[12주차] 비정형 텍스트 분석: 토픽, 군집, 분류와 감성분석

1. 토픽 모델링

토픽(Topic) 모델링이란?

- 텍스트에 숨겨져 있는 주제들을 찾아내기 위한 통계 추론에 기반한 분석기법
- 개별 문서는 다수의 주제, 혹은 토픽을 다룰 수 있다는 점을 가정
- 수집된 텍스트를 토픽의 확률적 혼합체로 간주하고,
  각 토픽을 추출된 키워드들의 분포로 나타냄으로써
  텍스트 내의 구조를 파악
- 토픽분석의 시초는 잠재 의미 분석(LSA: Latent Semantic Analysis)이며, 최근에는 잠재 디리클레 할당(LDA: Latent Dirichlet Allocation) 모델 기법이 대표적으로 사용되고 있음

토픽(Topic) 모델링이란?

- 텍스트에 숨겨져 있는 주제들을 찾아내기 위한 통계 추론에 기반한 분석 기법으로 의미론적 표현법의 새로운 패러다임으로 주목 받고 있음
- 토픽분석의 시초는 잠재 의미 분석(LSA: Latent Semantic Analysis)이며, 최근 잠재 디리클레 할당(LDA: Latent Dirichlet Allocation) 모델 기법이 대표적으로 사용되고 있음

잠재 의미 분석(Latent Semantic Analysis)

- 대량의 텍스트 문서에서 발생하는 단어들 간의 연관관계를 분석함으로써 잠재적인 의미 구조를 도출(Deerwester et al.,1990)
- 토픽 추출을 위해 수학적인 접근 방식인 행렬 인수분해 기법 중 특이값 분해(SVD, Singular Value Decomposition)를 활용하여 텍스트 문서 집합을 내용의 유사도에 따라 여러 개의 소집단으로 분할
- 텍스트 문서 집합 내에 숨겨진 주제를 도출하는 기법으로 문서 집합 내에서 연관성, 즉 동시 출현 빈도가 높은 단어들을 기준으로 유사한 문서를 추출

잠재 의미 분석(Latent Semantic Analysis)

- 잠재 의미 분석의 원리는 고차원의 단어 문서 행렬을 축소시켜 토픽을 추출하는데 적용시키는 것임.
- 잠재 의미 분석은 생성된 단어 문서 행렬(X: txd 직사각 행렬) 을 SVD을 이용하여 다음과 같이 3개의 독립적인 행렬로 분해를 수행함.

$$X = T_0 S_0 D_0'$$

T_0: X의 좌특이벡터로 구성된 t x m 직교 행렬
S_0 : 특이치로 구성된 m x m 대각행렬
D_0': X의 우특이벡터로 구성된 m x d 직교 행렬

→ 여기서 m(≤ min(t, d))은 단어 문서 행렬 X의 계수(rank)를 의미

- SVD에 의하여 분해된 행렬의 계수에 따라서 단어 문서 행렬의 차원을 축소하여 토픽이 추출됨.

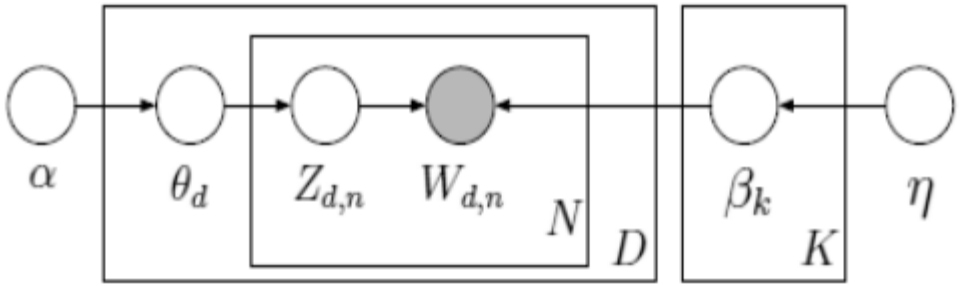
잠재 디리클레 할당(LDA) 모델

- 문서가 생성될 확률인 사후분포에 기반한 변수(hidden variable)를 추론하여 텍스트 내의 숨겨져 있는 주제를 찾아내는 방식 (Blei *et al.,* 2003)
- 문서를 작성하기 위해 각 문서에 어떤 주제들을 포함시킬 것인지, 또 그에 따라 어떤 단어들을 어떤 주제에서 선택하여 배치할 것인지를 토픽의 분포와 각 토픽 별로 단어가 생성될 확률인 사전분포에 기반한 변수 (observed variable)로 모델링
- 결과적으로 전체 텍스트 문서 집합의 주제(토픽)들, 각 텍스트 문서별 주제 (토픽) 비율, 각 단어들이 각 주제(토픽)에 포함될 확률을 도출

