환경 텍스트 데이터 분석 방법론

한국환경정책·평가연구원
환경경제연구실
부연구위원 진대용

KEI 한국환경정책·평가연구원
Korea Environment Institute

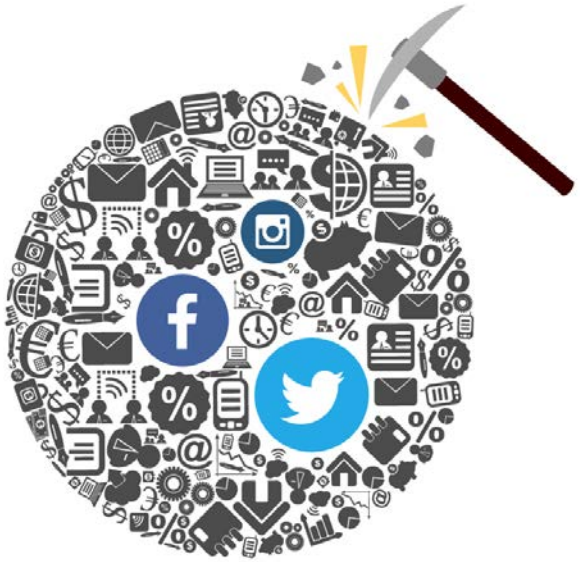
발표자 소개

- 이름 : 진대용 (dyjin@kei.re.kr)
  - 광주과학기술원 전기전자컴퓨터공학 박사 (2017)
  - 한국환경정책·평가연구원 환경경제연구실 부연구위원 (2017~)

- 관심분야
  - 기계학습 및 딥러닝, 자연어처리 및 텍스트 마이닝
  - 계산생물학(Computational Biology), 환경 빅데이터 분석
- 연구활동
  - 생활밀착형 환경 이슈에 대한 수요반영 개선 연구 : 민원 빅데이터 분석을 중심으로(2019)
  - 기후환경 이슈분석을 위한 텍스트 마이닝 활용 방안 (2018)
  - 대기이미지를 활용한 미세먼지 오염도 추정 (2018)
  - Jin, Daeyong, and Hyunju Lee. "Prioritizing cancer-related microRNAs by integrating microRNA and mRNA datasets." Scientific reports 6 (2016): 35350.
  - Jin, Daeyong, and Hyunju Lee. "A computational approach to identifying genemicroRNA modules in cancer." PLoS computational biology 11.1 (2015): e1004042

텍스트 마이닝

- 텍스트 마이닝은 글(텍스트)를 캐낸다는 의미, 여기서 마이닝은 'mining‘ 을 나타낸 것으로 ‘(광산을)채굴하다 ’ 의 뜻
- 단어의 출현 빈도, 단어 간 관계성 등을 파악하여 유의미한 정보를 추출하는 것
- 빈도수의 마법



http://media.fastcampus.co.kr/knowledge/about-text-mining/

텍스트 마이닝 하면 떠오르는것?

Crawling/ Scrapping
rvest
Beatifulsoup
Html/XML
Regular expression

Latent semantic analysis
Co-occurrence
Dimension reduction
(SVD,PCA)

Sentiment analysis
Lexicon based
Machine learning based

Language modeling
Uni/Bigram
N-gram

NLP
Lexical analysis
Syntax analysis
Semantic analysis

Word embedding
Word2Vec
Glove
Fasttext

Visualization
Keyword analysis
Association analysis

Preprocessing
Tokenization
Cleaning
Stemming/ Lemmatization
Stopword Filtering

Topic modeling
LDA
LSA/pLSA

Lexical analysis
Co-reference
POS tagging
Named entity recognition

Document/Sentence Classification
SVM, Naïve Bayes, CNN/RNN

Document representation
Bag-of-words (DTM, TDM,TF)
TF-IDF
Word embedding

텍스트 마이닝 과정

어떤 텍스트 자료를 활용하여 환경 이슈를 도출할 것인지 선택

문장을 형태소 단위로 나누고 품사 태깅을 수행

텍스트 분석에 필요 없는 단어를 처리

각 문서에서 개별 단어의 빈도수를 산출 한 뒤 각 단어의 중요도를 어떻게 표현할 것인지 설정

키워드 빈도수, 토픽 모델링, 연관키워드, 감성분석 적용

분석으로 도출된 결과에 대한 상세 분석

형태소 분석 예시

아버지가 방에 들어가신다

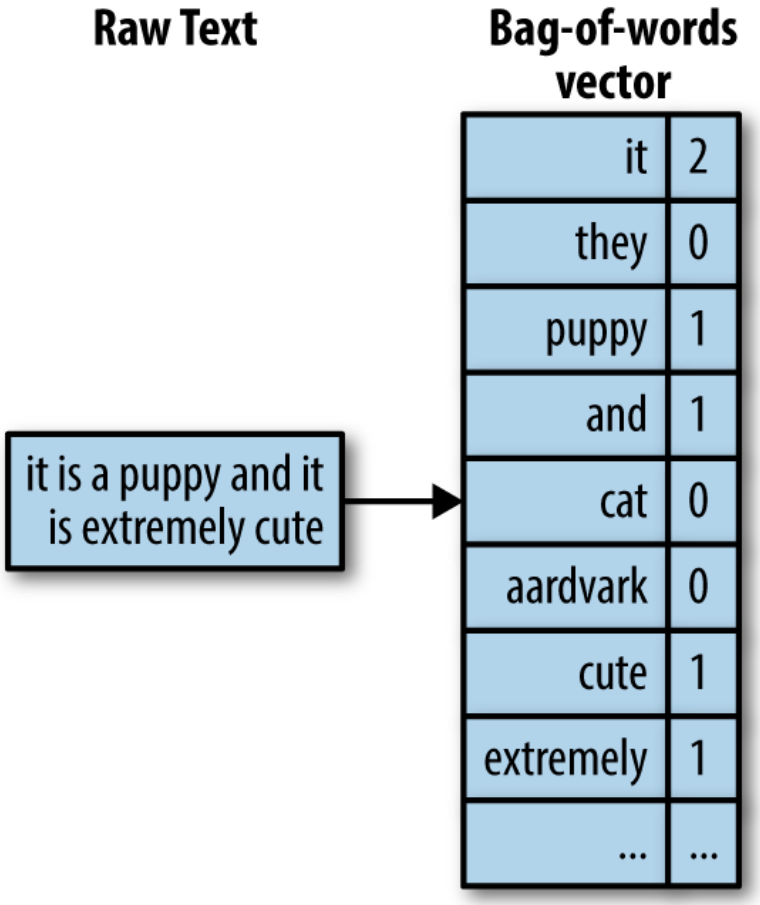
['아버지', '가', '방', '에', '들어가신다']

