an ordered sequence of terms

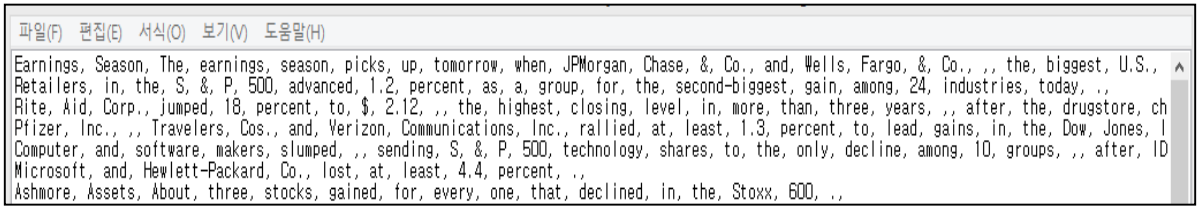
'U.S. stocks extended a four-day rally as retailers gained after reporting claims dropped more than forecast'

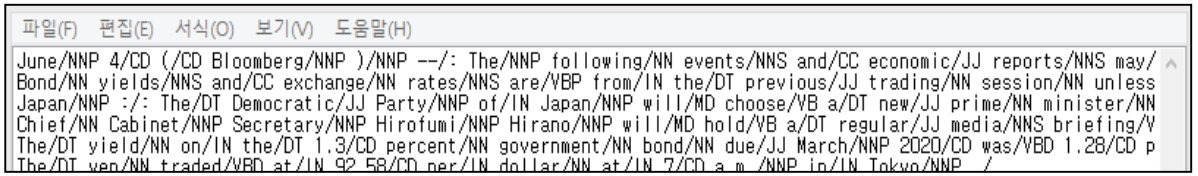
'U.S.', 'stocks', 'extended', 'a', 'four-day', 'rally', 'as', 'retailers', 'gained', '
after', 'reporting', 'claims', 'dropped', 'more', 'than', 'forecast'

텍스트 표현법 : Bag-of-words

PoS Tagging

- Assign each term to a PoS tag such as noun, verb adjective and adverb etc.





Form	Category	Tag
go	base	VB
goes	3rd singular present	VBZ
gone	past participle	VBN
going	gerund	VBG
went	simple past	VBD

Table 1. Morphology in PoS Tagsets

Ex)
- 'Hold', 'hold' → 'hold/VB'
- 'was' → 'was/VBD'

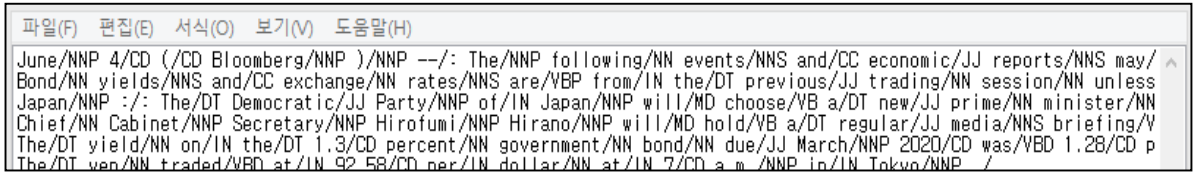
Format : Lower 'word' / Upper 'tag'

텍스트 표현법 : Bag-of-words

Term vector representation

- Make a numerical vector with words to represent documents

Document #1



Filtering

Adjective, preposition, adverb, conjunction, **noun**, numeral, particle, pronouns, **verb**, ...

'mexico/NNP', 'stock/NN', 'index/NN', 'demand/NN', 'Companies/NNS', 'retreat/NN', 'management/NNP', 'economists/NNS', 'supply/NNP', 'indicators/NNS', 'movil/NNP', 'enthusiasm/NN', 'latin/NNP', 'america/NNP', 'argentina/NNP', 'chile/NNP', 'colombia/NNP', 'peru/NNP'

Term vector

Term	stock/NN	demand/NN	economists/NNS	supply/NNP	...
Frequency	45	23	12	31	...

∴ **Document #1** = (45, 23, 12, 31, 0, 0, 0, 1, 0, 3, 2 ....)

텍스트 표현법 : Bag-of-words

Term vector representation

Doc1 : { "You are a very good boy." }

{ "You", "are", "a", "very", "good", "boy" }

Doc2 : { "You are a good girl too." }

{ "You", "are", "a", "good", "girl", "too" }

Doc3 : { "We are the world" }

{ "We", "are", "the", "world" }

- An unique set of all terms from news articles

{ "You", "are", "a", "very", "good", "boy", "girl", "too", "We", "the", "world" }

- Filtering 1 : Selecting PoS-tag terms

{ "You", "are", "a", "good", "boy", "girl", "We", "the", "world" }

- Filtering 2 : Minimum level of frequency

{ "You", "are", "a", "good", "boy", "girl", "the", "world" }

텍스트 표현법 : Bag-of-words

Term vector representation

Doc1 : { "You are a very good boy." }

{ "You", "are", "a", "very", "good", "boy" }

Doc2 : { "You are a good girl too." }

{ "You", "are", "a", "good", "girl", "too" }

Doc3 : { "We are the world" }

{ "We", "are", "the", "world" }

- **An unique set of all terms from news articles**
  { "You", "are", "a", "very", "good", "boy", "girl", "too", "We", "the", "world" }

- **Filtering 1 : Selecting PoS-tag terms**
  { "You", "are", "a", "good", "boy", "girl", "We", "the", "world" }

- **Filtering 2 : Minimum level of frequency**
  { "You", "are", "a", "good", "boy", "girl", "the", "world" }

텍스트 표현법 : Bag-of-words

Term vector representation

Doc1 : { "You are a very good boy." }

{ "You", "are", "a", "very", "good", "boy" }

Doc2 : { "You are a good girl too." }

{ "You", "are", "a", "good", "girl", "too" }

Doc3 : { "We are the world" }

{ "We", "are", "the", "world" }

- **An unique set of all terms from news articles**

{ "You", "are", "a", "**very**", "good", "boy", "girl", "**too**", "We", "the", "world" }

- Remove "adverb" terms

- **Filtering 1 : Selecting terms by PoS-tag**

{ "You", "are", "a", "good", "boy", "girl", "We", "the", "world" }

- **Filtering 2 : Minimum level of frequency**

{ "You", "are", "a", "good", "boy", "girl", "the", "world" }

텍스트 표현법 : Bag-of-words

Term vector representation

Doc1 : { "You are a very good boy." }

{ "You", "are", "a", "very", "good", "boy" }

Doc2 : { "You are a good girl too." }

{ "You", "are", "a", "good", "girl", "too" }

Doc3 : { "We are the world" }

{ "We", "are", "the", "world" }

- **An unique set of all terms from news articles**

{ "You", "are", "a", "very", "good", "boy", "girl", "too", "We", "the", "world" }

- **Filtering 1 : Selecting PoS-tag terms**

{ "You", "are", "a", "good", "boy", "girl", "**We**", "the", "world" }

- **Filtering 2 : Minimum level of frequency**

- Remove terms which "# ≤ 20"