* 디리클레(Dirichlet): 확률분포 명칭

잠재 디리클레 할당(LDA) 모델

■ 잠재 디리클레 할당(LDA) 구조

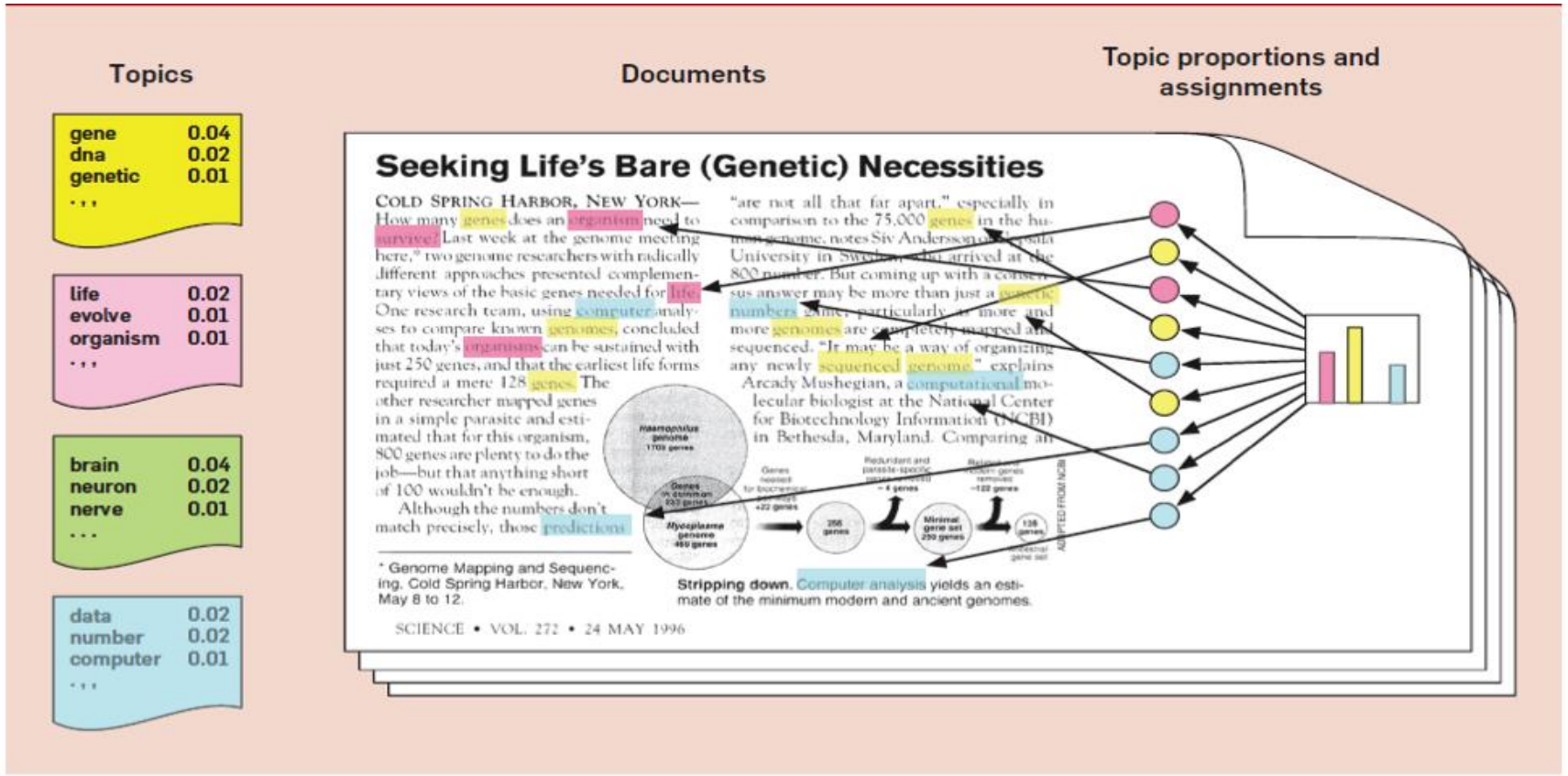


W: 해당 문서의 단어 인덱스
K: 추출할 토픽의 개수
(처음에 주어지는 고정값)
D: 문서의 개수
N: 문서 집합 전체 D의 길이
(모든 문서에 나타나는 단어의 총 개수)

출처: Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM 55(4): 77–84.

Hidden Variable	θ: 문서별 토픽 비율 Z: 해당 문서의 단어별 토픽 인덱스 β: 토픽별 단어의 생성확률
Hyper-parameter of Dirichlet Distribution	α: 각 문서가 어떠한 토픽 비율로 구성될지를 나타내는 θ값을 결정하는 파라미터 η: 각 단어가 어떠한 토픽들의 비율로 구성될지를 나타내는 β값을 결정하는 파라미터
Observed Variable	문서, 단어

잠재 디리클레 할당(LDA) 모델



출처: Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM 55(4): 77–84.

토픽 모델링 결과

- 하나의 토픽은 여러 개의 키워드 집합으로 표현되며, 키워드 집합을 통해 각 토픽에 대한 의미 파악이 가능함.
- 각 토픽에 대한 명명(Naming)은 연구자가 직접 판단해야 함.

범주	토픽 ID	문서 임계치	용어 임계치	토픽	용어 수	문서 수
다중	1	0.249	0.150	수립,전략,항공,협상,안전	9	17
다중	2	0.217	0.148	항공기,안전,대책,규제,소형	5	19
다중	3	0.245	0.147	개선,제도,항공,분야,자격증	7	14
다중	4	0.235	0.149	체계,관리,구축,운영,항공기	9	15
다중	5	0.216	0.151	개발,안전,국가,항공,관리	8	14
다중	6	0.203	0.147	수요,지역,조사,+지역#항공#수요,제주	6	10
다중	7	0.210	0.148	교통,평가,서비스,+교통#서비스#평가,분야	7	11
다중	8	0.223	0.150	산업,도입,항공,구축,지원	8	15
다중	9	0.197	0.150	항공사,시장,지역,대응,개발	9	9
다중	10	0.166	0.151	기준,기술,정비,항공업,주파수	7	14

* 문서 임계치: 문서가 해당 토픽에 포함되어야 하는 최소 토픽 membership
* 용어 임계치: 단어가 해당 토픽에 대한 단어로 사용되어야 하는 최소 토픽 가중치
* 토픽: 토픽을 기술하는 단어들

토픽 모델링의 활용

- 사회 문제를 다루고 있는 대용량 뉴스기사로부터 LDA 기반의 토픽 분석을 적용하여 사회적 이슈에 관한 키워드를 도출하는 시스템을 제안 (Jeong et al.,2013)
- 트위터(Twitter) 데이터를 대상으로 LDA 기반의 토픽분석을 적용하여 SNS 상에서의 주요 이슈를 추출하는 트위터 이슈 트래킹 시스템을 제안 (Bae et al., 2014)
- 국토교통, 안전, 정보통신기술, 건설과 철강산업 등의 분야에도 토픽 모델링을 적용하여 미래 핵심 기술과 이슈를 발견하고 트렌드를 분석하여 경제적•사회적 부가가치를 창출하고, 국가 전략 및 정책 수립 시 반영하는데 활용

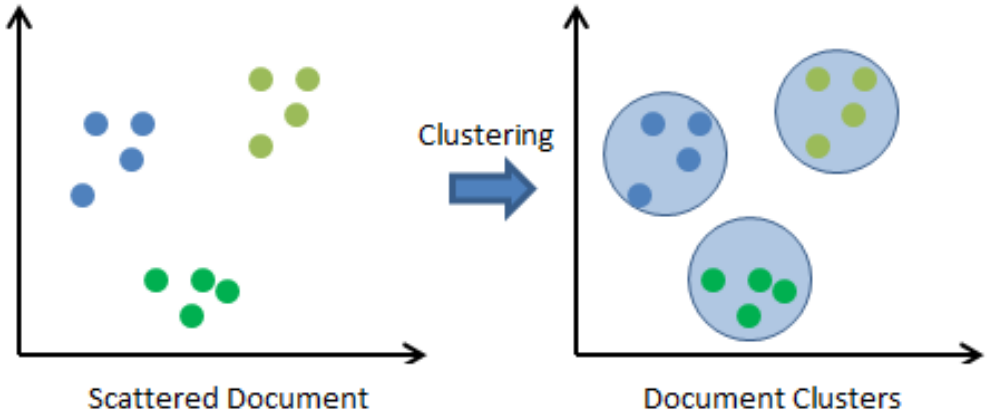
2. 텍스트 클러스터링

텍스트 클러스터링이란?

- 텍스트들 간의 유사도(similarity)를 계산하여 유사한 텍스트들을 몇 개의 집단으로 묶어주는 과정을 의미
- 텍스트 데이터 형태의 분석대상 개체들에 군집분석(cluster analysis)를 적용하는 것
- 대상 개체는 문서, 단락, 문장, 단어 등이 될 수 있음
- 가장 일반적인 개체는 문서

텍스트 클러스터링이란?...

- 대상을 유사한 속성을 지닌 몇 개의 집단으로 군집화 한 후, 각 집단의 성격을 파악하고자 하는 경우 사용함.



문서간 유사도는 어떻게 측정할까?

- **Term-Document Matrix로 측정**

Unstructured Document

Doc 1 : deposit the cash and check in the bank!!
Doc 2 : the river boat is on the bank
Doc 3 : borrow based on credit
Doc 4 : river boat floats up the river
Doc 5 : boat is by the dock near the bank

Pre-processing

Structured Matrix

Term-by-Document Frequency Matrix

	d1	d2	d3	d4	d5
cash	1	0	0	0	0
check	1	0	0	0	0
bank	1	1	0	0	1
river	0	1	0	2	0
credit	0	0	1	0	0

문서간 유사도는 어떻게 측정할까?