	환경	미세먼지	쓰레기	악취	소음
Doc1	5	0	2	2	0
Doc2	2	3	4	0	0
Doc3	1	0	0	0	2
DocN	2	4	0	1	2

문서-단어 행렬

Bag of Words

- 특정 문서에서 단어의 순서들을 생각하지 않고 단어가 출현하는 빈도 수만을 보는 방법
- NLP 기본 가정 중 하나인 Bag of words hypothesis
  - Turney & Pantel (2010:153): Bag of words hypothesis
  - "The frequencies of words in a document tend to indicate the relevance of the document to a query (Salton et al., 1975). – If documents and pseudo-documents (queries) have similar column vectors in a term–document matrix, then they tend to have similar meanings."



https://wdprogrammer.tistory.com/31
http://uc-r.github.io/creating-text-features

키워드 빈도수 분석

- TF vs TF-IDF
  - TF (term frequency) : 단순 빈도수
  - TF-IDF (term frequency inverse document frequency)
    - 문서에 자주 등장하는 단어에 패널티

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

- 유니그램(unigram) vs 바이그램(bigram)
  - 온실가스 감축 : unigram (온실가스, 감축), bigram (온실가스 감축)
  - 기후변화 협약 : unigram (기후변화, 협약), bigram (기후변화 협약)

명사인식 (Noun recognition)

- 전문가 지식
- 명사인식 알고리즘 -> 검증

문장 1: 나는 오늘 'N'에 갈 예정이야.

문장 2: 9시에 'N'에서 만나자

문장 3: 학생들이 'N'으로 모였다.

단어를 반으로 쪼개서 오른쪽 단어의 분포를 보고서 명사인지 판단

예를 들어 'N'이 '도담동' 이라고 가정

1) '도담동' 을 명사로 취급

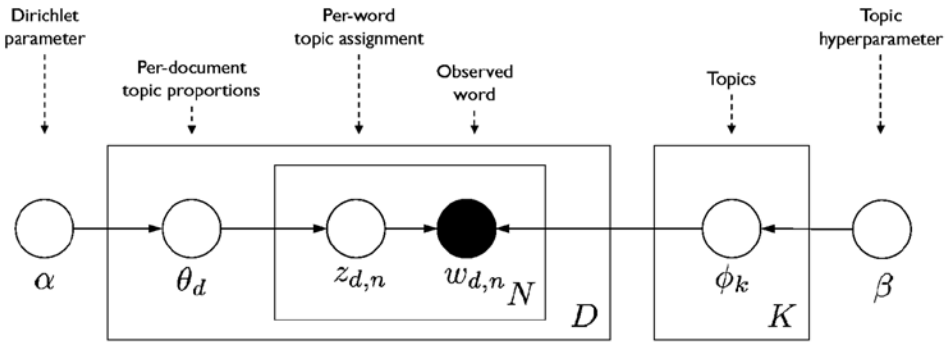
'도담동'에, '도담동'에서, '도담동'으로

2) '도담' 을 명사를 취급

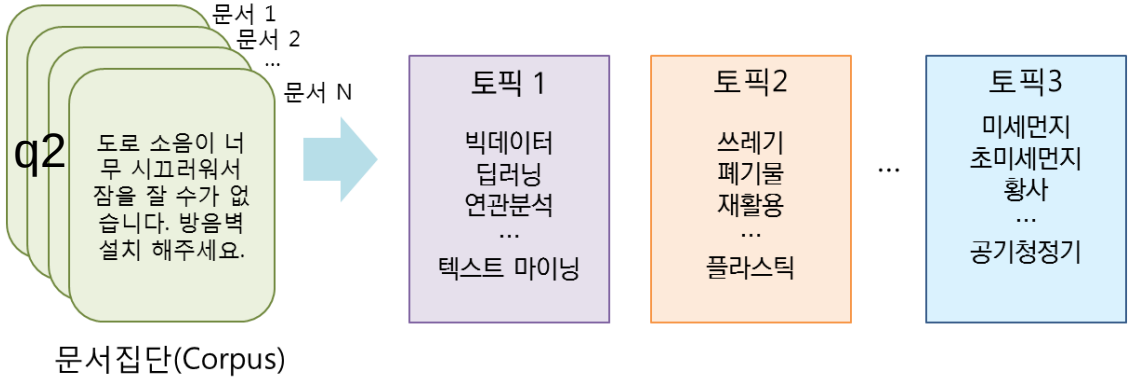
'도담'동에, '도담'동에서, '도담'동으로

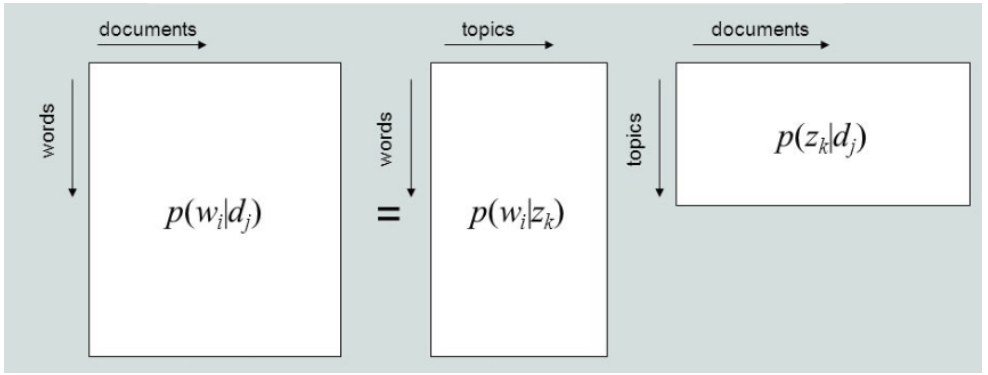
https://github.com/lovit/soynlp/blob/master/notes/unskonlp.pdf

토픽 모델링 (Topic Modeling)



- LDA (Latent Dirichlet Allocation)
  - 토픽 모델링 방법
  - 문서의 집합(코퍼스)으로부터 (숨겨진) 토픽의 등장확률을 계산하는 텍스트 마이닝 방법론
  - 단어(Word) -> 토픽(Topic) -> 문서(Document)
  - 각 문서에 대해 확률이 가장 큰 값을 가지는 주제를 문서의 주제로 설정



