

# News\_ight

뉴스 자동 분류 및 토픽 요약 서비스 |



2019-05-23

김현호  
노성문  
백승현  
안성윤  
#장정훈



프로젝트 개요

웹 페이지 시연

처리 단계

개선 방안 및 기대효과

참고문헌

QnA

Newsight



Media service

뉴스사이트(Newsight)는 뉴스 자동 분류 및 토픽 요약 서비스다.

**Owner** : Team Newsight

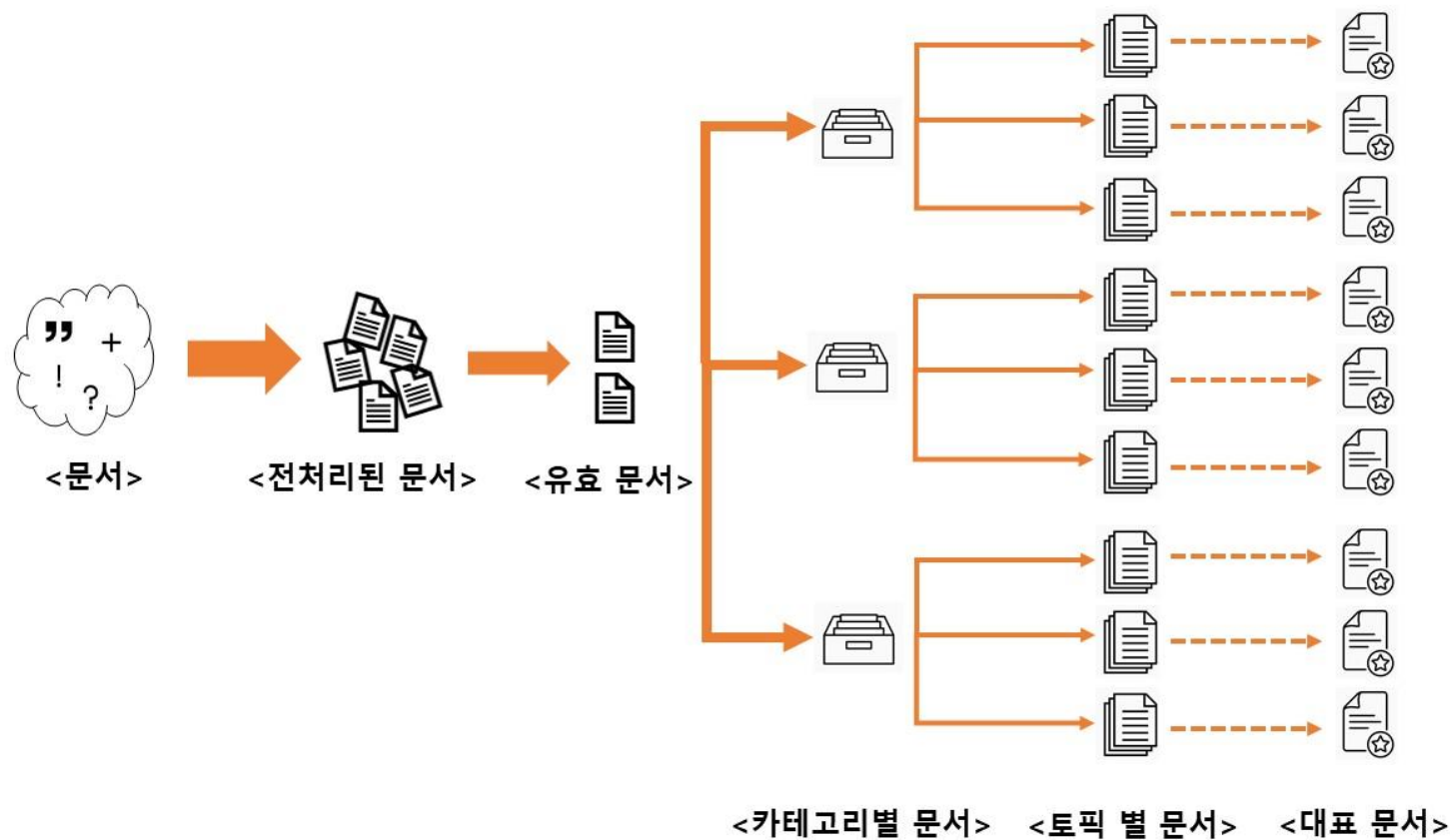
**Date Launched** : May 23, 2019

**Type of site** : Web search engine

**Written in** : Python

People also search for





- 2단계(**카테고리, 토픽**) 자동 분류
- **토픽**별 대표 기사 추출
- 시각화

**BIG Kinds**  
NEWS BIGDATA & ANALYSIS

사람중심 취업사이트  
**saramin**

**incruit**

**NAVER** **News\_ight**  
**JOBKOREA**

## AS-IS

- 미분류 카테고리
- 중복 기사
- 요약 필요



## TO-BE

- 자동 분류
- 대표기사 추출
- 시간순 정렬



프로젝트 개요

웹 페이지 시연

처리 단계

개선 방안 및 기대 효과

참고문헌

QnA

Newsight



Media service

뉴스사이트(Newsight)는 뉴스 자동 분류 및 토픽 요약 서비스다.

**Owner** : Team Newsight

**Date Launched** : May 23, 2019

**Type of site** : Web search engine

**Written in** : Python

People also search for





프로젝트 개요

웹 페이지 시연

처리 단계



개선 방안 및 기대 효과

참고문헌

QnA

Newsight



Media service

뉴스사이트(Newsight)는 뉴스 자동 분류 및 토픽 요약 서비스다.