{ "You", "are", "a", "good", "boy", "girl", "the", "world" }

텍스트 표현법 : Bag-of-words

Term vector representation

Doc1 : { "You are a very good boy." }

{ "You", "are", "a", "very", "good", "boy" }

Doc2 : { "You are a good girl too." }

{ "You", "are", "a", "good", "girl", "too" }

Doc3 : { "We are the world" }

{ "We", "are", "the", "world" }

- **An unique set of all terms from news articles**

{ "You", "are", "a", "very", "good", "boy", "girl", "too", "We", "the", "world" }

- **Filtering 1 : Selecting PoS-tag terms**

{ "You", "are", "a", "good", "boy", "girl", "**We**", "the", "world" }

- **Filtering 2 : Minimum level of frequency**

{ "You", "are", "a", "good", "boy", "girl", "the", "world" }

Term vector Feature set

텍스트 표현법 : Bag-of-words

Term vector representation

- Doc1 : { "You are a very good boy." }
- Doc2 : { "You are also a good girl too." }
- Doc3 : { "We are the world" }

X1 "You"	X2 "are"	X3 "a"	X4 "good"	X5 "boy"	X6 "girl"	X7 "the"	X8 "world"
1	1	1	1	1	0	0	0
1	1	1	1	0	1	0	0
0	1	0	0	0	0	1	1

Bow matrix

	X1	X2	X3	X4	X5	...	Xn
Doc1	1	0	0	1	2	...	0
Doc2	0	2	1	1	0	...	3
...	...	...	...	...	...	...	...
Doc m	4	0	0	1	3	...	0

분석 기법

- TF-IDF : 특정 단어가 문서 내에서 얼마나 중요한지를 나타내는 수치
- TF-IDF = Term frequency-inverse document frequency

- TF - 'Term Frequency'
  - 특정 단어가 특정 문서에 얼마나 나오는가를 나타냄

X1 "You"	X2 "are"	X3 "a"	X4 "good"	X5 "boy"	X6 "girl"	X7 "the"	X8 "world"
1	1	1	1	1	0	0	0
1	1	1	1	0	1	0	0
0	1	0	0	0	0	1	1

- IDF - 'Inverse Document Frequency'
  - 특정 단어를 포함한 문서가 얼마나 많은가를 나타냄

$$= log \frac{N-n}{n}$$

- N = 전체 문서의 개수
- n = 특정 단어가 포함된 문서의 개수

- TF- IDF = TF X IDF
  - 중요한 단어는 특정 문서에서는 많이 나와야 하지만 다른 문서에는 많이 포함되지 않아야 함
  - TF - IDF 값이 크면 중요한 단어로 생각할 수 있음

분석 기법

- TF-IDF : 특정 단어가 문서 내에서 얼마나 중요한지를 나타내는 수치
- TF-IDF = Term frequency-inverse document frequency

- 예시 – '북한 관련 뉴스기사'

TF가 높고 IDF가 낮은 경우

- '**북한**' 이라는 단어
  - ✓ 특정 문서에서 발견되는 Frequency가 높음(TF ↑)
  - ✓ 북한 관련 문서 corpus에서 많은 문서가 '북한'이라는 단어를 포함하고 있음 (IDF ↓)
  - ✓ TF X IDF 값이 크지 않음

TF가 높고 IDF가 높은 경우

- ' **5차 핵실험** ' 이라는 단어
  - ✓ 최근 문서에서 발견되는 Frequency가 높음(TF ↑)
  - ✓ 북한 관련 문서 corpus에서 일부 문서만 '5차 핵실험' 이라는 단어를 포함하고 있음 (IDF ↑)
  - ✓ TF X IDF 값이 큼

	TF	n	IDF	TF X IDF
북한	**30**	2,000,000	0.176	5.282
5차 핵실험	12	100,000	1.770	**21.25**

분석 기법

- TF-IDF : 특정 단어가 문서 내에서 얼마나 중요한지를 나타내는 수치
- TF-IDF = Term frequency-inverse document frequency

- 예시 : TF-IDF 변형을 이용한 전자 뉴스에서의 키워드 추출 방법
  - 각각의 TF와 IDF의 가중치 계산을 변형하여 중요한 정도를 달리 할 수 있다.

순위	기존 TF-IDF	기존 TF-IDF의 수정		
		BTF	NTF1	NTF2
1	일부	이명박 후보	이명박 후보	서울
2	박수	후보	후보	무단전재
3	협력	기자	기자	한국언론뉴스허브
4	한나라당	에리카 김	에리카 김	뉴시스통신사
5	서울	무단전재	무단 전재	재배포 금지
6	대표	광고	광고	모바일연합뉴스
7	오후	서울	서울	재배포금지
8	조성	재배포 금지	재배포 금지	저작권자연합뉴스
9	통합	검찰	검찰	오전
10	협상	이명박	이명박	오후

분석 기법 – 실습 예시

- 실제 연합 뉴스 기사

"아시아 대중음악 축제인 아시아송페스티벌이 10월 7일부터 9일까지 부산아시아드 경기장에서 열린다. 부산시는 10월 7일과 8일 부산아시아드 보조경기장에서 2016 아시아송페스티벌 전야행사를 하고, 10월 9일 부산아시아드 주경기장에서 본행사를 한다고 밝혔다. 2014년부터 시작한 아시아송페스티벌은 대중음악을 매개체로 아시아의 화합과 문화교류 활성화를 도모한다. 전체기사 본문배너 올해 축제는 대한민국을 대표하는 K팝 아티스트와 중국, 일본, 베트남, 싱가포르, 인도네시아, 필리핀 등 아시아 주요 국가들의 최정상급 아티스트가 대거 출연한다. 10월 9일 열리는 본행사에는 엑소, NCT 127, 세븐틴, 트와이스, AOMG 사이먼 도미닉 등이 참여를 확정해 아시아 한류팬들의 관심을 불러일으킬 것으로 기대된다. 행사를… "

<TF기반 워드클라우드>

<TF-IDF기반 워드클라우드>

분석 기법

- 이외에도 Bigram, Trigram(연어, collocation)을 고려한 모델도 있음
  - 단어를 하나씩만 보지 말고 두 개, 세 개씩 보자

  Ex) "남녀 관계는 정말 모르겠다.", "대북 관계에서 큰 합의점을 찾지 못했다."

  → 일반적인 word vector로 표시하면 두 문장 모두 '관계'라는 단어가 포함됨

  → "남녀 관계 " 와 "대북 관계"는 전혀 다른 표현

  → 연어(collocation)을 고려한다면?

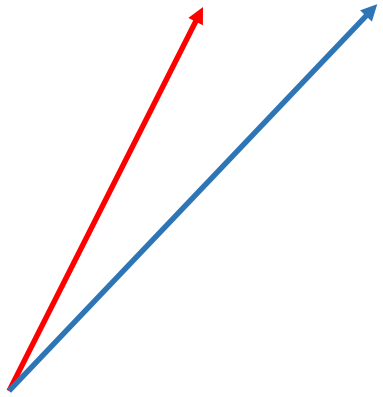
  → 형태소 분석의 한계점을 보완해 줄 수 있다.