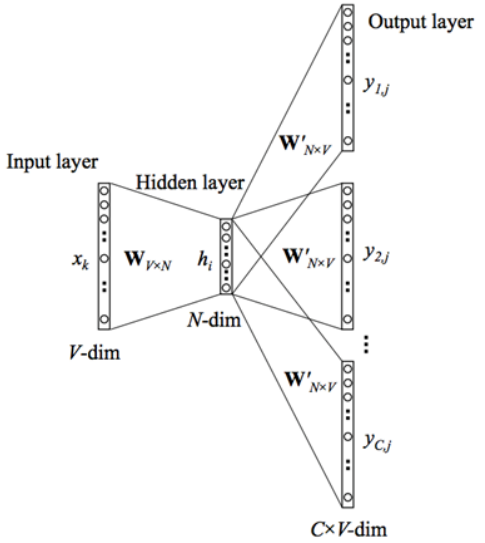


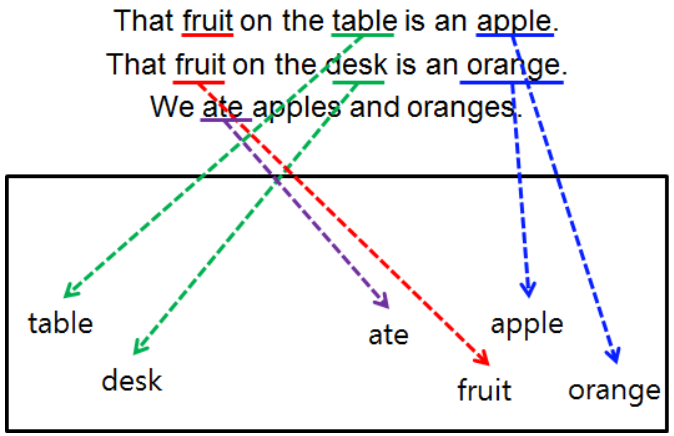
LDA 분석 사례 : 네이버 환경 뉴스 (2005~2017)

기타1	가축	기후변화 (재난)	원전	기름유출	기후변화 (재난)	생태계	기후변화 (온실가스)	쓰레기/폐기물	화학물질	수질오염	기후변화 (재난)	기후변화 (재난)	국립공원	기타2	기타3	미세먼지	교통
아이	농가	폭염	원전	방제	한파	식물	온실가스	쓰레기	지진	수질	얼음	태풍	국립공원	강수량	철새	미세먼지	소음
프로그램	구제역	어린이	방폐	유출	대설주의보	습지	기후변화	산림	석면	하천	폭설	강풍	동물	기압골	소나무	황사	교통
참가	중국	무더위	결정	해역	규칙	생태	에너지	폐기물	화학물질	녹조	장마	호우	야생동물	번개	천연기념물	오존	호선
참여	바이러스	초여름	원자력	양식	특보	생물	감축	친환경	공장	오염	절정	폭우	돌고래	천둥	저수지	중국	수돗물
어린이	중국망	열대야	검토	어민	대설	서식	탄소	재활용	오염	상류	온도	가뭄	고래	천둥 번개	연못	초미세먼지	철도
그린	오리	임영주	의견	연안	건조	생태계	기후	자원	가습기	지하수	시기	침수	방사	미세먼지	번식	미세먼지 미세먼지	개통
생태	조류인플루엔자	폭염 특보	소송	산과학	철원	멸종위기	온실가스 감축	도시	위반	조류	강수량	통제	야생	지역별	산시	마스크	자전거
생명	가축	여름	부지	태안군	정읍	자연	이산화탄소	음식물	살균제	방류	한반도	초속	등산	남해안	날개	계절	노선
운동	축사	장맛비	입장	해변	적설량	보전	목표	매립	가습기 살균제	가뭄	온난화	파도	포획	위치	그루	기질	요금
자연	감염	소나기	판결	해안	안동	보호	배출량	음식물 쓰레기	보건	하수	평균기온	부근	사냥	소나기	머리	유차	항공
이야기	축산	야외	의회	원유	평지	서식지	투자	생산	가스	하류	예년	집중호우	보호	교차	이동	외출	택시
생각	돼지	특보	논란	유조선	중인	자생	친환경	순환	폐수	정화	리기	특보	멸종위기	돌풍	취하	석탄	상수도
마음	예방	불볕	위원	오염	양양	국립공원	태양광	소각	환경오염	수위	서리	총주	동물보호	남북	조류	호흡기	사흘
마당	당국	찜통	방조제	속보	고창	곤충	중국	녹지	준치	수문	중순	복구	자연	충청	어미	자제	정체
에코	총리	지면	토론회	자원봉사자	구미	야생	참여	시범	실내	미군	한파	제천	개체	건강	산책로	오염	서구
청소년	재난	한낮	시민단체	수온	분포	확인	전력	선정	중금속	강변	해수면	강수량	치료	조업	월동	자동차	방향
관람객	방지	폭염 폭염	허가	방사	양구	개체	아시아	투기	발암	금강	변화	산사태	그물	서부	부리	실외	항공기
실천	조선	지수	위원장	어선	원주	희귀	논의	보급	화학	향기	여름	풍속	국립공원 국립공원	산발	도래	발원	한경
주최	베이징	비롯	합의	방사성	대한민국 희망	갯벌	생산	자원 순환	허용	미군기지	대륙	준비	천연기념물	선박	무리	환경과학	차로
선정	의심	햇볕	제기	선박	희망	동물	도시	에너지	확인	용수	해발	동반	멸종위기종	남쪽	줄기	착용	고속

Word2vec : 단어 임베딩

- Word2vec
  - CBOW / Skip-gram
    - CBOW(Continuous Bag of Words) : 주변 단어를 이용하여 특정 단어를 예측
    - Skip-gram : 특정 단어를 이용하여 주변 단어를 예측
  - 문자로 이루어진 단어를 숫자 (벡터) 로 변환
  - 단어 유사도 계산, 유사어 처리, 연관 키워드 분석 등