**Owner** : Team Newsight

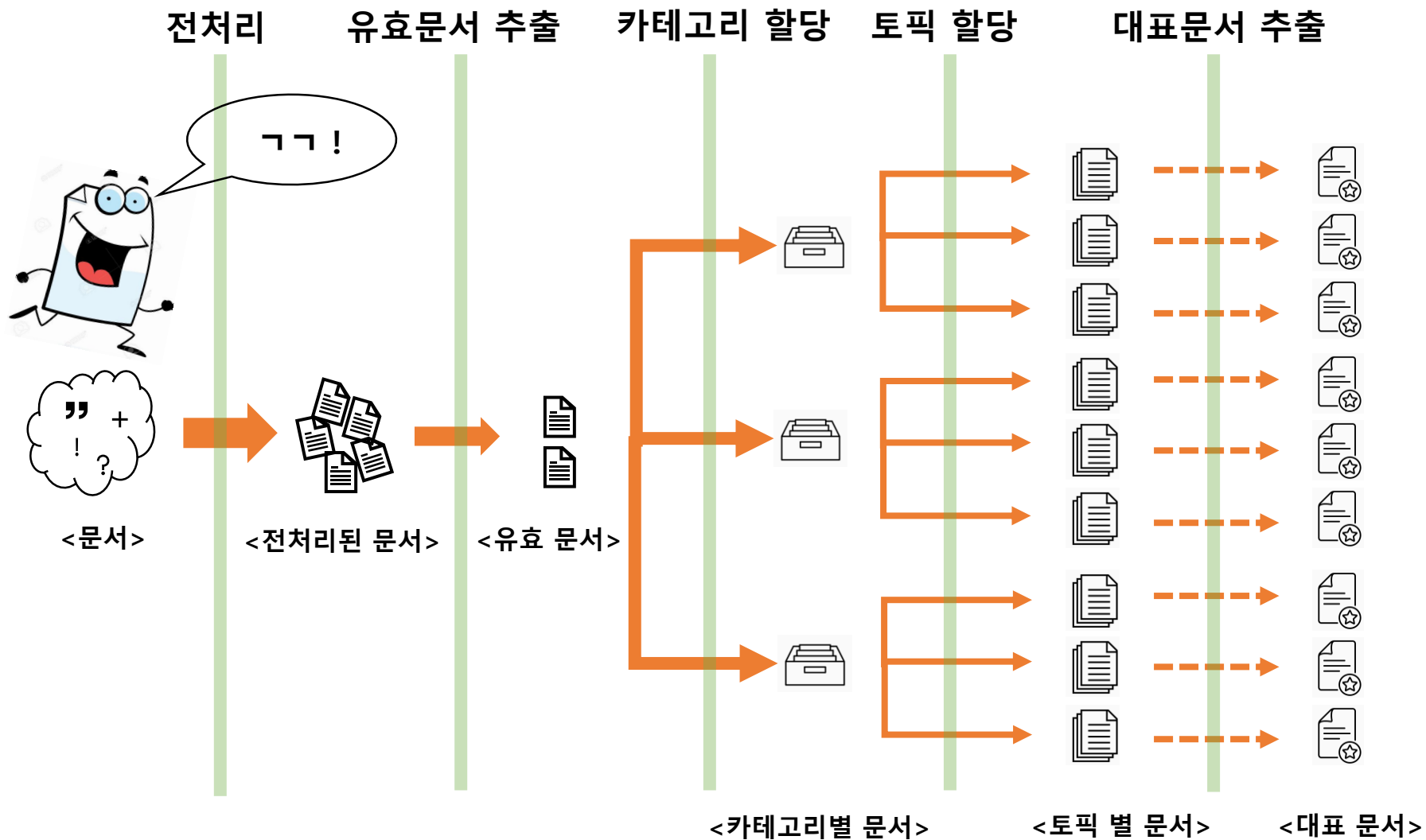
**Date Launched** : May 23, 2019

**Type of site** : Web search engine

**Written in** : Python

People also search for





- 전처리
- 유효문서 추출
- 카테고리 할당
- 토픽 할당
- 대표 문서 추출



- 검색어 : 미래에셋대우, 삼성전자, SK, 소카, 카카오
- 데이터 크기 : 검색어 당 10,000건

	date	title	content	category
0	2019-04-19	미래에셋대우-에트리홀딩스, 중소·벤처기업 성장지원 업무 협약	[이승렬 기자] \n\n미래에셋대우는 미래에셋센터원에서 에트리홀딩스(주)와 '...	[경제>산업_기업]
1	2019-04-19	미래에셋대우, 에트리홀딩스와 '중소·벤처기업 성장 지원' MOU 체결	미래에셋대우는 지난 18일 서울 중구 미래에셋센터원에서 에트리홀딩스와 '중소·벤처 ...	[경제>산업_기업, 경제>취업_창업]
2	2019-04-19	홍콩 오피스 빌딩에 2800억 투자한 미래에셋대우	[디지털타임스 차현정 기자] 미래에셋대우는 18일 홍콩 CBD2(이스트 카우룽)에 ...	[경제>국제경제, 경제>금융_재테크, 경제>산업_기업]
3	2019-04-18	홍콩 오피스 빌딩에 2800억 투자한 미래에셋대우	[디지털타임스 차현정 기자] 미래에셋대우는 18일 홍콩 CBD2(이스트 카우룽)에 ...	[경제>국제경제, 경제>금융_재테크, 경제>산업_기업]

## Bigkinds



Media service

빅카인즈(Bigkinds)는 한국언론진흥재단의 뉴스데이터 분석 서비스다.

**Owner** : 한국언론진흥재단

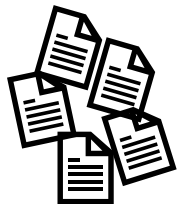
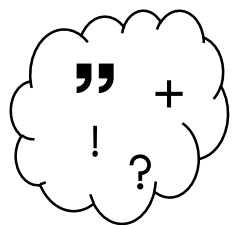
**Year Launched** : 2017

**Type of site** : Web search engine

People also search for

News\_ight NAVER Daum





[검사, 확인, 검출...]