분석 기법 – 유사도 문서 검색

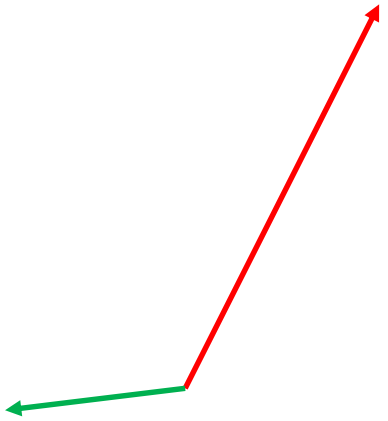
- **문서 비교 – 유사도 계산**
  - 두 벡터의 유사도(similarity)를 계산하는 방법은 여러가지 있다.
  - 문서가 Bag-of-word를 통해 벡터로 표현 가능하기 때문에 문서의 유사도 또한 계산할 수 있다.

코사인 유사도(Cosine Similarity)

- 벡터의 내적 값을 이용하여 코사인 값을 계산하는 것으로 두 벡터가 유사할 수록 값이 큼(0~1)
- $\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$



< 코사인 값이 큼/ 두 벡터 유사>



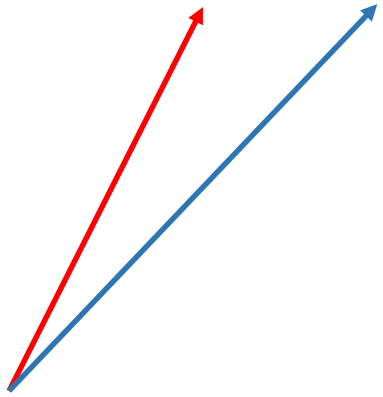
< 코사인 값이 작음/ 두 벡터 유사 X>

분석 기법 – 유사도 문서 검색

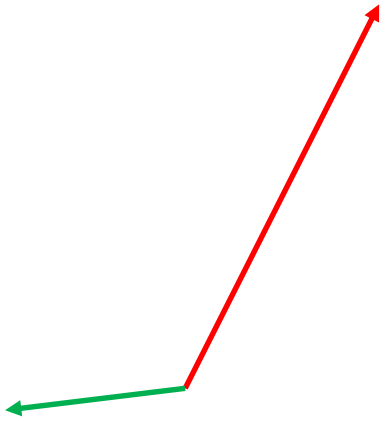
- **문서 비교 – 유사도 계산**
  - 두 벡터의 유사도(similarity)를 계산하는 방법은 여러가지 있다.
  - 문서가 Bag-of-word를 통해 벡터로 표현 가능하기 때문에 문서의 유사도 또한 계산할 수 있다.

코사인 유사도(Cosine Similarity)

- 벡터의 내적 값을 이용하여 코사인 값을 계산하는 것으로 두 벡터가 유사할 수록 값이 큼(0~1)
- $\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$



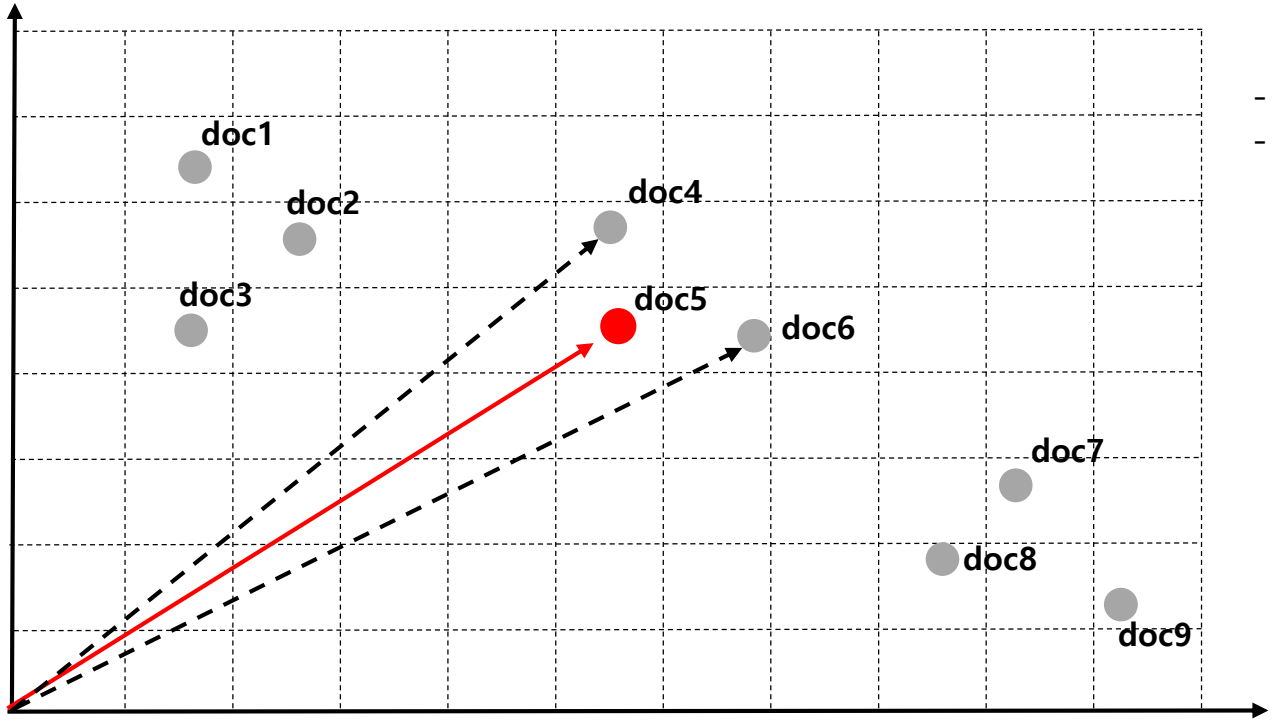
< 코사인 값이 큼/ 두 벡터 유사>



< 코사인 값이 작음/ 두 벡터 유사 X>

분석 기법 – 유사도 문서 검색

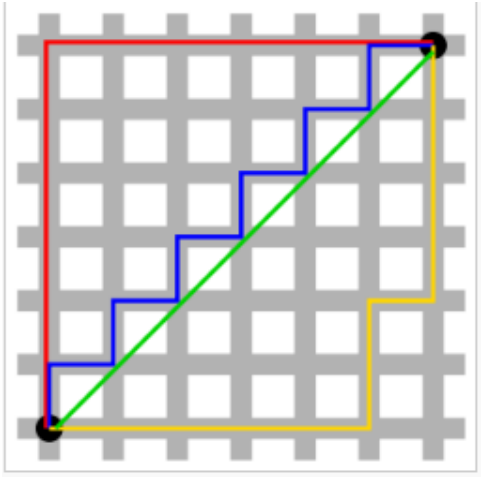
- **문서 비교 – 유사도 계산**
  - 두 벡터의 유사도(similarity)를 계산하는 방법은 여러가지 있다.
  - 문서가 Bag-of-word를 통해 벡터로 표현 가능하기 때문에 문서의 유사도 또한 계산할 수 있다.