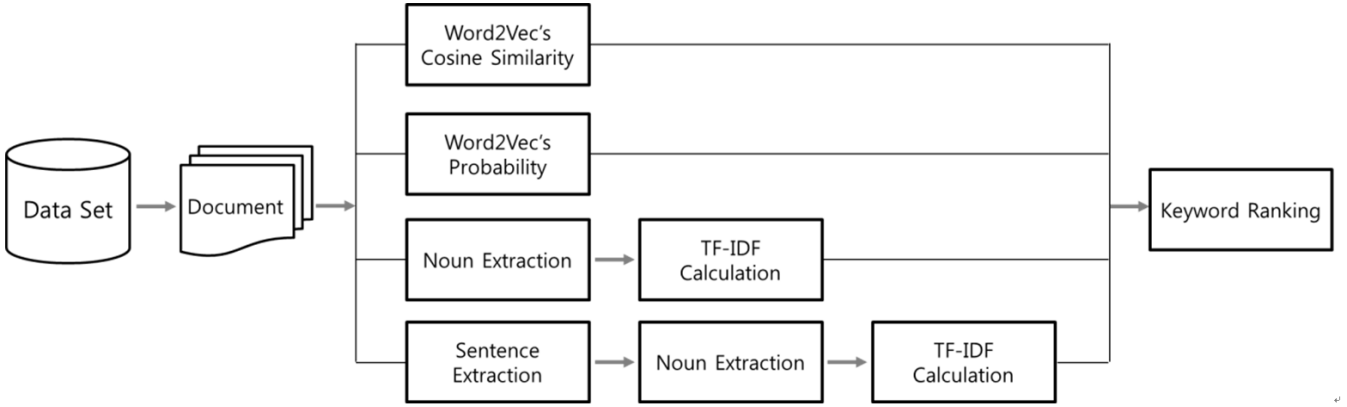


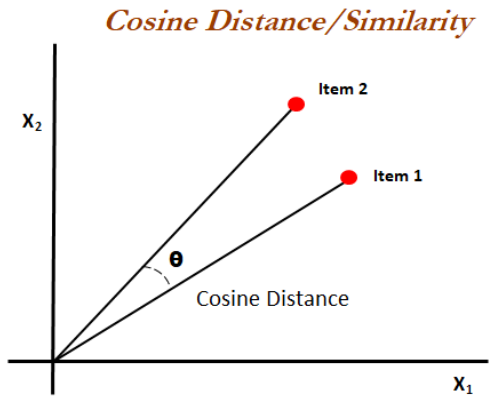
"유사한 문맥에 있는 단어는 유사한 좌표로 변환"

참고강의 : https://www.youtube.com/watch?v=sY4YyacSsLc

연관 키워드 분석 : '친환경' 연관키워드 찾기

- Word2vec (좌표상의 거리)
  - 좌표상의 거리를 활용
  - '친환경' 단어와 거리가 가까운 단어는?
- Word2vec (단어의 분포)
  - 단어의 분포를 활용한 방법
  - '친환경' 단어가 근처에서 주로 어떤 단어가 나타나는가?
- 문서추출 + 빈도수 조사
  - '친환경' 단어가 포함된 문서를 추출한 뒤 빈도수 조사
- 문장추출 + 빈도수 조사
  - 문서를 문장으로 쪼갠 뒤, '친환경' 단어가 포함된 문장을 추출 후 빈도수 조사





연관 키워드 분석 : '친환경' 연관키워드 찾기

순위	키워드	순위	키워드
1	녹색	31	건축
2	장려	32	표방
3	혁신	33	생태도시
4	그린	34	편리
5	에코	35	전환
6	아이디어	36	에코디자인
7	촉진	37	창의
8	로컬	38	지속가능성
9	푸드	39	이미지
10	무공	40	옥상녹화
11	환경마크	41	각광
12	드라이브	42	꾀
13	지향	43	접목
14	제로	44	유도
15	전기자동차	45	주력
16	세제	46	절약
17	위주	47	로하스
18	클린	48	자재
19	브랜드	49	탄소
20	대체에너지	50	환경
21	실천	51	녹색건축
22	녹색교통	52	제품
23	신재생에너지	53	대체연료
24	디자인	54	차세대
25	포장	55	실생활
26	합리화	56	인센티브
27	공공	57	그린카드
28	동력	58	효율
29	계시	59	패러다임
30	전기차	60	물류

순위	키워드	순위	키워드
1	제품	31	유통업체
2	운전	32	전환
3	상품	33	성장
4	인증	34	아이디어
5	건축물	35	유통
6	친환경	36	조성
7	타운	37	그린카드
8	소비	38	비문
9	경영	39	캠페인
10	구매	40	빌딩
11	공법	41	기아차
12	농산물	42	시책
13	명절	43	비즈
14	자동차	44	자재
15	생태도시	45	교통
16	에너지	46	수상
17	건축	47	에코
18	전기차	48	명품
19	농법	49	브랜드
20	포장	50	유차
21	생활	51	주행
22	전기자동차	52	대중교통
23	매장	53	혜택
24	각광	54	우수
25	타이어	55	확산
26	설제	56	정착
27	보급	57	표방
28	무공	58	메카
29	페인트	59	개발
30	실천	60	도시

순위	키워드	순위	키워드
1	에너지	31	타운
2	녹색	32	전기차
3	탄소	33	절감
4	친환경 에너지	34	에너지 타운
5	온실가스	35	서울시
6	자동차	36	건축
7	도시	37	이산화탄소
8	상품	38	참여
9	그린	39	절약
10	확대	40	교통
11	생산	41	농업
12	보급	42	투자
13	경차	43	전환
14	실천	44	태양광
15	친환	45	친환경 자동차
16	경영	46	자원
17	친환 경차	47	소비자
18	운전	48	노력
19	친환경 상품	49	온실가스 감축
20	기후변화	50	하이브리드
21	소비	51	시스템
22	감축	52	탄소 친환경
23	친환경 운전	53	건축물
24	연료	54	미세먼지
25	도입	55	친환경 소비
26	생태	56	전자
27	효율	57	친환경 경영
28	구축	58	에너지 절약
29	효과	59	선정
30	에코	60	미래

순위	키워드	순위	키워드
1	기후변화	31	전력
2	감축	32	자원
3	생산	33	구축
4	도시	34	비용
5	자동차	35	규제
6	그린	36	에코
7	확대	37	자전거
8	이산화탄소	38	오염
9	미세먼지	39	노력
10	온실가스 감축	40	가스
11	참여	41	발전소
12	연료	42	하천
13	효과	43	일본
14	생태	44	경영
15	배출량	45	폐기물
16	태양광	46	습지
17	목표	47	절약
18	효율	48	절감
19	투자	49	연간
20	실천	50	신재생에너지
21	전기차	51	기존
22	기후	52	피해
23	서울시	53	재활용
24	도입	54	부문
25	보급	55	선정
26	전기	56	친환경 에너지
27	자연	57	증가
28	시스템	58	상품
29	소비	59	하이브리드
30	중국	60	교통

연관 분석

- SUPPORT(지지도)
  - 키워드 A, B가 같은 문서에서 동시에 발생할 확률 $P(A \cap B)$
- CONFIDENCE(신뢰도)
  - 키워드 A가 나타났을 때, 키워드 B가 같은 문서 내에서 발생할 확률 $P(B|A)$
- LIFT(향상도)
  - 키워드 A와 B가 동시에 나타날 확률(지지도)값에 각각의 키워드가 나타날 문서에 확률을 나누어 주는 패널티로써의 역할 $P(A \cap B)/P(A)P(B)$

〈표 2-3-14〉 NAVER 환경뉴스(2013~2016년) 키워드 연관 분석

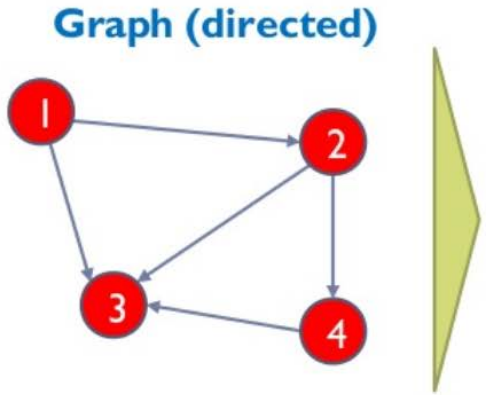
순번	lhs		rhs	지지도	신뢰도	향상도
1	간접	⇒	태풍	0.002	0.960	51.279
2	태풍	⇒	간접	0.002	0.081	51.279
3	대설	⇒	특보	0.001	0.380	40.323
4	특보	⇒	대설	0.001	0.128	40.323
5	최강	⇒	한파	0.001	0.688	62.212
6	한파	⇒	최강	0.001	0.105	62.212
7	고니	⇒	태풍	0.001	0.675	36.055
8	태풍	⇒	고니	0.001	0.061	36.055
9	반달	⇒	지리	0.001	0.685	153.626
10	지리	⇒	반달	0.001	0.237	153.626
11	가슴	⇒	지리	0.001	0.667	149.529
12	지리	⇒	가슴	0.001	0.237	149.529

자료: 저자 작성.