- 다양한 거리 척도가 사용됨
  - 레코드간 거리의 측정
    - √ 유클리드 거리 (Euclidean distance)
    - √ 마할라노비스 거리 (Mahalanobis distance)
    - √ 맨하탄 거리(Manhattan distance)
    - √ 최대좌표거리 (Maximum coordinate distance) 등
  - 군집간 거리의 측정
    - √ 최단거리
    - √ 최대거리
    - √ 평균거리
    - √ 중심거리 등

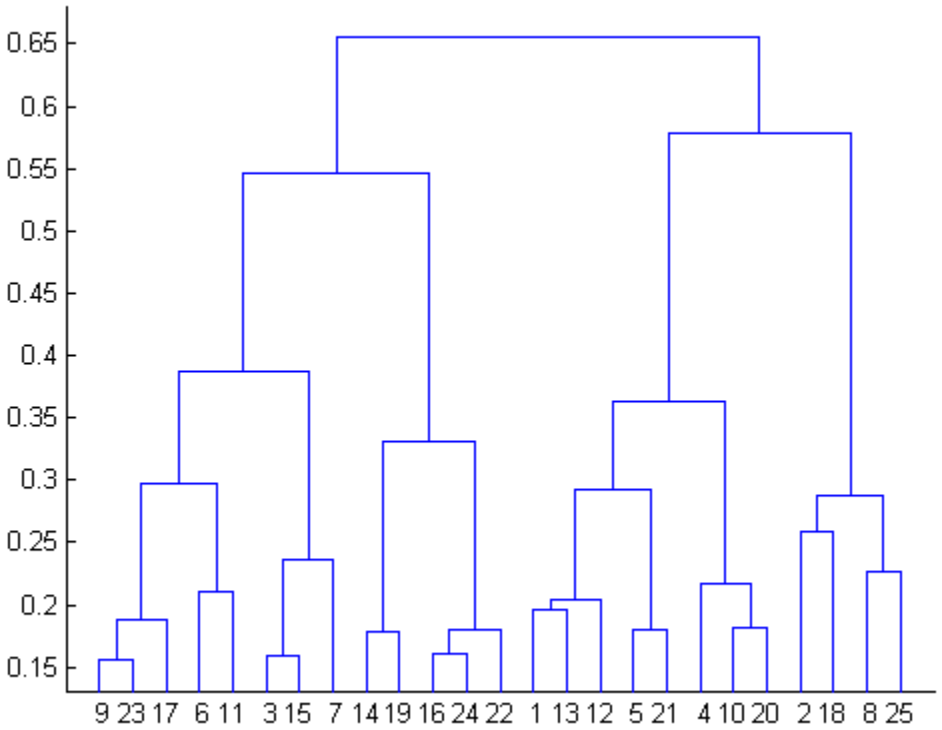
텍스트 클러스터링 방법

- 계층적 클러스터링(hierarchical clustering)
- 비계층적 클러스터링(non-hierarchical clustering)

계층적 클러스터링

- 개체들 간의 거리에 의하여 가장 유사한 개체들 부터 결합하여 나무모양의 계층적 구조를 형성해 가는 방법
- 거리계산 방법에 따라
  - 최단 연결법(Single Linkage Method)
  - 최장 연결법(Complete Linkage Method)
  - 평균 연결법(Average Linkage Method)
  - Ward의 방법(Ward Linkage Method) 등

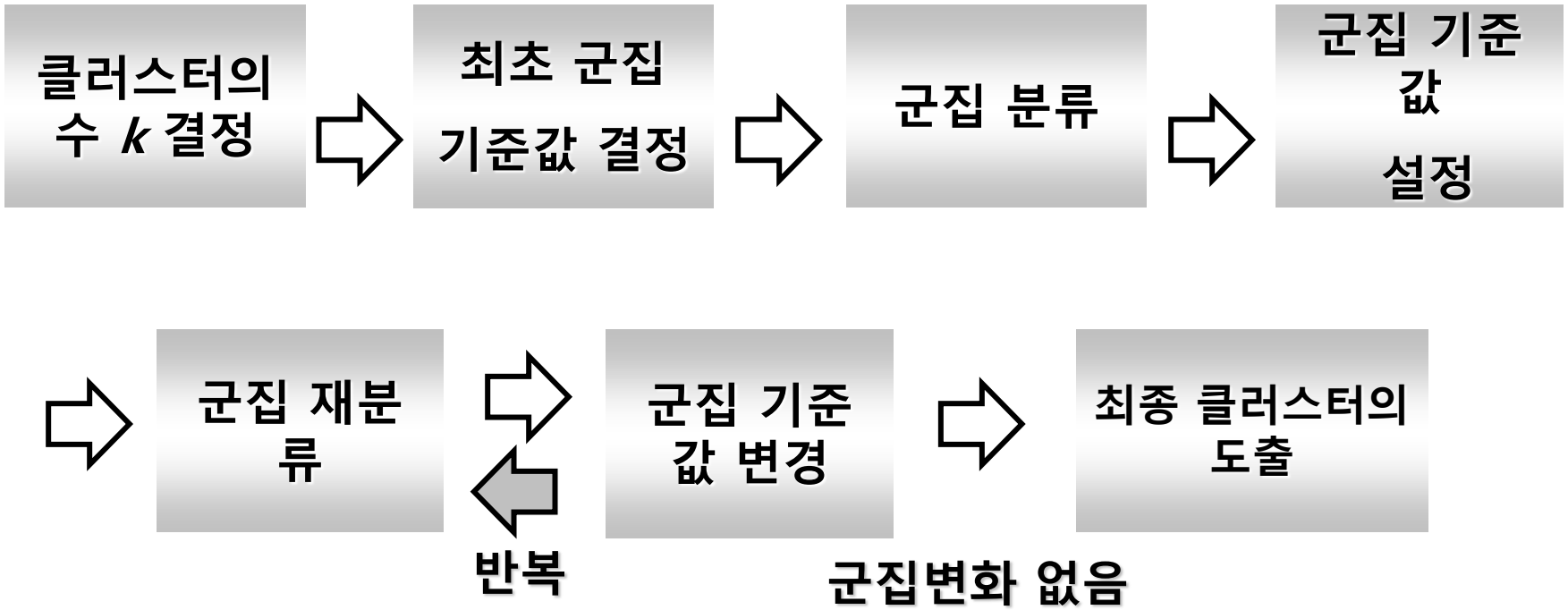
계층적 군집분석과 덴드로그램(Dendrogram)



비계층적 클러스터링

- 비계층적 클러스터링(non-hierarchical clustering)
  - 군집의 수를 미리 정한 상태에서 설정된 군집의 중심에서 가장 가까운 개체를 하나씩 포함해 가는 방법
  - 대표적인 방법
    - √ K-평균 클러스터링(K-means clustering)

K-평균 클러스터링


클러스터의 수 k 결정
최초 군집 기준값 결정
군집 분류
군집 기준 값 설정
군집 재분류
군집 기준 값 변경
최종 클러스터의 도출
반복
군집변화 없음

K-평균 클러스터링의 예시

	d1	d2	d3	d4	d5
Bank(X1)	2	4	6	8	10
River(X2)	4	4	6	6	9

개체	X_1	X_2
1	2	4
2	4	4
3	6	6
4	8	6
5	10	9

개체	Ⅰ	Ⅱ
1	1	45
2	1	25
3	13	5
4	29	1
5	74	8

- ✓ Step 1: 임의로 개체 1, 2를 클러스터 Ⅰ, 개체 3, 4, 5를 클러스터 Ⅱ로 분류함.
- ✓ Step 2: 각 개체와 클러스터 평균과 제곱합을 계산하면 표. 1.2와 같음.
- ✓ Step 3: 각각의 개체에 대해 클러스터 Ⅰ, Ⅱ 중 거리가 짧은 그룹에 분류함. 따라서 개체 1, 2는 클러스터 Ⅰ, 개체 3, 4, 5는 클러스터 Ⅱ에 분류함.
- ✓ Step 4: 더 이상의 분류가 필요 없으므로 개체 1, 2를 하나의 클러스터로, 개체 3, 4, 5를 하나의 클러스터로 분류함.

모형 구축은 어떻게 할까?

Unstructured Document

Doc 1 : deposit the cash and check in the bank!!
Doc 2 : the river boat is on the bank
Doc 3 : borrow based on credit
Doc 4 : river boat floats up the river
Doc 5 : boat is by the dock near the bank

Pre-processing

Structured Matrix

Term-by-Document Frequency Matrix

	d1	d2	d3	d4	d5
cash	1	0	0	0	0
check	1	0	0	0	0
bank	1	1	0	0	1
river	0	1	0	2	0
credit	0	0	1	0	0

모형 구축은 어떻게 할까?