- doc5는 doc4와 유사!
- doc5는 doc6와 유사!

분석 기법 – 유사도 문서 검색

- **문서 비교 - 유사도계산**
  - 문서의 유사도는 코사인을 이용한 방법 이외에 다양한 특성을 가진 값들이 존재
  - 유클리디안 거리
  - 해밍 거리
  - 맨하탄 거리 등등

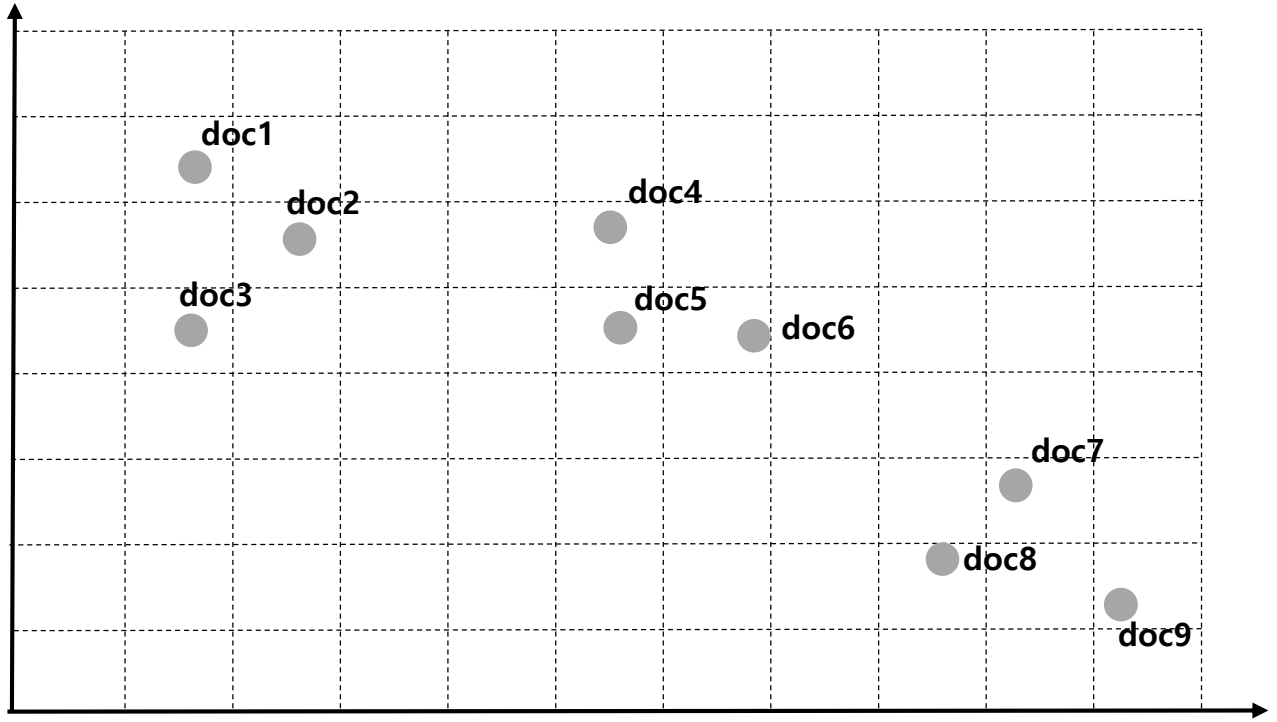


< 맨하탄 거리 예시(빨,파,노)>

→ 주어진 연합뉴스 "한류" 기사를 이용하여 유사도를 계산, 유사 문서를 찾아보자!

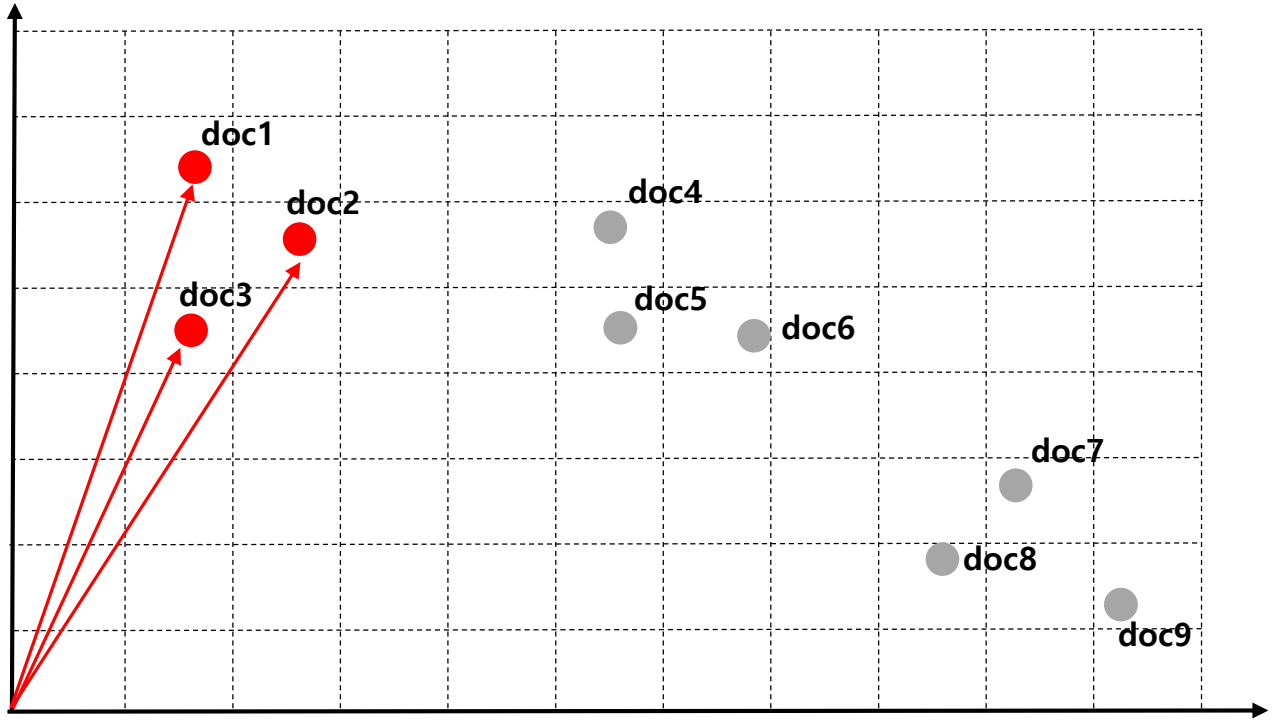
분석 기법 – 유사도 문서 군집화

- **문서 유사도를 이용한 군집 분석**
  - 두 벡터의 유사도(similarity)를 계산할 수 있으면 데이터들이 모아져 있는 군집(cluster)분석이 가능
  - 비슷한 문서 집단을 식별할 수 있다.