자료 : http://www.hmwu.idv.tw/web/R/C04-hmwu_R-AssociationRule.pdf

네트워크 분석의 기본

- 네트워크 분석
  - 그래프 구성
    - 노드와 엣지를 정의
    - 예시
      - 노드 : 키워드
      - 엣지 : 코사인 유사도



Edge list

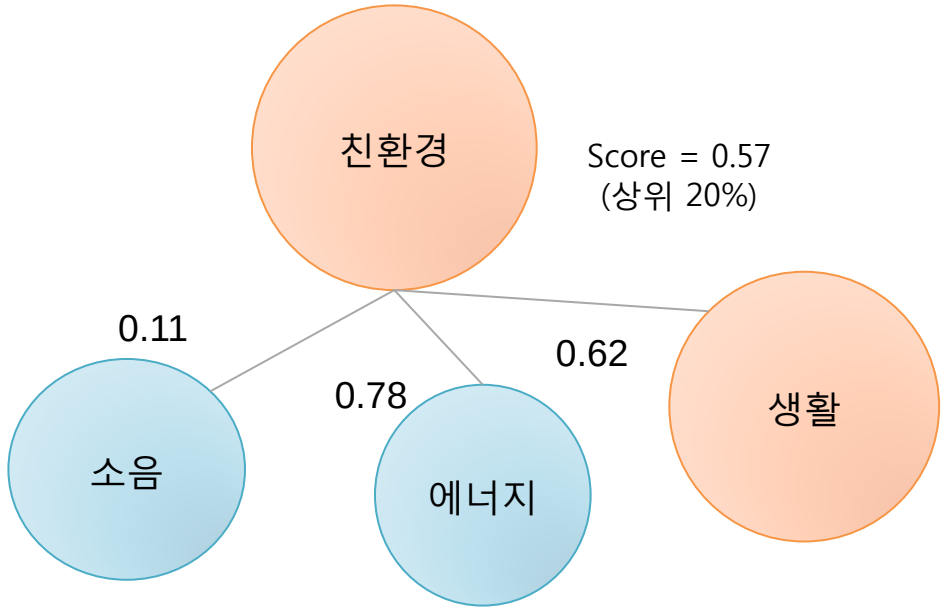
Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