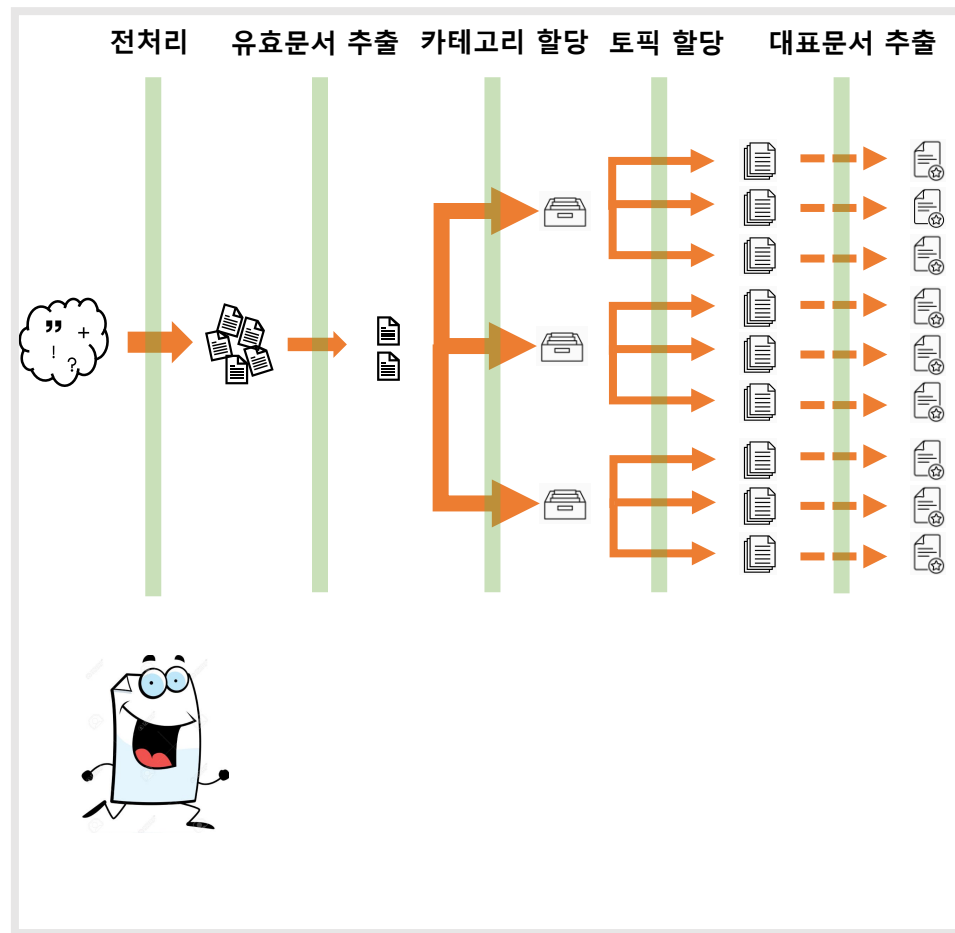


확인

'뉴스ایت 넘모 좋은  
사이트에요!'



[뉴스ایت, 사이트]





## 미래에셋대우 홍콩법인, 에미레이트 항공기 2대 매각

기사입력 2019-05-20 10:54:49. 폰트 + -