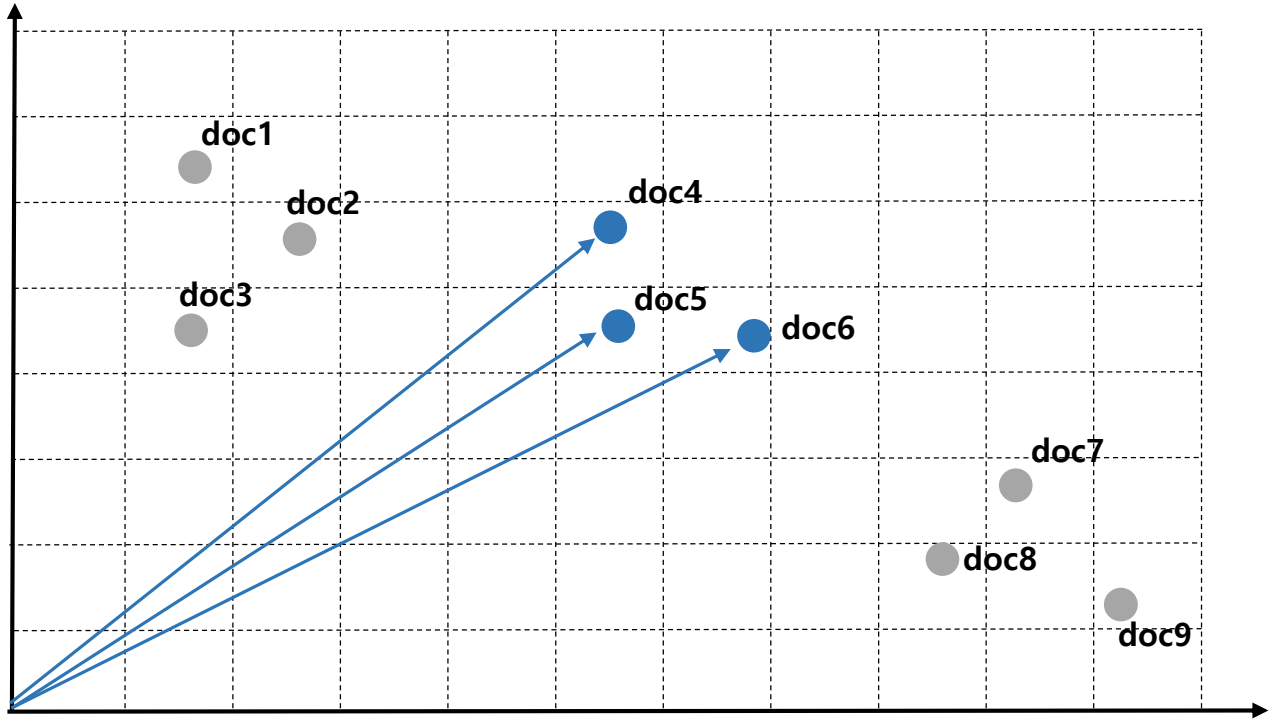
분석 기법 – 유사도 문서 군집화

- **문서 유사도를 이용한 군집 분석**
  - 두 벡터의 유사도(similarity)를 계산할 수 있으면 데이터들이 모아져 있는 군집(cluster)분석이 가능
  - 비슷한 문서 집단을 식별할 수 있다.



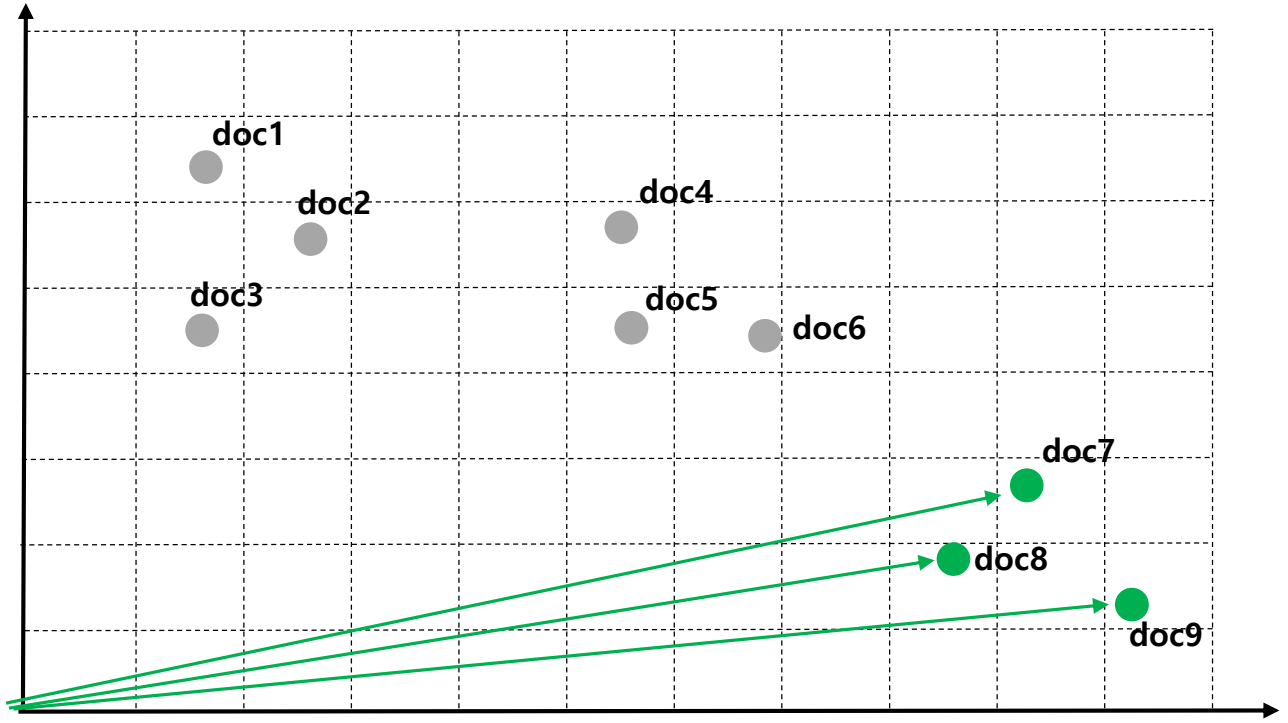
분석 기법 – 유사도 문서 군집화

- **문서 유사도를 이용한 군집 분석**
  - 두 벡터의 유사도(similarity)를 계산할 수 있으면 데이터들이 모아져 있는 군집(cluster)분석이 가능
  - 비슷한 문서 집단을 식별할 수 있다.