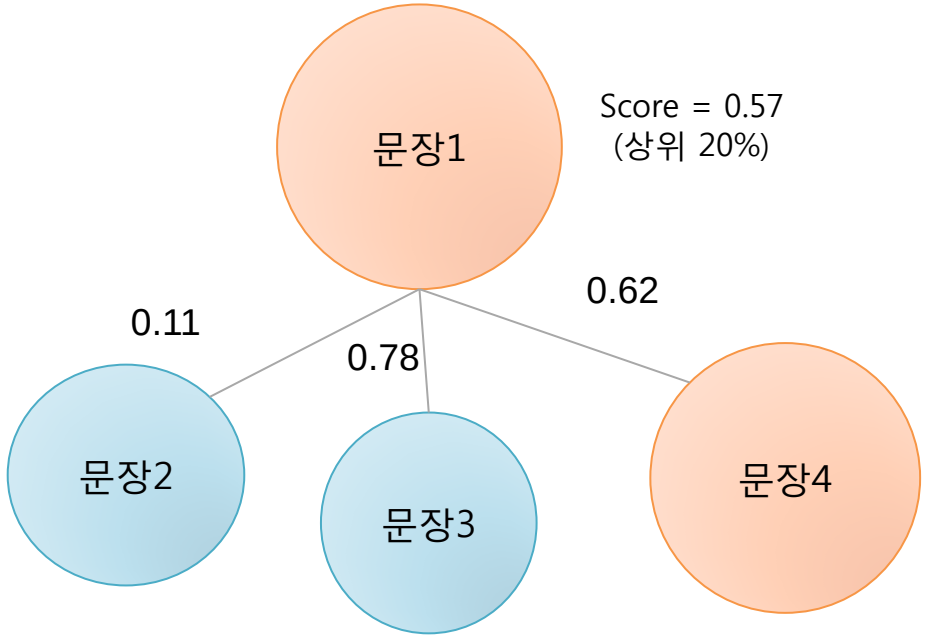
Adjacency matrix

Vertex	1	2	3	4
1	-	1	1	0
2	0	-	1	1
3	0	0	-	0
4	0	0	1	-

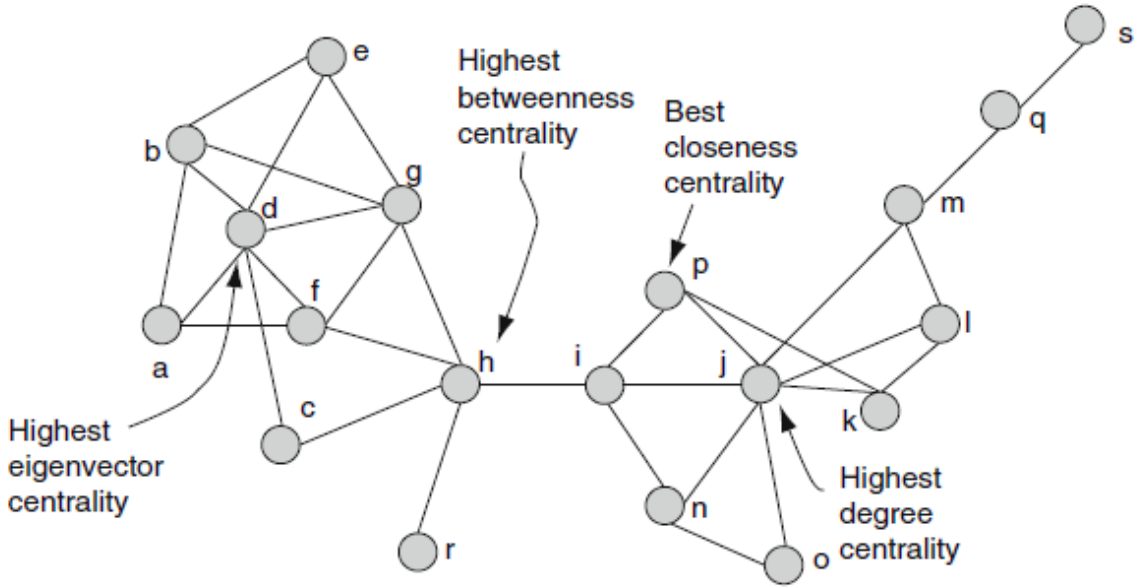
키워드 네트워크 시각화

- 키워드들 사이의 연관성을 계산하여 네트워크 시각화
- 예 : '친환경' 네트워크
  - 노드와 엣지는 연구자의 판단에 따라 여러 방법으로 정의할수 있음
  - 노드 (node)
    - 노드 크기 : In-Degree (노드로 들어오는 엣지의 개수)
  - 엣지 (edge)
    - Word2vec 코사인 유사도 (상위 20%)





네트워크 중심성 : 중요노드 선정방법



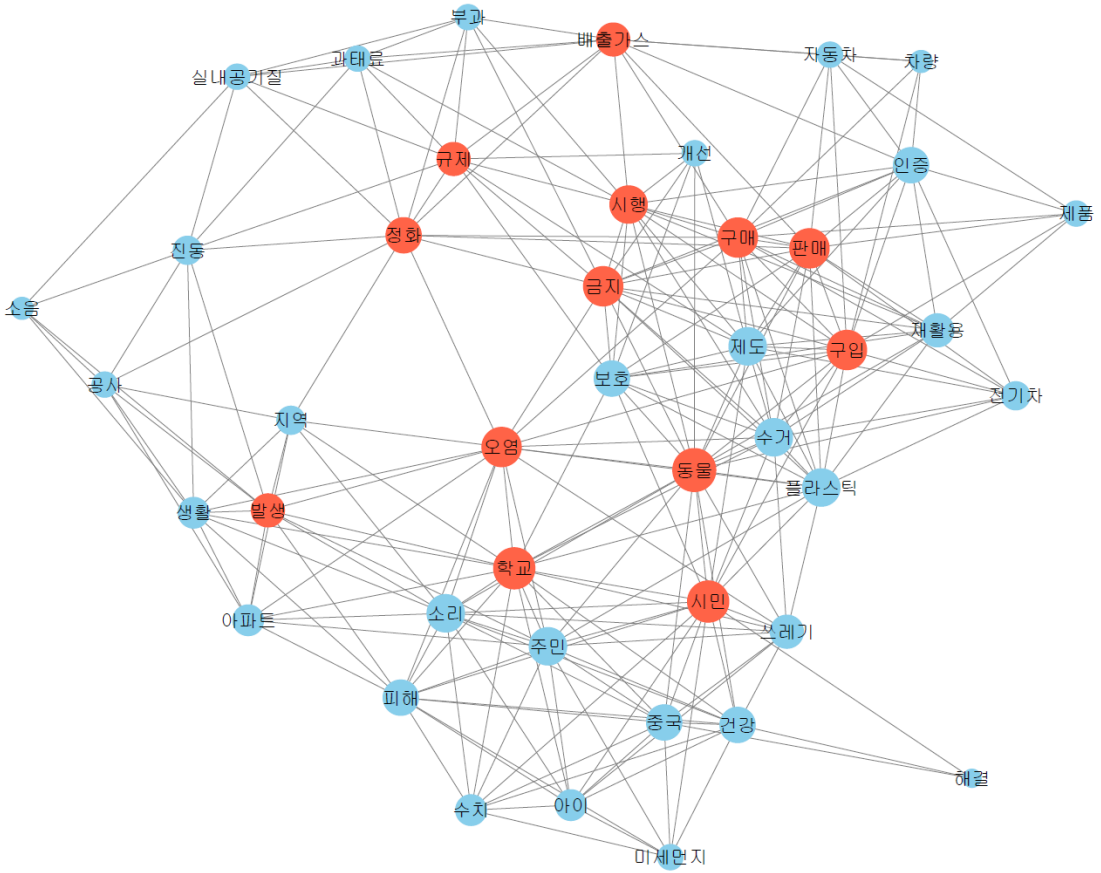
중심성 지표	강조 포인트
Degree centrality	직접적인 연결관계 한 꼭지점에 연결된 간선의 수
Closeness centrality	연결하는 경로길이 특정 꼭지점이 그를 제외한 다른 꼭지점과 얼마나 가까이에 있는지를 나타내는 지표
Betweeness centrality	노드들이 최단경로상 위치 여부 한 꼭지점의 중개중심성은 그 꼭지점을 제외한 다른 두 꼭지점을 잇는 최단거리에 해당 꼭지점이 얼마나 많이 등장하는지 빈도로
Eigenvector centrality	노드의 연결정도 중요한 꼭지점에 연결된 꼭지점일 수록 그 중요도가 높아지는 지표

Text Rank는? $TR(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} TR(V_j)$

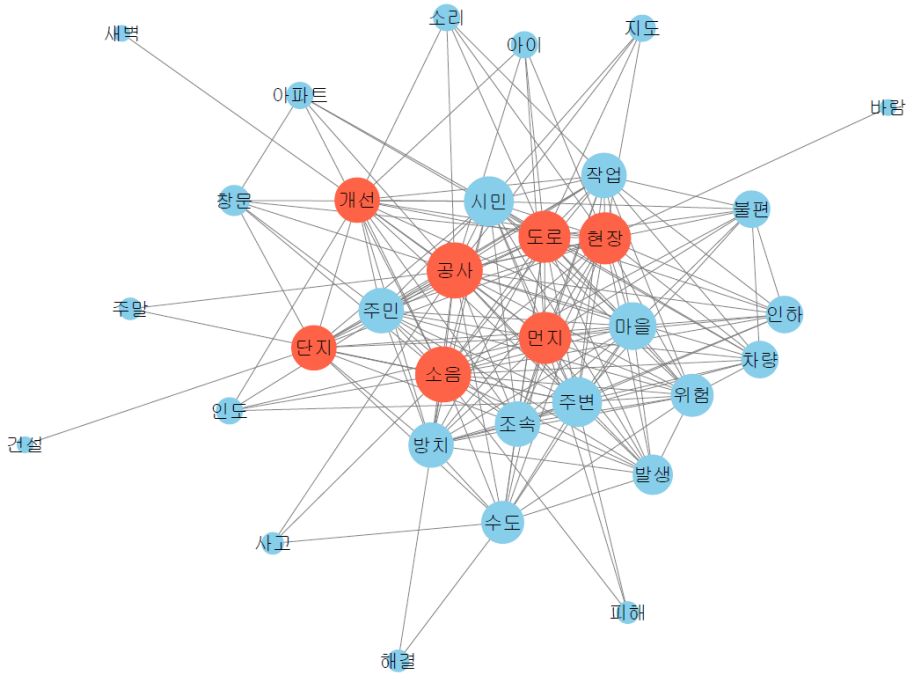
https://www.quora.com/What-are-the-limitations-of-graph-centrality-measures