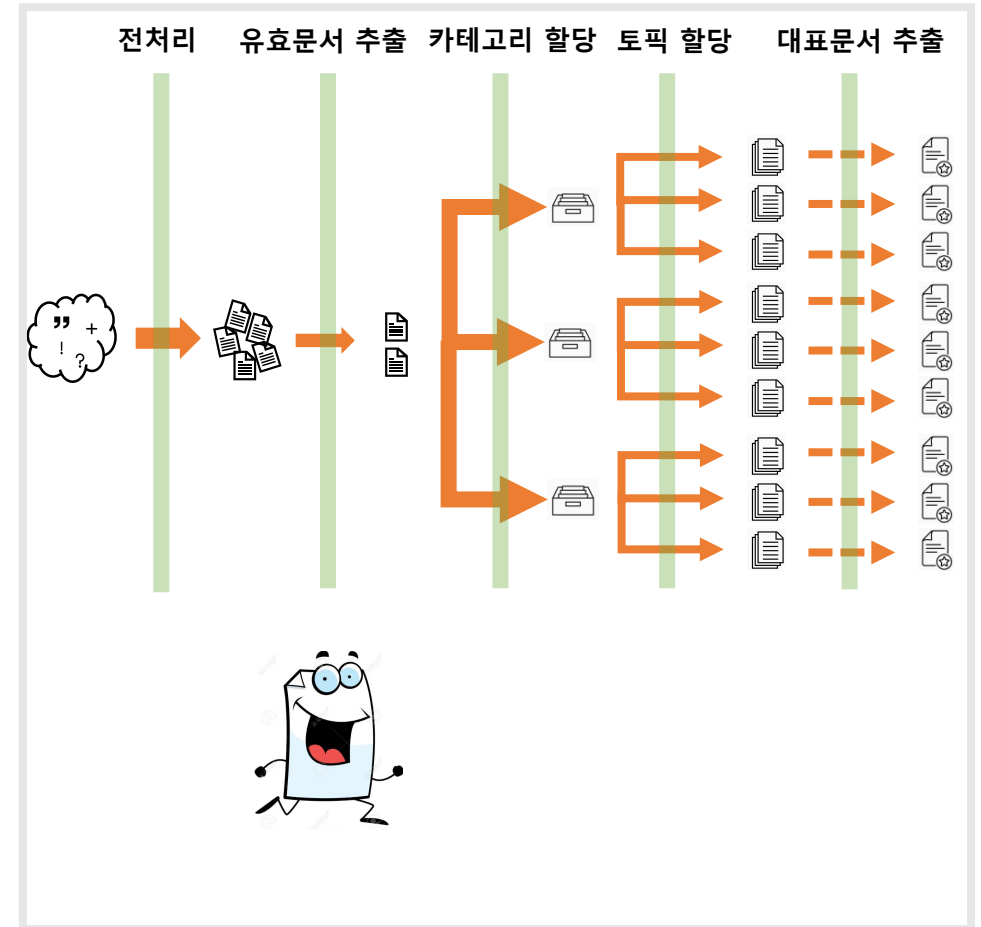
미래에셋대우 홍콩법인은 2015년부터 보유한 에미레이트항공의 B777-300ER 항공기 2대를 일 본계 리스사에 매각 완료했다고 20일 밝혔다.