- **Term-Document Matrix**

Term-by-Document Frequency Matrix

	d1	d2	d3	d4	d5
cash	1	0	0	0	0
check	1	0	0	0	0
bank	1	1	0	0	1
river	0	1	0	2	0
credit	0	0	1	0	0

군집분석 수행

	Cash	Check	Bank	River	Credit
D1	1	1	1	0	0
D2	0	0	1	1	0
D3	0	0	0	0	1
D4	0	0	0	2	0
D5	0	0	1	0	0

계층적 클러스터링 장단점

- 장점
  - 덴드로그램을 통해 군집이 형성되는 과정을 살펴볼 수 있음
  - 분석대상이 되는 개체들의 수가 비교적 작을 때 유용
- 단점
  - 개체 수가 커지면 유용성이 떨어질 뿐만 아니라 거리 행렬의 계산에 매우 많은 시간과 컴퓨터 용량이 필요하므로 적용에 제약이 따름

비계층적 클러스터링 장단점

- 장점
  - 주어진 데이터의 내부구조에 대한 사전정보 없이 의미 있는 자료구조를 찾을 수 있는 방법
- 단점
  - 사전에 클러스터의 수를 정하는 부담
  - 초기 분류에 의해 영향을 많이 받음

텍스트 클러스터링 활용분야

- 문서 요약
- 유사 중복 문서 검출
- 검색 엔진 최적화
- 추천 시스템
- 기타 양적 연구데이터 적용분야

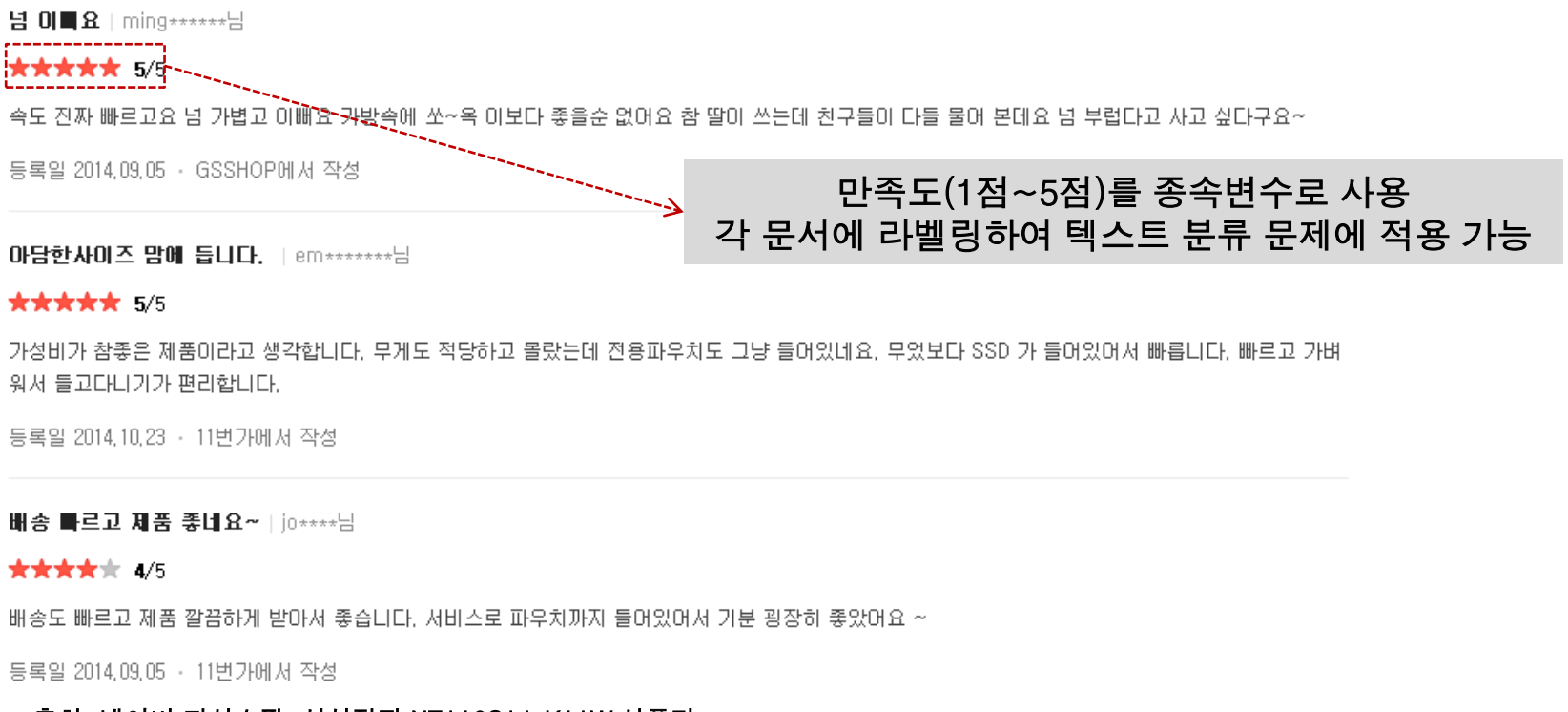
3. 문서 분류

분류(Classification)란?

- 서로 다른 그룹에 속하는 문서의 예가 주어지면
  그 문서의 속성을 사용하여 모델링하고,
  이를 통해 새로운 문서의 카테고리를 예측
- 지도 학습(Supervised Learning)에 의해 모델링
- 분석 기법의 선정, 수렴 및 일반화, 안정성의 점검 등을 수행
- 인공신경망, 의사결정나무, SVM, Naïve Bayes 등 기계학습 기법 이용