https://ratsgo.github.io,

키워드 네트워크 시각화 사례



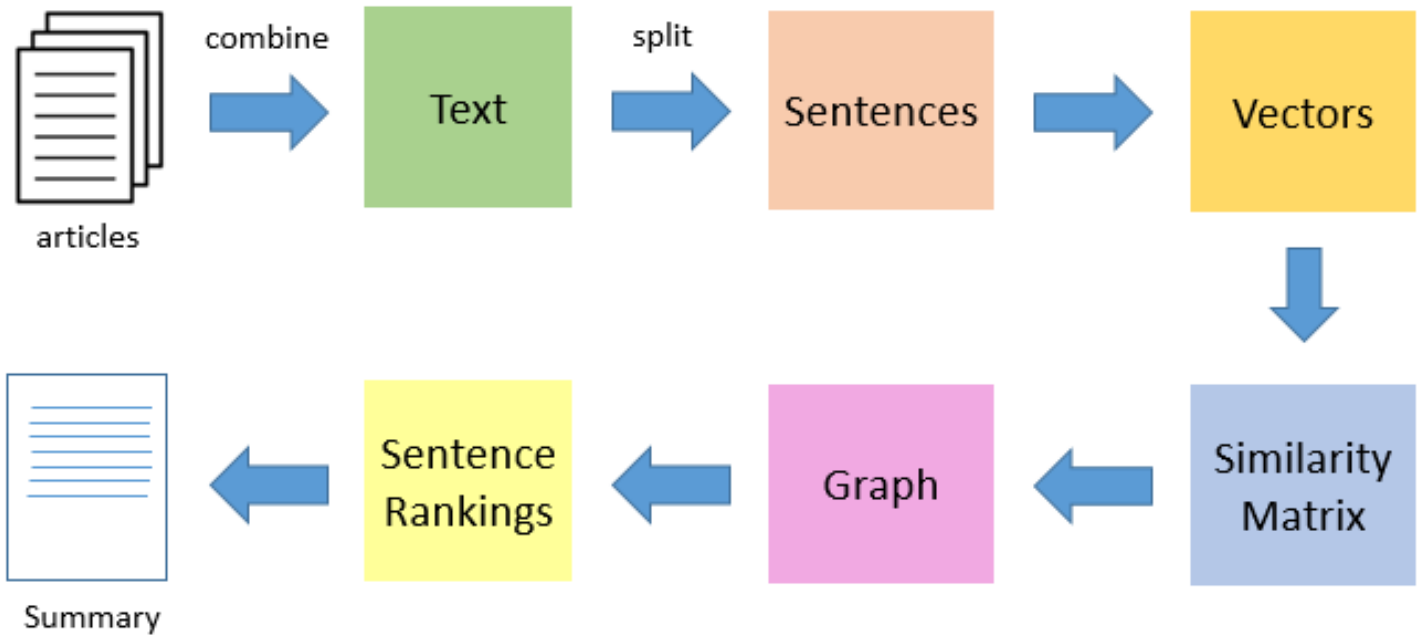
환경부 민원 분석



세종시 시민의창 민원분석

문서요약 알고리즘

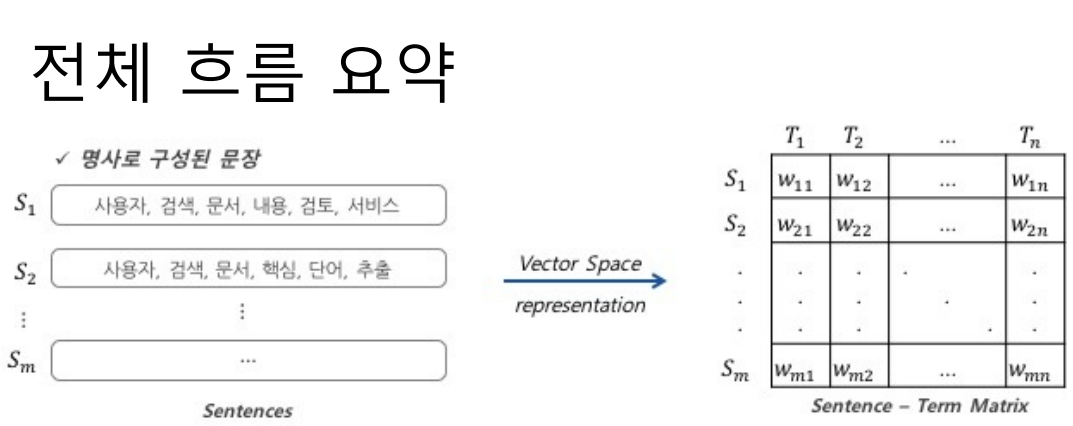
- 그래프 기반 문서요약 알고리즘의 흐름

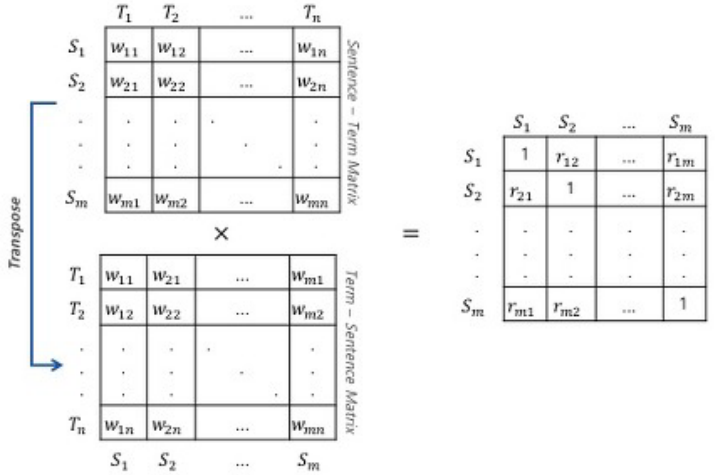


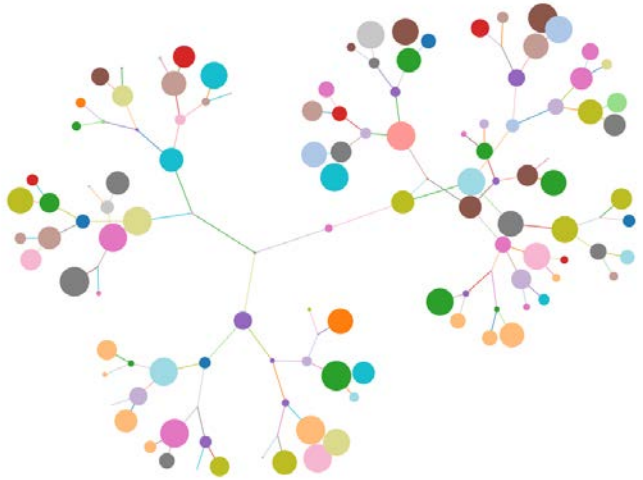
https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/