미래에셋대우는 이번 항공기 매각으로 15% 이상의 수익을 낼 것으로 기대했다.

2015년 미래에셋대우 홍콩법인은 에미레이트항공이 인도받아 사용하던 B777-300ER 항공기 를 매입해 재임대하는 "세일즈 앤드 리스백" 계약을 체결했다.

당시 외국계 은행과 국내 연기금을 비롯한 기관투자자들이 약 3억2000만달러 규모의 매입자 금 조달에 공동 참여했다. 앞서 미래에셋대우 홍콩법인은 2017년에도 핀에어 항공기 매각에 성공했다.

미래에셋대우 홍콩법인 관계자는 "이번 항공기 매각도 성공적으로 마무리해 항공기 금융시장 에서 의미 있는 트랙 레코드(실적)를 확보하게 됐다"고 말했다. 임성엽기자



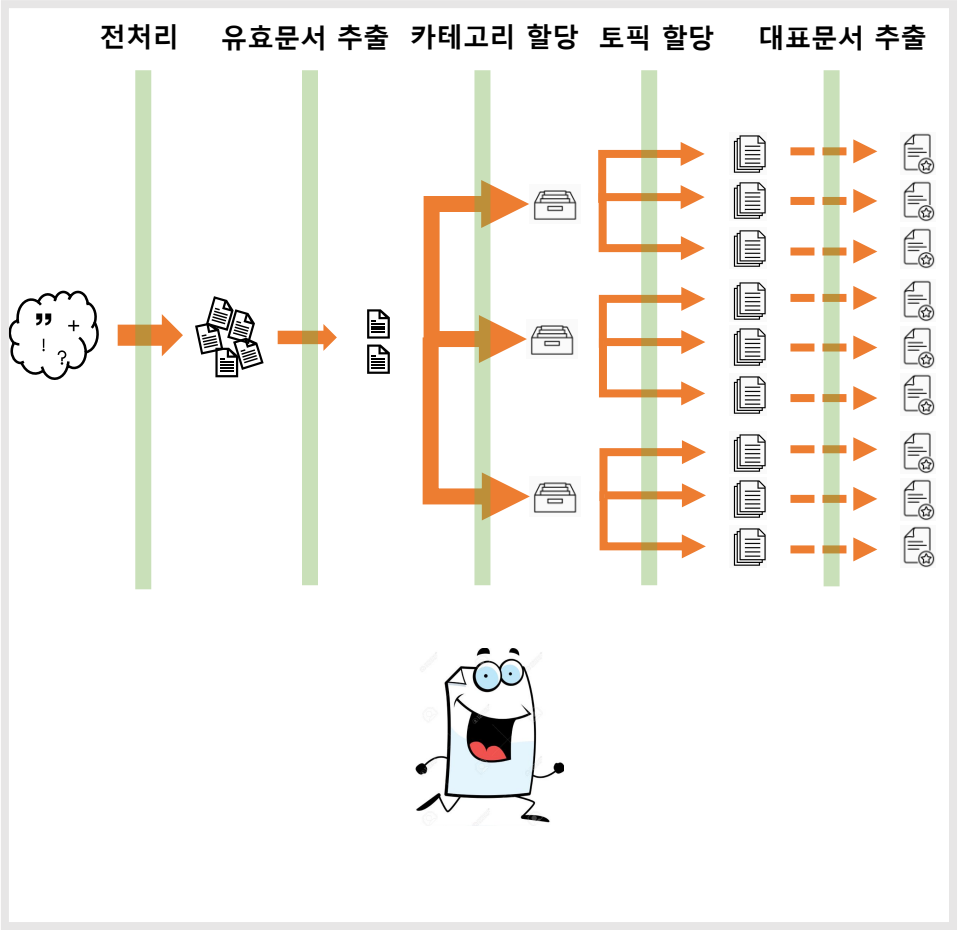
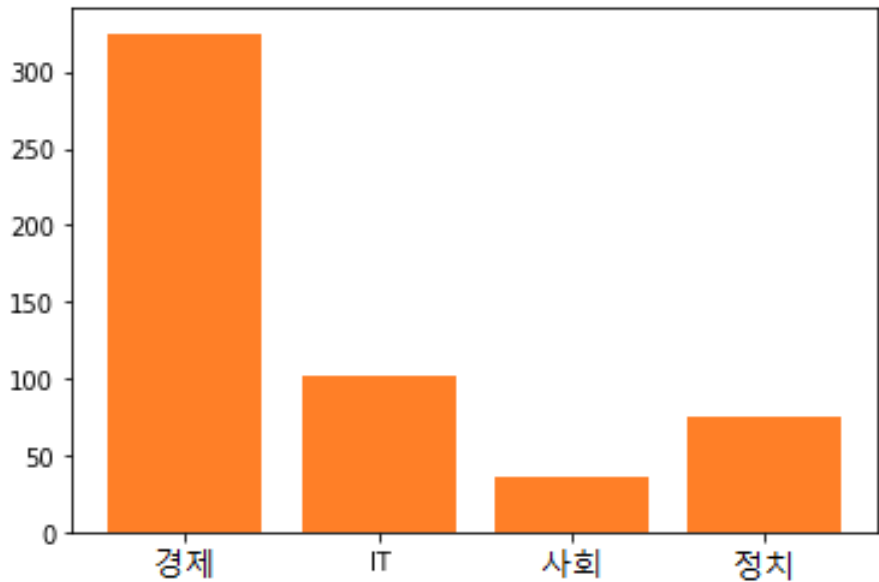