문서 분류의 예

■ 온라인 리뷰를 이용한 만족도 분류 예시



만족도(1점~5점)를 종속변수로 사용
각 문서에 라벨링하여 텍스트 분류 문제에 적용 가능

출처: 네이버 지식쇼핑, 삼성전자 NT110S1J-K11W 상품평

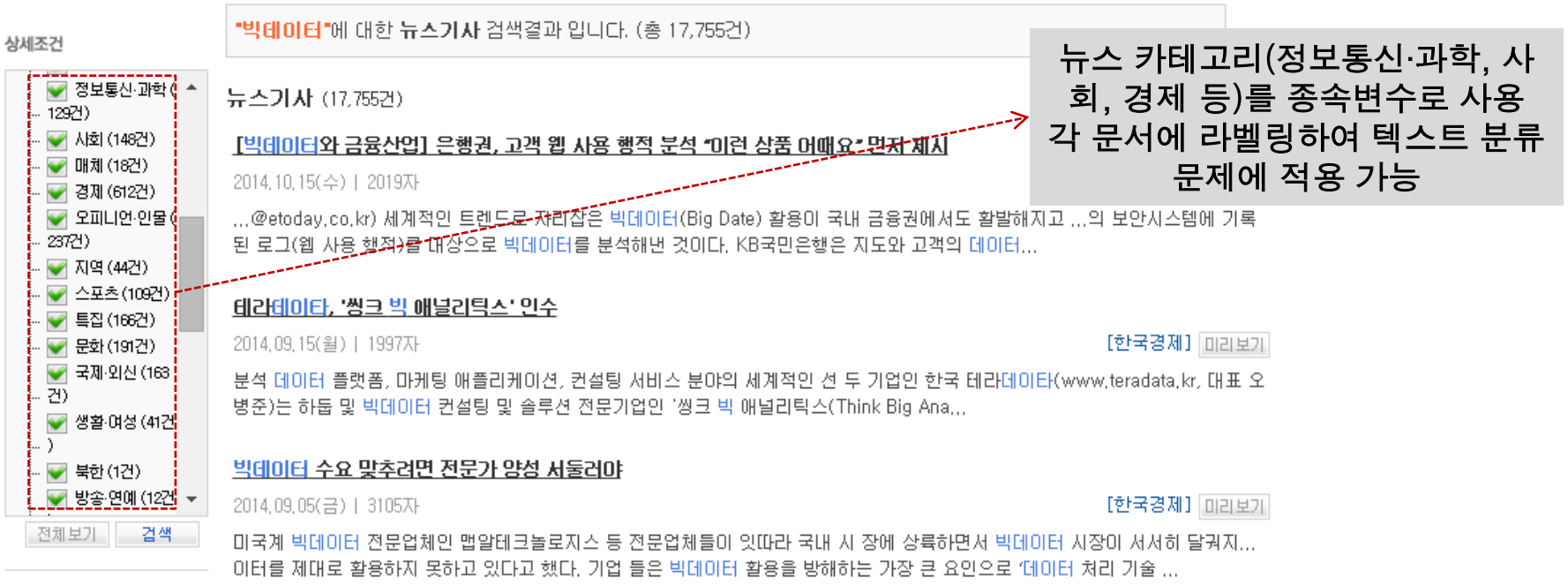
문서 분류의 예

■ 온라인 리뷰를 이용한 만족도 분류 예시

	Inputs	Targets
ID	TEXT	만족도 (별점: 1~5점)
1	속도 진짜 빠르고요 넘 가볍고 이뻐요 가방에 쏘~옥 이보다 좋을순 없어요 참 딸이 쓰는데 친구들이 다들 물어 본대요 넘 부럽다고 사고 싶다구요~	5
2	가성비가 참좋은 제품이라고 생각합니다. 무게도 적당하고 몰랐는데 전용파우치도 그냥 들어있네요. 무었보다 SSD 가 들어있어서 빠릅니다. 빠르고 가벼워서 들고다니기가 편리합니다.	5
3	배송도 빠르고 제품 깔끔하게 받아서 좋습니다. 서비스로 파우치까지 들어있어서 기분 굉장히 좋았어요 ~	4
4	부팅도엄청 빠르고 작고 귀여워요 하지만무게는 꽤나가는느낌이에요ㅠㅠ 여자라서 백에이것저것넣으니..하지만 굉장히 만족하구요 저는 기사님도 엄청 친절하시구 집에들어와주셔서설명도해주시고 엄청 만족합니다^^	4

문서 분류의 예

- 뉴스 기사 카테고리 분류 예시



출처: 한국언론진흥재단 기사통합검색서비스 KINDS , 빅데이터에 대한 뉴스기사 검색결과

문서 분류의 예

■ 뉴스 기사 카테고리 분류 예시

	Inputs	Targets		
ID	TEXT	정보통신 과학	사회	경제
1	정부는 내년부터 범죄가 발생하는 시간과 장소를 예측하는 등에 '빅데이터'(big data·축적된 다량의 정보를 통해 가치를 찾고 결과를 분석하는 기술)를 활용하기로 했다.	1	0	0
2	정보산업진흥원은 올해 10월 '정보통신기술(ICT) 산업전망 콘퍼런스'에서 국내 전문가 556명의 설문 조사를 통해 얻은 '2014년 주목해야 할 ICT 10대 이슈'를 발표했다.	1	0	0
3	건보공단은 2011년 서울고등법원이 흡연과 일부 폐암·후두암의 인과관계를 인정했다는 점에 주목하고, 건보공단이 보유한 빅데이터를 동원해 흡연과 암 발생의 인과관계를 통계적으로 입증할 계획이다.	0	1	0
4	국내 카드업계 1위인 신한카드의 '빅데이터 경영'이 업계 안팎의 주목을 받고 있다. 2200만 고객의 빅데이터 분석을 통해 만들어낸 상품개발 체계인 '코드나인'과 이를 기반으로 만든 새 카드는 선보인 지 두 달 만에 큰 호응을 얻고 있다.	0	0	1
5	정보통신(IT) 강국이라는 한국이 빅데이터 활용에선 약소국인 것으로 나타났다. 대한상공회의소는 14일 국내기업 500개사를 대상으로 빅데이터 활용 현황을 조사한 결과, 응답기업 81.6%가 '활용하지 않고 있다'고 답했다고 밝혔다	0	0	1

분류모형을 어떻게 구축할까?

Unstructured Document

Doc 1 : deposit the cash and check in the bank!!
Doc 2 : the river boat is on the bank
Doc 3 : borrow based on credit
Doc 4 : river boat floats up the river
Doc 5 : boat is by the dock near the bank

Pre-processing

Structured Matrix

Term-by-Document Frequency Matrix

	d1	d2	d3	d4	d5
cash	1	0	0	0	0
check	1	0	0	0	0
bank	1	1	0	0	1
river	0	1	0	2	0
credit	0	0	1	0	0

분류모형을 어떻게 구축할까?

- **Term-Document Matrix**

Term-by-Document Frequency Matrix

	d1	d2	d3	d4	d5
cash	1	0	0	0	0
check	1	0	0	0	0
bank	1	1	0	0	1
river	0	1	0	2	0
credit	0	0	1	0	0

분류모형 구축 수행

	Cash	Check	Bank	River	Credit
D1	1	1	1	0	0
D2	0	0	1	1	0
D3	0	0	0	0	1
D4	0	0	0	2	0
D5	0	0	1	0	0

4. 감성 분석

감성 분석이란?

- 사람들이 작성한 텍스트를 분석하여 특정 주제에 대해 의견이 긍정, 부정, 또는 중립인지를 분류하는 방법으로, 오피니언 마이닝(opinion mining) 이라고도 불림
- 트위터나 페이스북과 같은 SNS 및 인터넷 댓글 등 사안에 대한 오피니언(Opinion)이 잠재되어 있는 문서들을 주로 분석
- 텍스트의 주제가 무엇인지 추출하기 보다는, 그 텍스트를 작성한 사람들이 주제에 대하여 어떠한 태도, 의견, 성향과 같은 주관적인 감정을 가지고 있는지 판단하는 분석기법