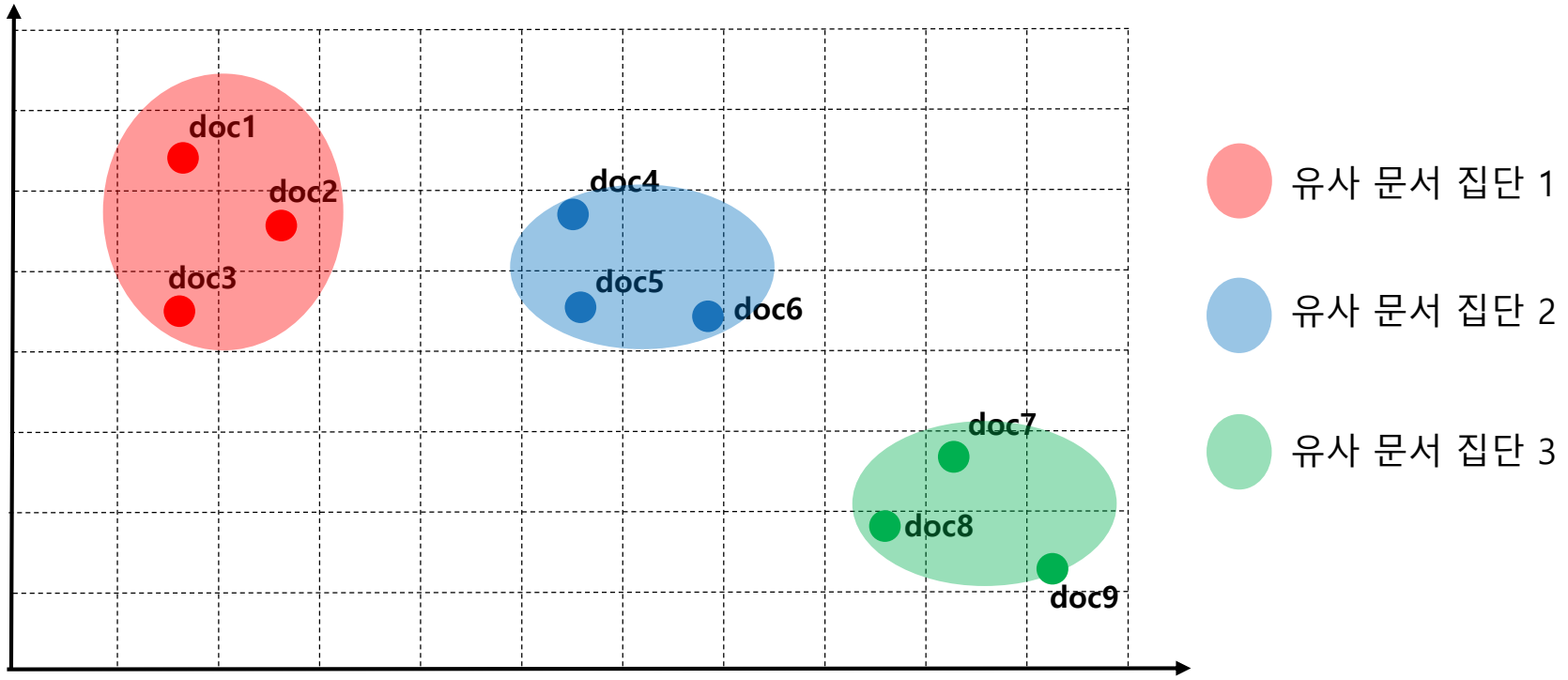
분석 기법 – 유사도 문서 군집화

- **문서 유사도를 이용한 군집 분석**
  - 두 벡터의 유사도(similarity)를 계산할 수 있으면 데이터들이 모아져 있는 군집(cluster)분석이 가능
  - 비슷한 문서 집단을 식별할 수 있다.



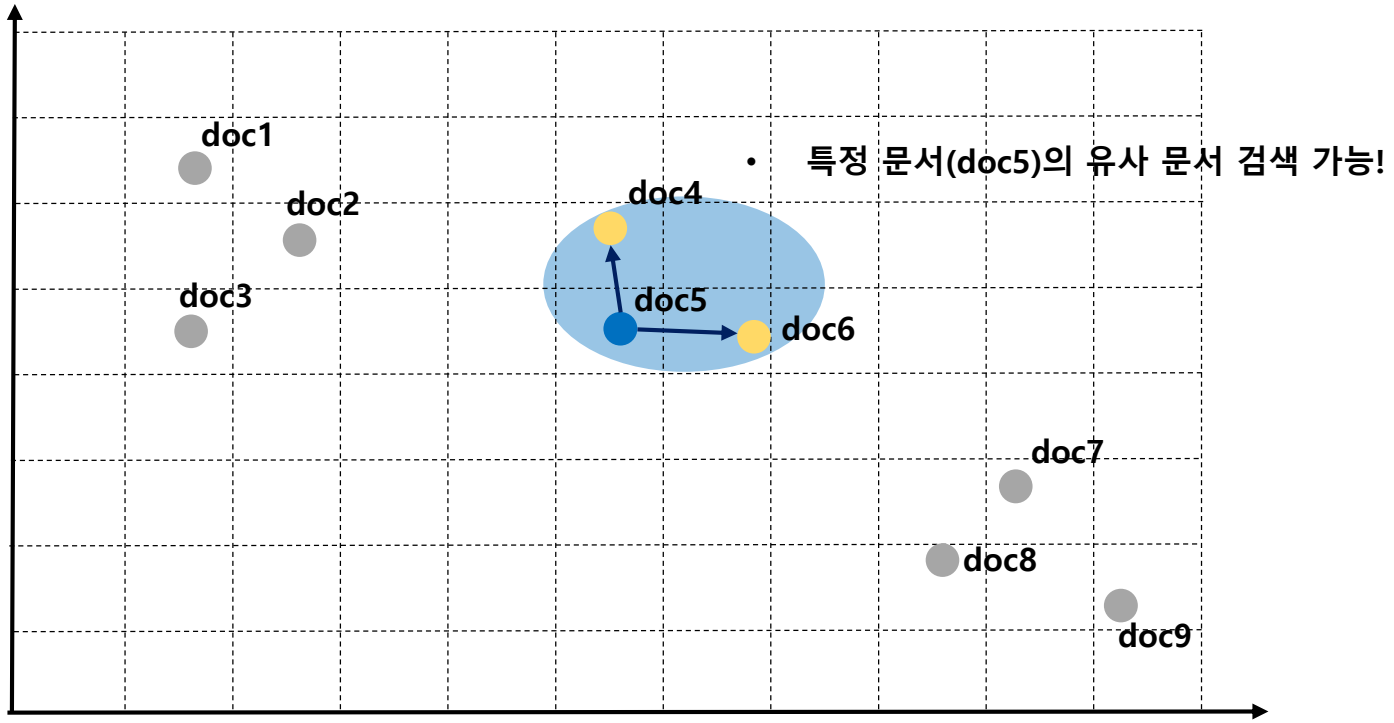
분석 기법 – 유사도 문서 군집화

- **문서 유사도를 이용한 군집 분석**
  - 두 벡터의 유사도(similarity)를 계산할 수 있으면 데이터들이 모아져 있는 군집(cluster)분석이 가능
  - 비슷한 문서 집단을 식별할 수 있다.