- 경제>증권\_증시, 경제>서비스\_쇼핑, 경제>금융\_재테크
- IT\_과학>인터넷\_SNS, 경제>산업\_기업
- 지역>전북, 경제>증권\_증시, 지역>대구

→ 경제

→ IT

→ 지역





- 토픽 수를 몰라요
- 실시간으로 서비스하고 싶어요
- 뭘로 검색하던 분류가 잘됐으면해요



- 군집 수 지정 불필요
- 빠른 처리속도
- 최소 파라미터



OPTICS

DBSCAN

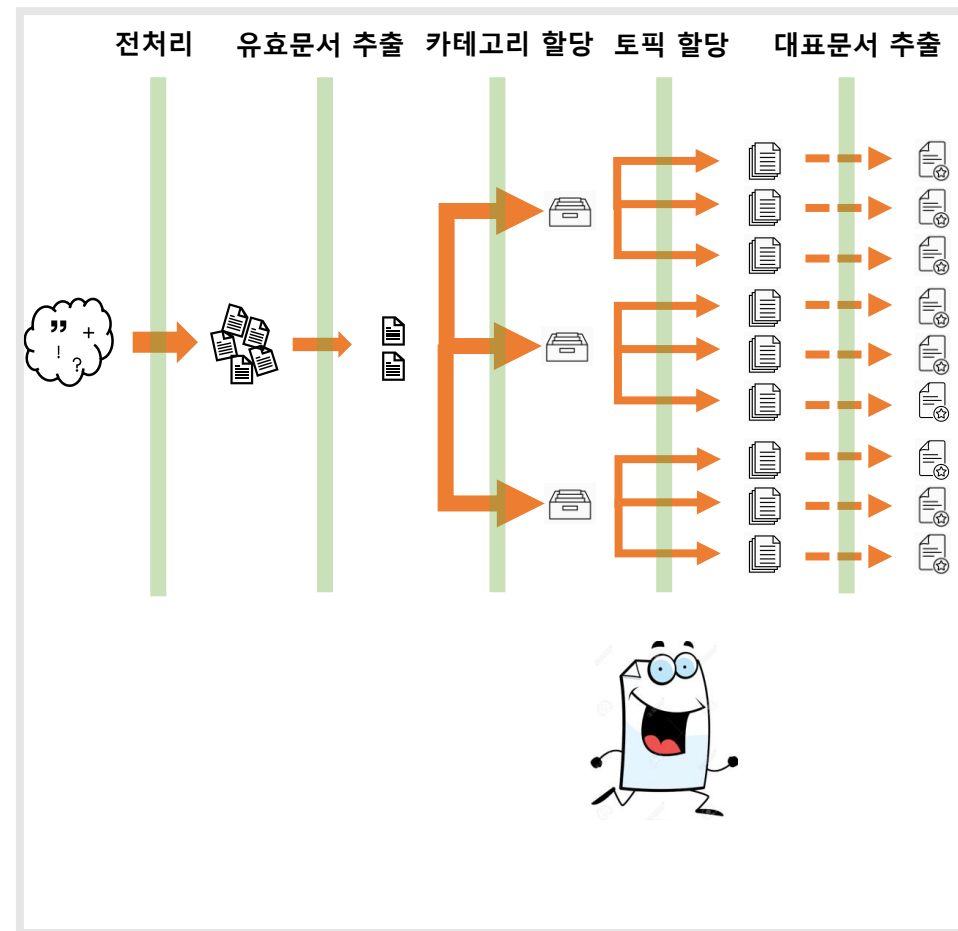
COMMUNITY DETECTION

LDA

Spherical KMeans

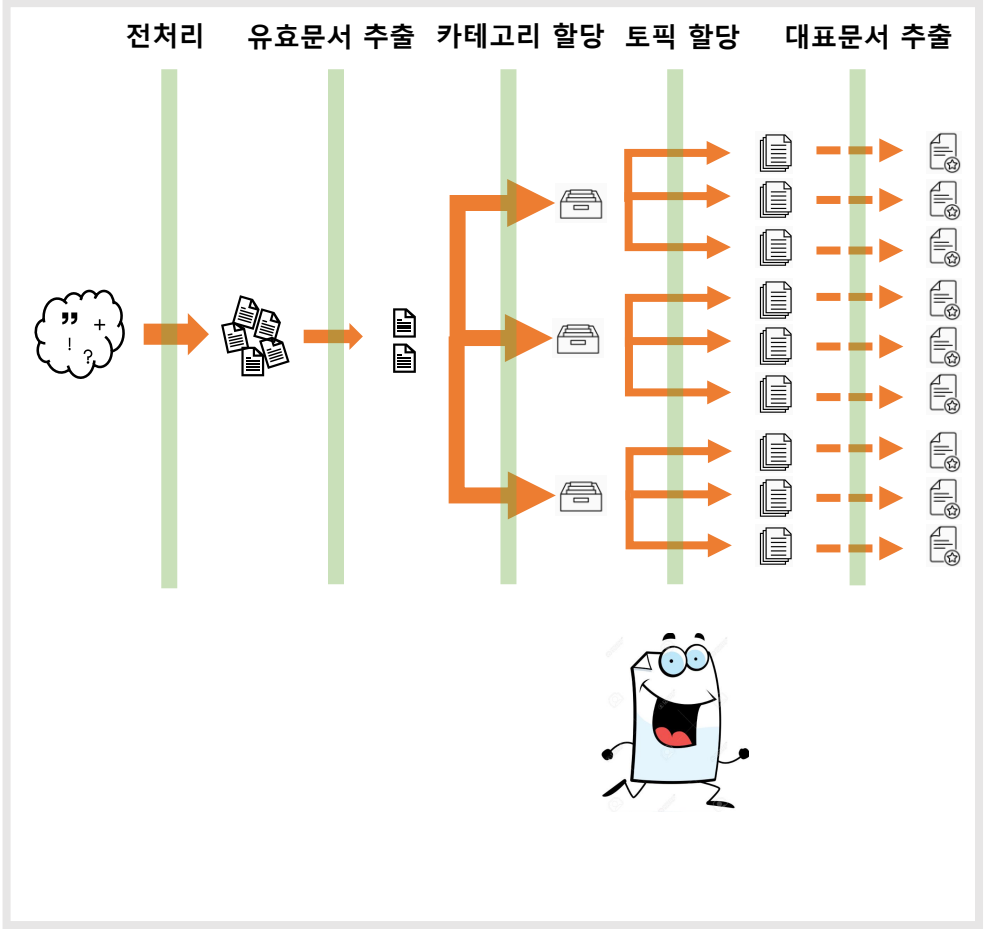
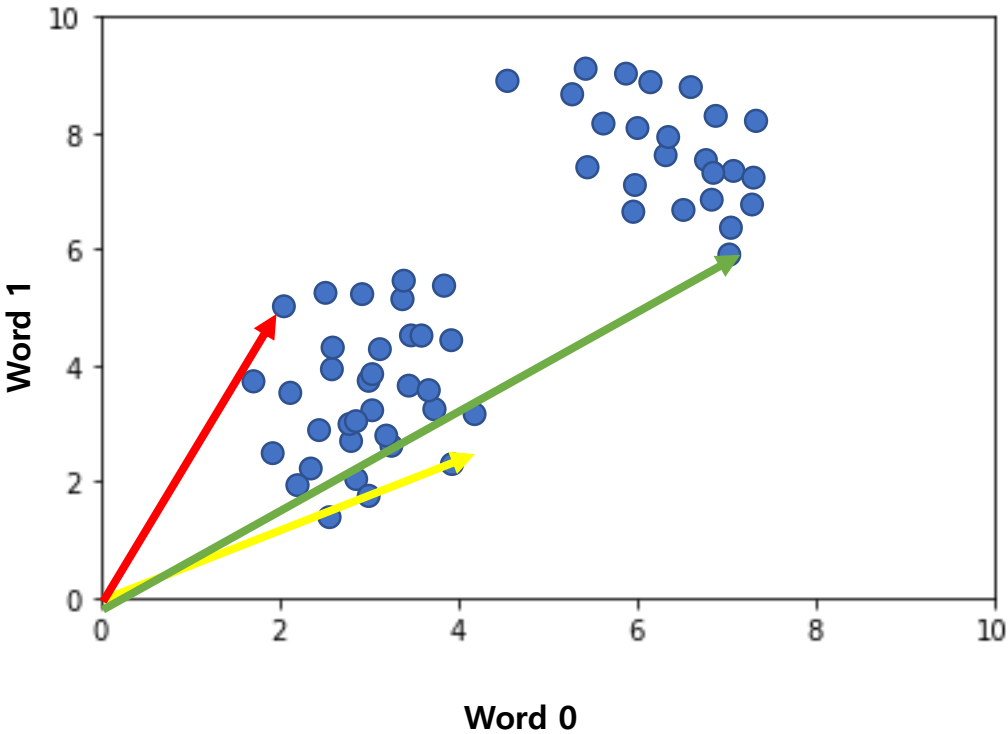
LSA