감성 분석 예시

나는 오늘 행복하다→ 긍정

나는 R을 매우 좋아하고, R의 활용력을 높이 평가한다→긍정

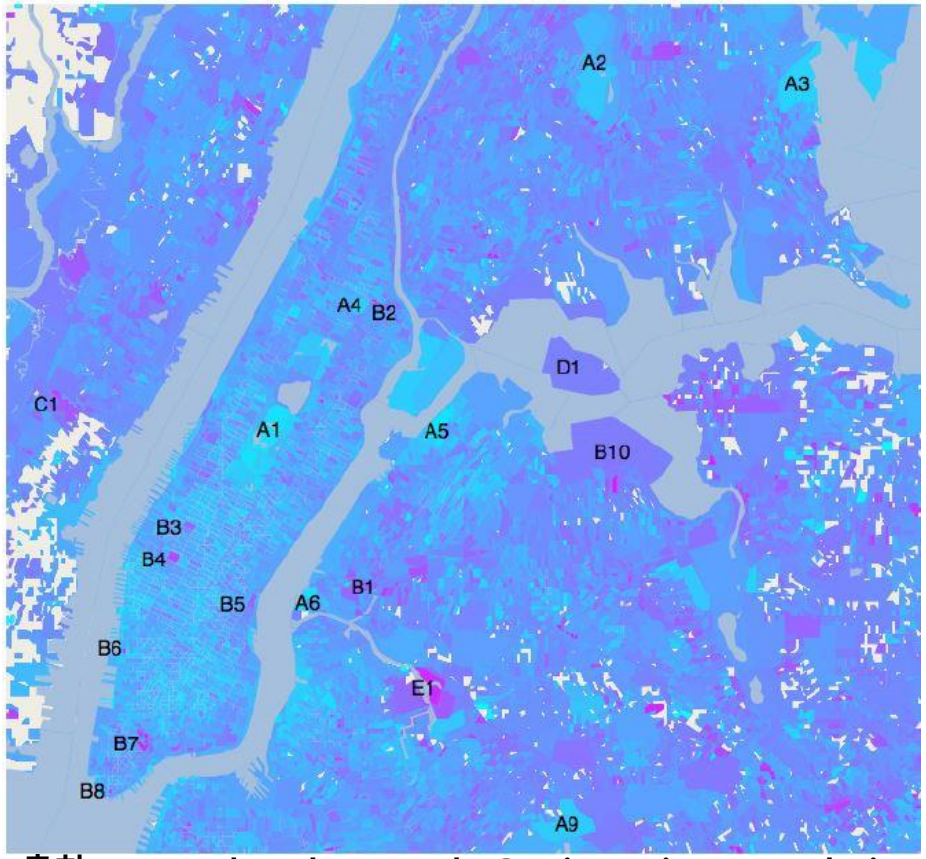
나는 R을 좋아하지 않고, 사용하는 것도 매우 복잡하다→부정

감성 분석 예시

- 분석 대상
  - 2012년 뉴욕시의 맨하튼과 인근지역에 사는 사람들을 대상으로 603,954 트윗에 대하여 조사
- 분석 방법
  - 트윗의 감성을 결정하기 위하여 트위터 키워드, 어휘, 이모티콘을 이용하여 분석 수행
- 분석 결과
  - 지역별로 다른 감정이 나타남
    - √ 공원: 긍정적인 강한 감성
    - √ 교통 허브: 부정적인 강한 감성

감성 분석 예시

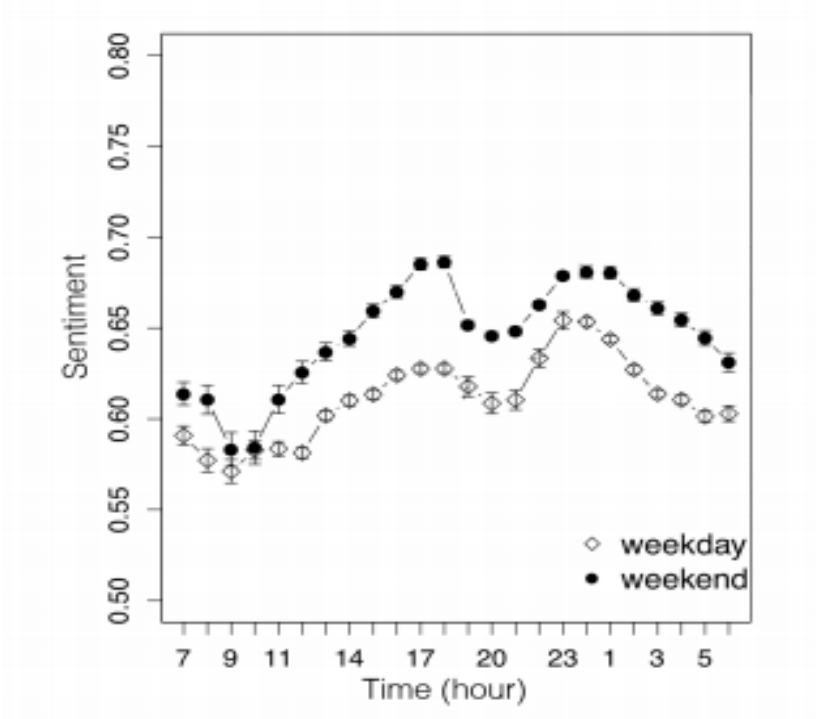
뉴욕시의 감성 맵



출처: Bertrand, Karla Z., et al. "Sentiment in new york city: A high resolution spatial and temporal view." arXiv preprint arXiv:1308.5010 (2013)

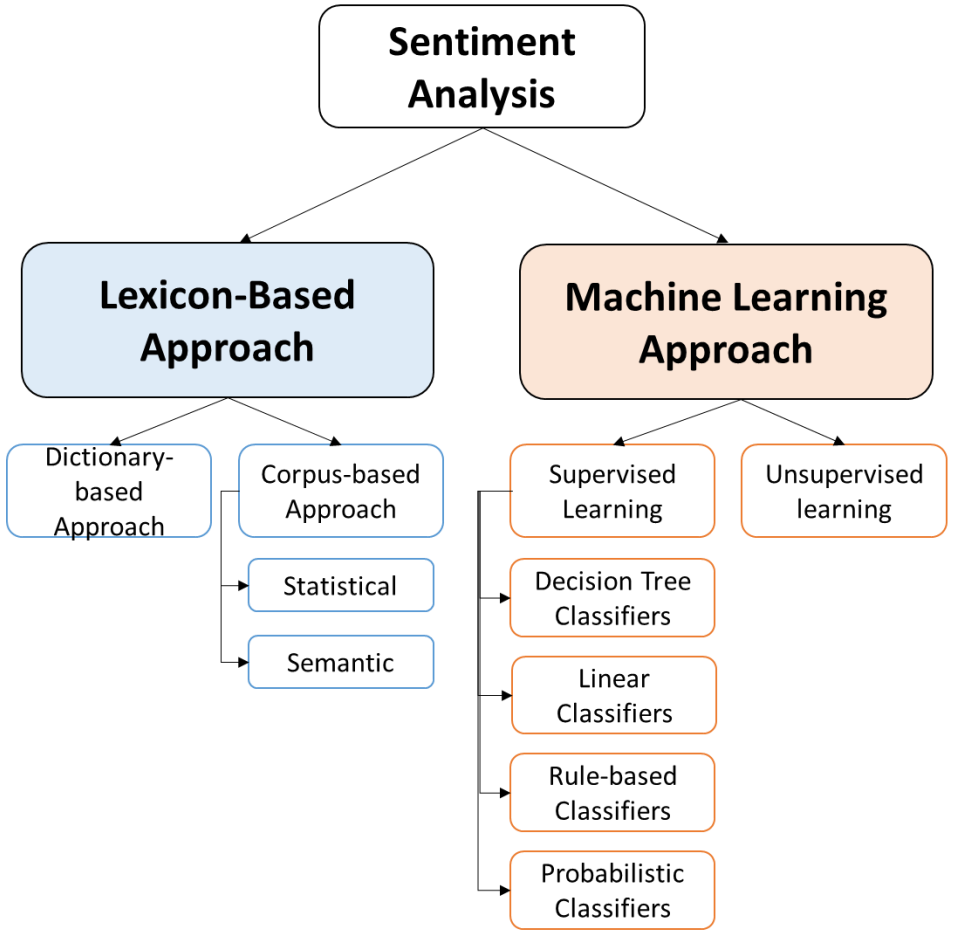
감성 분석 예시

- 주중과 주말의 시간대별 감성패턴을 분석



출처: K.Z. Bertrand, M. Bialik, K. Virdee, Andreas Gros, Y. Bar-Yam, Sentiment in New York City: A High Resolution Spatial and Temporal View. arXiv:1308.5010 (August 20, 2013).

두 가지 감성 분석 방법론


Sentiment Analysis
Lexicon-Based Approach
Machine Learning Approach
Dictionary-based Approach
Corpus-based Approach
Statistical
Semantic
Supervised Learning
Unsupervised learning
Decision Tree Classifiers
Linear Classifiers
Rule-based Classifiers
Probabilistic Classifiers

Lexicon-Based Approach란?