분석 기법 – 유사도 문서 군집화

- **문서 유사도를 이용한 군집 분석**
  - 두 벡터의 유사도(similarity)를 계산할 수 있으면 데이터들이 모아져 있는 군집(cluster)분석이 가능
  - 비슷한 문서 집단을 식별할 수 있다.



분석 기법 – 실습 예시(2)

- 북한의 자본주의 문화에 대한 단속과 통제에 관련된 뉴스

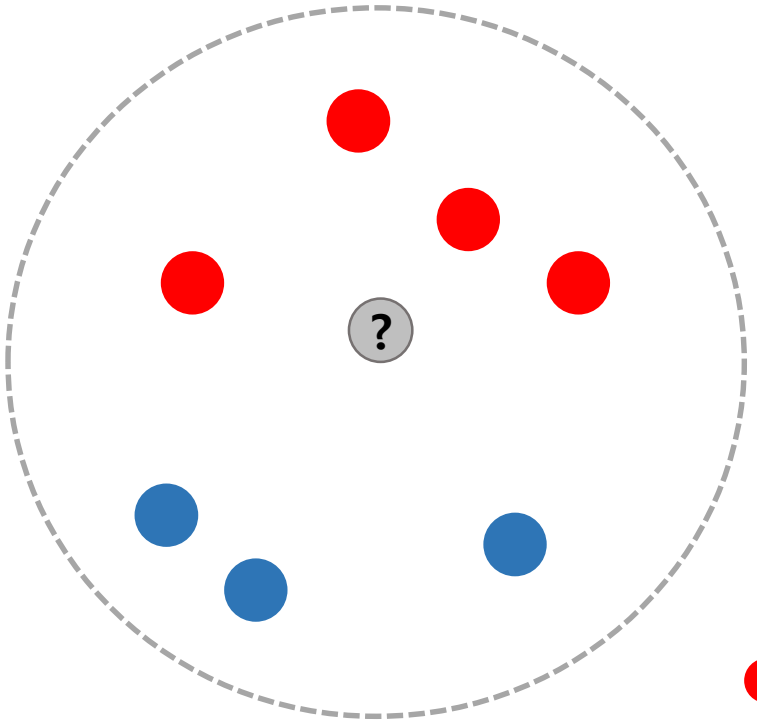
> 북한당국이 주민을 상대로 자본주의 문화에 대한 단속과 통제를 강화하고 있지만, '태양의 후예'와 같은 한국 드라마를 몰래 시청하는 북한 주민들이 많다고 북한 전문매체인 데일리 NK가 4일 보도했다. 평안남도 소식통은 이 매체에 "최근 젊은 청년들 속에서 \'태양의 후예\'라는 한국 드라마가 인기를 끌면서 날이 새는 줄도 모르고 시청하고 있다"면서 "이 드라마에 대한 소문이 퍼지자 어른·아이 할 것 없이 너도나도 (드라마를) 보기 위해 애쓰고 있는 상황"이라고 전했다

- 유사 문서

> 배정호 민주평화통일자문회의 사무처장은 27일 "북한의 전통적인 우방국이자 여전히 정치, 경제, 군사적 교류를 이어가는 베트남의 대북제재 이행은 북한의 태도변화에 큰 전환점이 될 수 있다"며 베트남의 적극적인 제재 동참을 촉구했다. 배 사무처장은 이날 베트남 하노이 롯데호텔에서 민주평통 주최로 열린 \'2016 한·베트남 평화통일포럼\'에서 기조연설을 통해 "유엔 제재가 빈틈없이 이행되도록 한국과 베트남이 적극적으로 협력하는 것이 무엇보다 필요하다"고 강조했다

분석 기법 – 분류

1. k-NN(k nearest neighbors)



< k-NN 예시 그림>

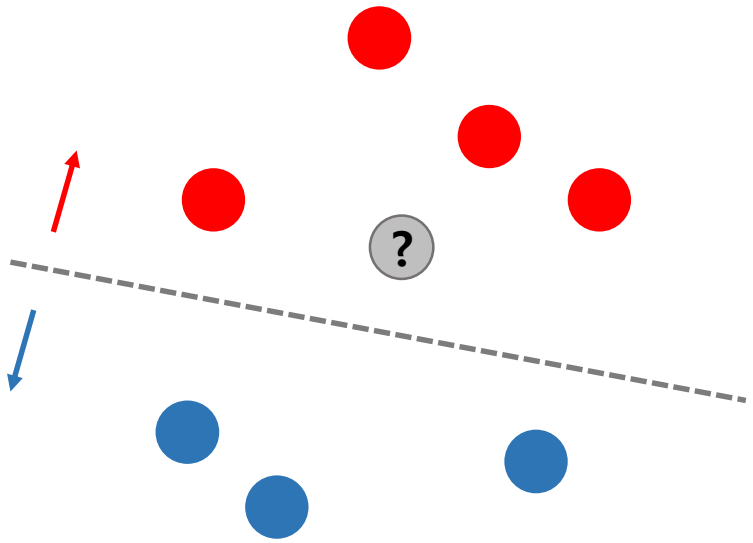
- K nearest neighbors(k-NN)은 주변에 가까운 이웃 데이터를 살펴보고 가장 많은 label의 값을 할당하는 알고리즘
- 왼쪽처럼 K=7인 knn의 예시를 살펴보면 label이 없는 뉴스 기사(하나의 점)의 가장 가까운 7개의 뉴스 기사의 label은 그림과 같다.
- KNN으로 해당 뉴스 기사가 어떤 뉴스 기사인지 분류하면 주변에 'Food' 관련 기사가 더 많기 때문에 해당 뉴스 기사도 'Food' 관련 뉴스 기사로 분류 할 수 있다!

: Food 관련 기사

: Energy 관련 기사

분석 기법 – 분류

2. Logistic regression



< Logistic regression 예시 그림>

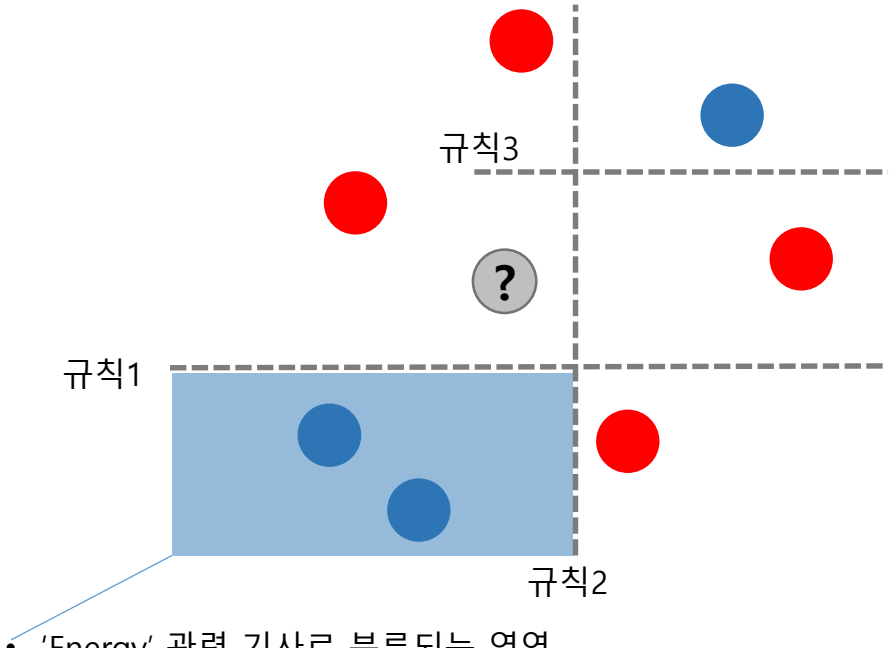
- Logistic regression은 두 label를 가장 잘 구분해 줄 수 있는 선형 구분 선을 찾아 주는 알고리즘
- Logistic regression은 knn과는 달리 결과 값이 0과 1 사이의 소수로 나타나 특정 class에 속할 확률로 사용할 수 있다.
- Label를 모르는 뉴스기사는 'Food'로 분류될 확률 값이 더 높기 때문에 'Food' 관련 뉴스 기사로 분류 할 수 있다!