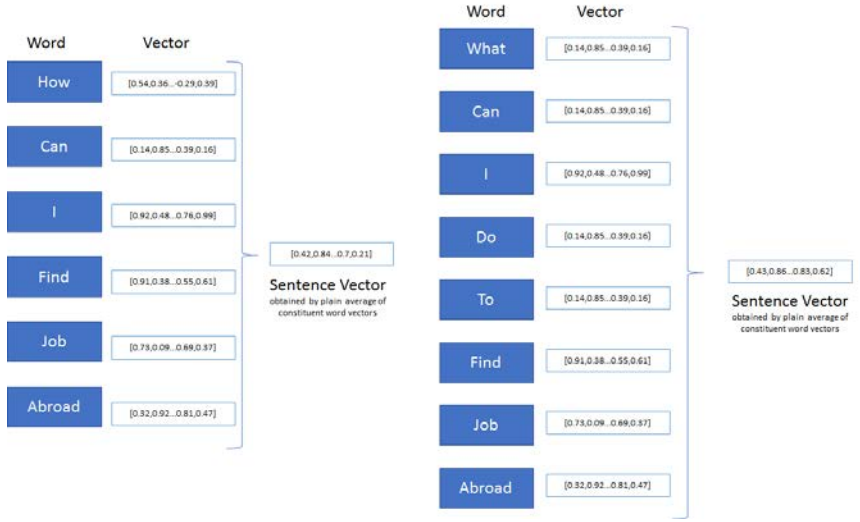
문서요약 알고리즘

- 전체 흐름 요약





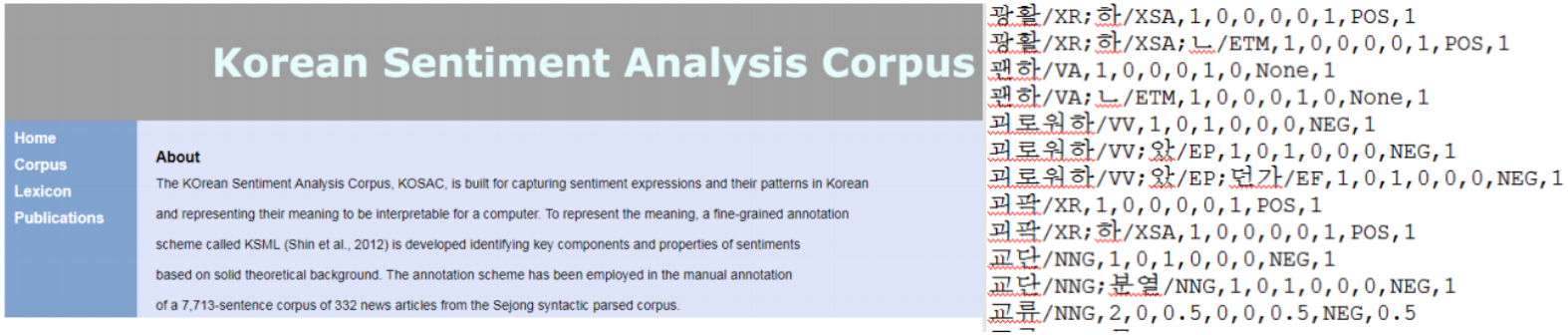




https://github.com/stanleyfok/sentence2vec
https://excelsior-cjh.tistory.com/93

감성분석

- 감성 사전 활용
  - 긍정 키워드 : + 점수
  - 부정 키워드 : - 점수



KOSAC 강성분석 코퍼스

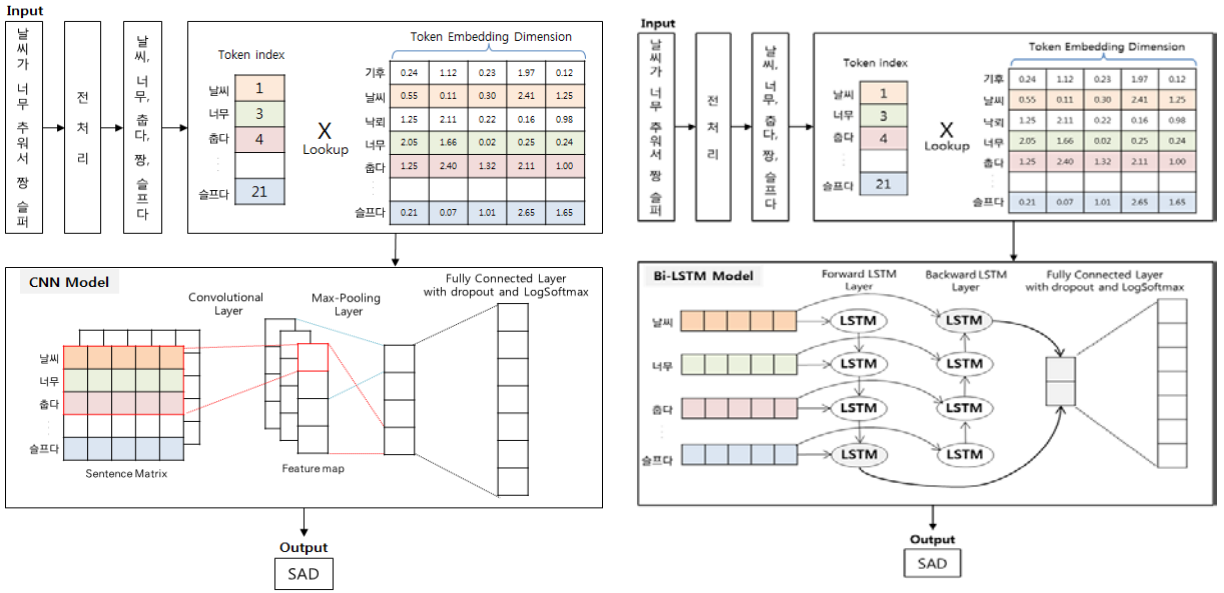
군산대학교 한국어 감성 사전

감성분석

- 학습에 기반한 방식

	환경	미세먼지	짜증	…	소음	상쾌	LABEL
Doc1	5	0	0		0	3	긍정
Doc2	2	3	4		3	0	부정
Doc2	1	0	1		1	2	긍정
…							긍정
DocN	2	4	3		2	1	부정

고전적 방식



딥러닝에 기반한 방식