- 감성 어휘 사전 (Opinion/Sentiment Lexicon)을 활용하는 방식
- 감성 어휘 사전 (Opinion Lexicon) 이란?
  - 어떤 단어가 긍정적 단어인지, 부정적 단어인지 분류해 놓은 사전
  - 영어의 경우 Hu and Liu (2004)에 의해 만들어진 약 6,800개의 단어로 구성된 감성어휘사전이 있음.
  - 어휘사전 다운로드 사이트:

    http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

    . 긍정적 단어 예시:

    Love, Best, Cool, Great, Good, Adore, Amazing, Beautiful..

    . 부정적 단어 예시:

    Hate, Worst, Abolish, awful, nightmare, wrong, kill, death..

Lexicon-Based 감성분석

■ 감성점수(sentiment score)의 산출

1 단계:
- 문서의 각각의 문장들에서 단어를 추출

2 단계:
- 각 문장별 감성어휘사전과 비교를 통해 긍/부정 단어의 수에 따른 비자율학습기반의 감성 점수를 계산

3 단계:
– 2단계에서 산출한 감성 점수에 의해 각 문장들에 대해 긍정, 중립, 부정으로 분류

4 단계:
- 문장들로 이루어진 문서에 대한 평균 감성 점수를 산출하거나 그래프를 통해 텍스트를 작성한 사람이 주제에 대하여 어떠한 태도, 의견, 성향과 같은 감정을 가지고 있는지 판단

Lexicon-Based 감성분석

- 감성점수(sentiment score) 산출식 예시

문서의 감성 점수는 긍정적 단어가 나타나면 +1,
부정적 단어가 나타나면 -1을 부여한 후, 합산함.

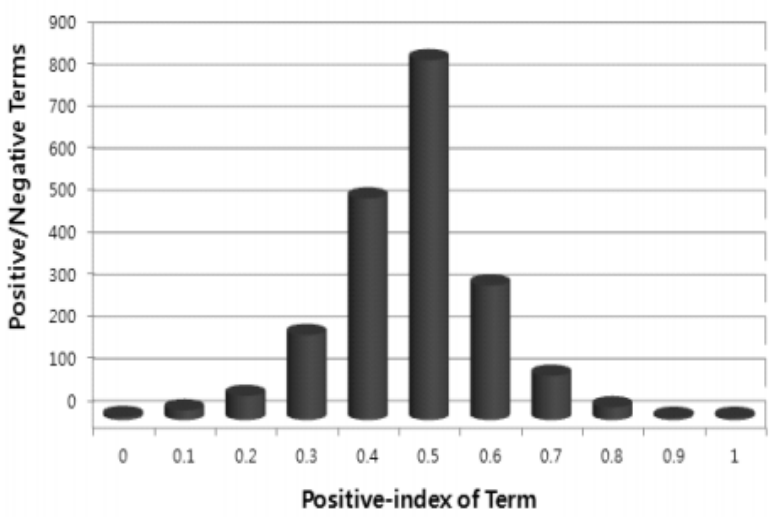
감성 점수 > 0,
  긍정적 의견 (positive opinion)을 나타내는 것으로 간주함.
감성 점수 < 0,
  부정적 의견 (negative opinion)을 나타내는 것으로 간주함.
감성 점수 = 0,
  중립적 의견 (neural opinion)을 나타내는 것으로 간주함.

Lexicon-Based 감성분석

■ 감성점수(sentiment score) 산출식 예시

해당 도메인에 따라 단어별 극성 강도를 표현하는 것도 가능

긍정 어휘		중립 어휘		부정 어휘	
어휘	긍정 지수	어휘	긍정 지수	어휘	긍정 지수
진출	0.7	가늠	0.5	역부족	0.1
안도	0.7	가능	0.5	급락세	0.2
급반등	0.7	가량	0.5	난항	0.2
안도감	0.7	개월	0.5	내림세	0.2
이익률	0.7	거래일	0.5	두려움	0.2
상승률	0.7	경우	0.5	심각	0.2
낙관론	0.7	기간	0.5	하락률	0.2
낙관적	0.7	너스	0.5	먹구름	0.3
희망	0.6	년간	0.5	불투명	0.3
호재	0.6	년래	0.5	쇼크	0.3
호전	0.6	누구	0.5	직격탄	0.3
호조	0.6	다음	0.5	추락	0.3
호평	0.6	다음날	0.5	침체	0.3
확장	0.6	모습	0.5	감소	0.4
상승세	0.6	사실	0.5	부정적	0.4
성공	0.6	사실상	0.5	불안	0.4
긍정적	0.6	수년	0.5	불안감	0.4
기대감	0.6	시각	0.5	위축	0.4
오름세	0.6	시간	0.5	위험	0.4



출처: 유은지, 김유신, 김남규, & 정승렬. (2013). 주가지수 방향성 예측을 위한 주제지향감성사전 구축 방안. 지능정보연구, 19(1), 95-110.

Lexicon-Based Approach의 장단점

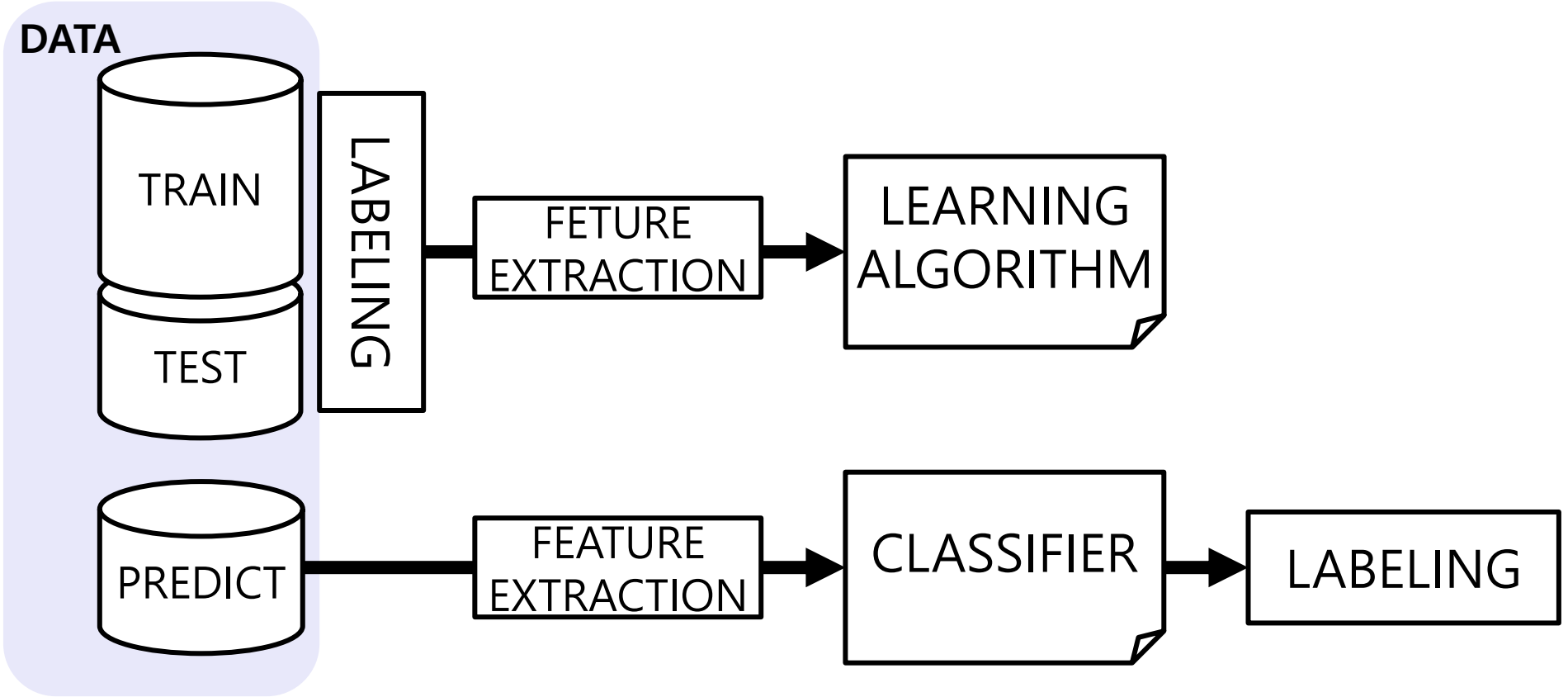
- 장점
  - 감성어휘사전이 있을 경우 사용하기 간편하다는 장점
- 단점
  - 역설이나 풍자를 내포하고 있을 경우 다중의미 파악이 어렵다는 점
  - 감성어휘사전 구축에 비용이 많이 든다는 점
  - 감성어휘사전이 분야마다 상이하다는 점
  - 지속적인 신조어 관리가 필요하다는 점

Machine Learning Approach란?