: Food 관련 기사

: Energy 관련 기사

분석 기법 – 분류

3. Decision tree



< Logistic regression 예시 그림>

- Decision Tree는 두 label를 구분해 줄 수 있는 규칙을 찾아 주는 알고리즘
- Label를 모르는 뉴스기사는 'Food'로 분류된 구역에 속해 있기 때문에 'Food' 관련 뉴스 기사로 분류 할 수 있다!

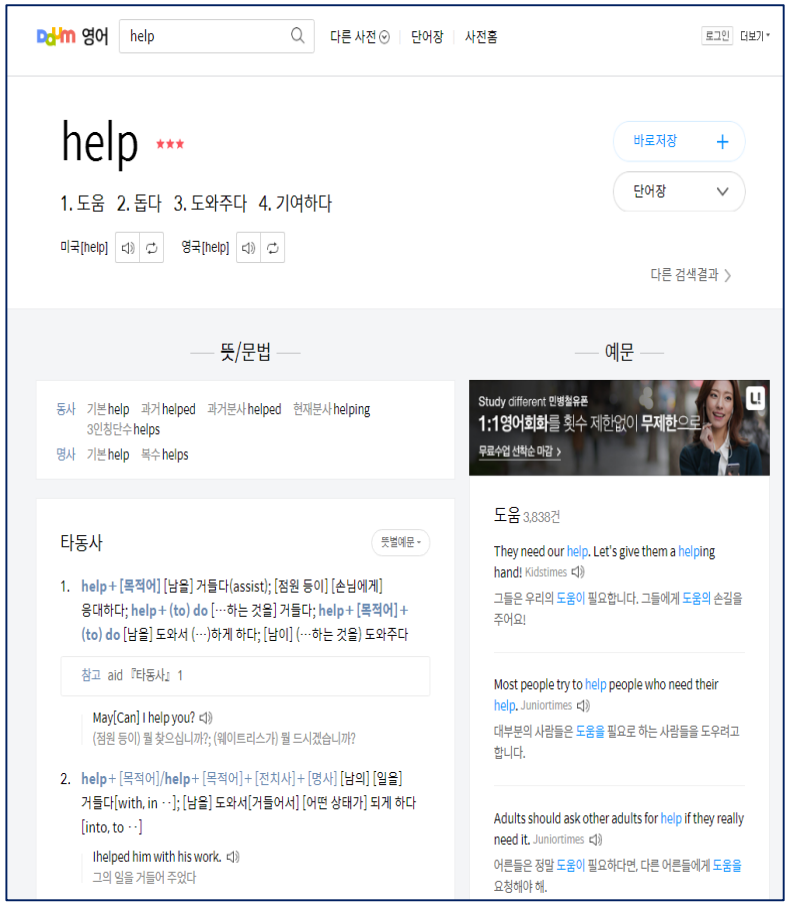
- **예시**
- **규칙 1** : 석유 연료와 관련된 단어가 나오면 Energy 관련 기사
- **규칙 2** : 닭과 관련된 단어가 나오면 Food 관련 기사

: Food 관련 기사

: Energy 관련 기사

+Data scraping

- Data scraping from website – 다음 영어사전



< Daum 영어사전>

```python
from bs4 import BeautifulSoup
import urllib.request

import numpy as np

word_list = []
meaning_list = []
pronounce_list = []


word_index = np.random.random_integers(1,100000,100)
start = "http://dic.daum.net/word/view.do?wordid=ekw"
end = "&q="

for index in word_index:
    num = 1000000000+index
    url = ""
    url = start + str(num)[1:] + end
    doc = ""

    with urllib.request.urlopen(url) as url:
        doc = url.read()


    soup = BeautifulSoup(doc, "html.parser")

    word = soup.find_all("span", class_="txt_cleanword")
    meaning = soup.find_all("span", class_="txt_mean")
    pro = soup.find_all("span", class_="txt_pronounce")


    word_list.append(word)
    meaning_list.append(meaning)
    pronounce_list.append(pro)

    #print(word[0].text+' copied')
```

```
C:₩Users₩Hooney₩Anaconda3₩lib₩site-packages₩ipykernel₩__ma
(1, 100000 + 1) instead
```

< Scraping code >

```
1 . dormitive

        (1). 잠이 오게 하는
        (2). 최면성의
        (3). 최면성
발음
미국식 =  dɔ́ːrmitiv
영국식 =  dɔ́ːrmitiv


2 . bruchid

        (1). 콩바구미
        (2). 콩바구밋과의
발음
미국식 =  brúːkid, -kəd
영국식 =  brúːkid, -kəd


3 . chuff chums

        (1). 동성애자들


4 . Apurimac

        (1). 아푸리막 강
발음
미국식 =  aːpríːmaːk
영국식 =  aːpríːmaːk
```

< Result >


thank you
danke
謝謝
gracias
merci
obrigado
dziękuję
bedankt
go raibh maith agat
teşekkür ederim
ngiyabonga
спасибо
Баярлалаа
рахмат
dank je
tapadh leat
mochchakkeram
sukriya
kop khun krap
terima kasih
arigatô
dakujem
takk
grazie
tack
mauruuru
hvala
sagolun
obrigada
asante
manana
tenki
murakoze
chokrano
mamnun
дякую
хвала
merci
мерси
감사합니다
ευχαριστώ
xiexie
dhanyavadagalu
shukriya
trugarez
tanemirt
rahmet
najis tuke
didi madloba
kam sah hamnida
তোমাকে ধন্যবাদ
sobodi
dėkuji
mèsi
ありがとう
chnorakaloutioun
gràcies
gratias ago
sulpáy
djiere dieuf
tau
akun
dankon
ačiū
köszönöm
bayarlalaa
graçie
nanni
nandri
dhanyavad
kiitos
dankie
faafetai lava
спасибі
vinaka
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
welalin
barka
kia ora
mersi
spas
enkosi