- 텍스트에 극성 (긍정/부정) 라벨링이 존재하는 경우에 사용할 수 있는 감성 분석 방법
- 문서 분류 기법과 유사
- 인공신경망, 의사결정나무, SVM, Naïve Bayes 등 지도 학습 기반의 (Supervised Learning) 기계학습 기법을 이용하여 문서 전체가 긍정인지 부정인지 분류

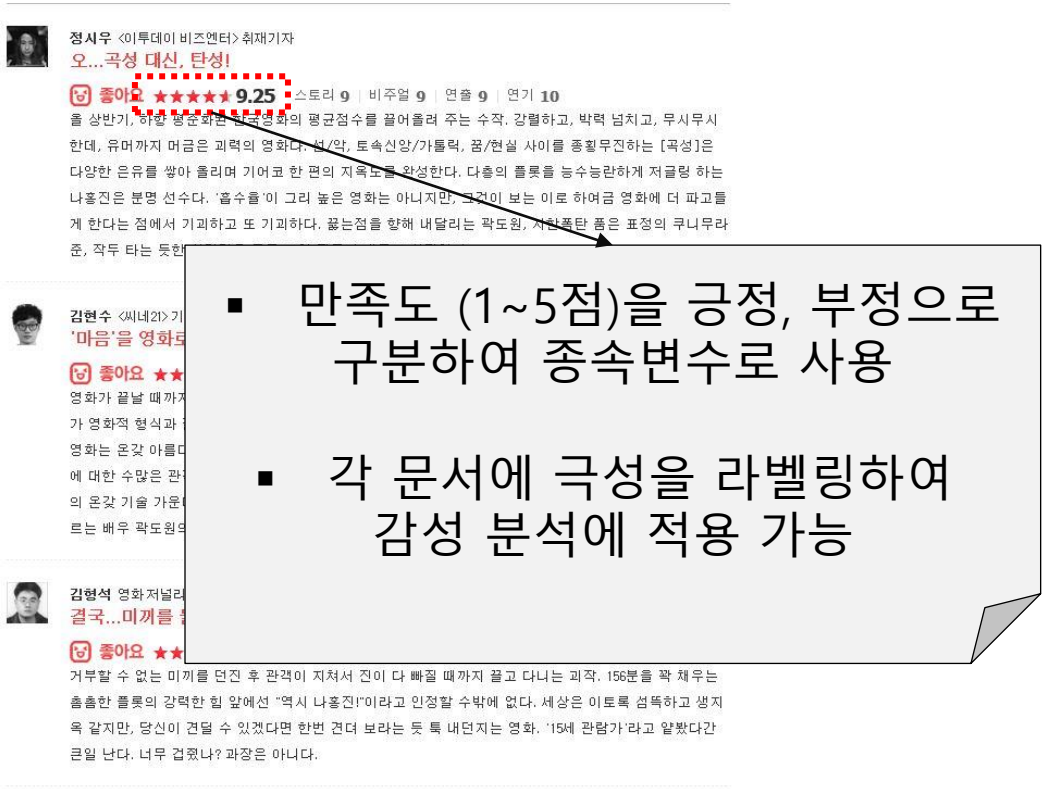
Machine Learning Approach란?

- 기계학습(machine learning)을 이용한 감성 분류 모델



Machine Learning 감성분석 방법 예시

- “곡성” 온라인 영화평



- 만족도 (1~5점)을 긍정, 부정으로 구분하여 종속변수로 사용
- 각 문서에 극성을 라벨링하여 감성 분석에 적용 가능

Machine Learning 감성분석 방법 예시

■ 수집된 데이터

INPUT TARGET

ID	TEXT	PALARITY
1	올 상반기, 하향 평준화된 한국영화의 평균점수를 끌어올려 주는 수작. 강렬하고, 박력 넘치고, 무시무시한데, 유머까지 머금은 괴력의 영화다. 선/악, 토속신앙/가톨릭, 꿈/현실 사이를 종횡무진하는 [곡성]은 다양한 은유를 쌓아 올리며 기어코 한 편의 지옥도를 완성한다.	POSITIVE
2	영화가 끝날 때까지 보이는 어떤 것도 믿을 수 없는 영화다. 이야기와 캐릭터가 담고 있는 주제, 혹은 의미가 영화적 형식과 잘 어우러진다. 들리는 것 역시 마찬가지다. 즉, 영화를 다 보고 돌이켜 생각해보면 이 영화는 온갖 아름다운 영화적 기법으로 관객을 현혹시키고 있다.	POSITIVE
3	거부할 수 없는 미끼를 던진 후 관객이 지쳐서 진이 다 빠질 때까지 끌고 다니는 괴작. 156분을 꽉 채우는 촘촘한 플롯의 강력한 힘 앞에선 "역시 나홍진!"이라고 인정할 수밖에 없다. 세상은 이토록 섬뜩하고 생지옥 같지만, 당신이 견딜 수 있겠다면 한번 견뎌 보라는 듯 툭 내던지는 영화. '15세 관람가'라고 얕봤다간 큰일 난다. 너무 겁줬나? 과장은 아니다.	POSITIVE

Machine Learning 감성분석 방법 예시

- Term-Document Frequency Matrix 로 변환

	d1	d2	d3	~	D m
Term1	1	0	0	~	0
Term2	1	1	1	~	3
Term3	1	1	0	~	1
Term4	0	1	0	~	2
Term5	0	0	1	~	0
~	~	~	~	~	~
Term n	2	2	0		1
Label Y	P	N	P		N

Machine Learning 감성분석 방법 예시

	d1	d2	d3	~	D m
Term1	1	0	0	~	0
Term2	1	1	1	~	3
Term3	1	1	0	~	1
Term4	0	1	0	~	2
Term5	0	0	1	~	0
~	~	~	~	~	~
Term n	2	2	0		1
Label Y	P	N	P		N

감성
분류모형
구축
수행

	Term1	Term2	Term3	~	Term n	Label Y
D1	1	1	1	~	0	P
D2	0	0	1	~	0	N
D3	0	0	0	~	1	P
~	~	~	~	~	~	
D m	0	0	1	~	0	N

Machine Learning Approach의 장단점

- 장점
  - 분류 예측력 우수
- 단점
  - 텍스트의 극성(출력 변수)을 알 수 없는 경우, 수동으로 라벨링(Labeling)하여 감성 분석을 위한 데이터를 생성해야 하는 어려움

활용 분야

- Voice of Customer 청취 및 분석
  - 소비자들의 제품에 대한 반응 탐지
  - 브랜드 이미지에 대한 모니터링 등
- 여론
  - 국가 정책에 대한 국민의 생각 분석
  - 선거 예측 등
- 각종 예측
  - 상품이나 서비스의 판매예측
  - 자본시장 예측
  - 영화 등 문화 상품의 흥행예측 등