KMeans



KMeans

Spherical KMeans

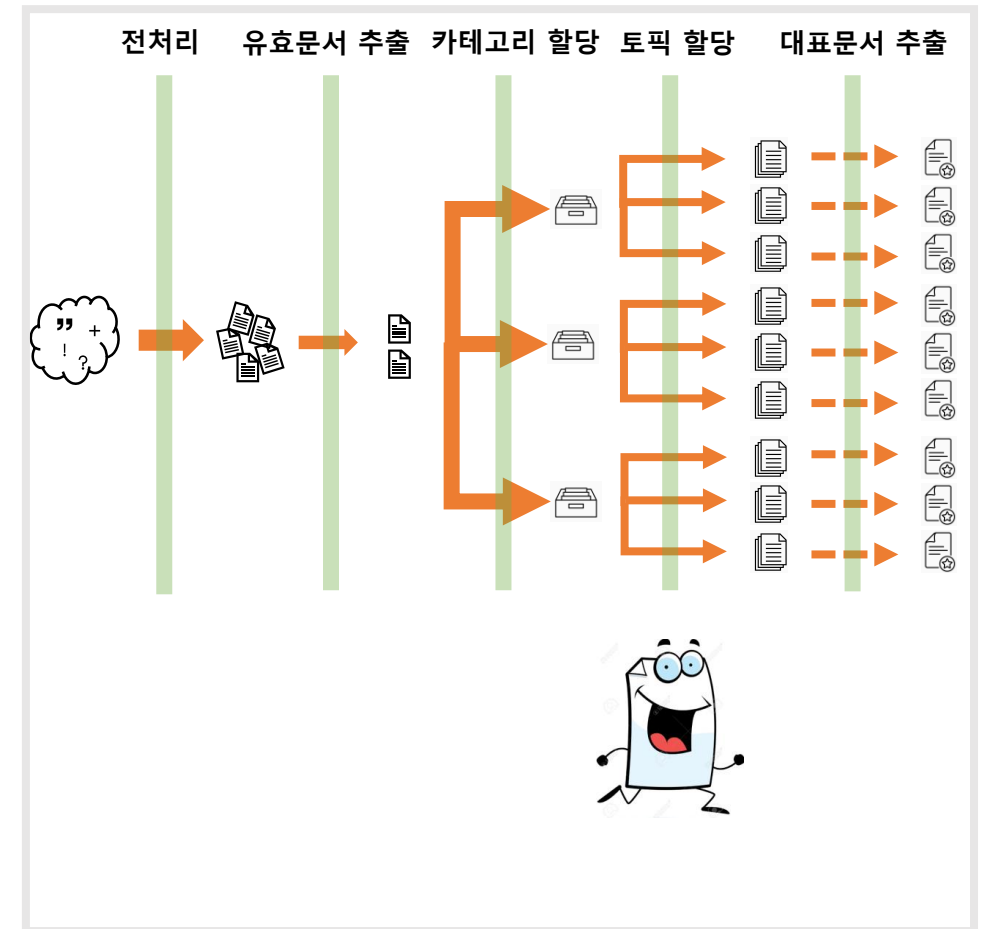
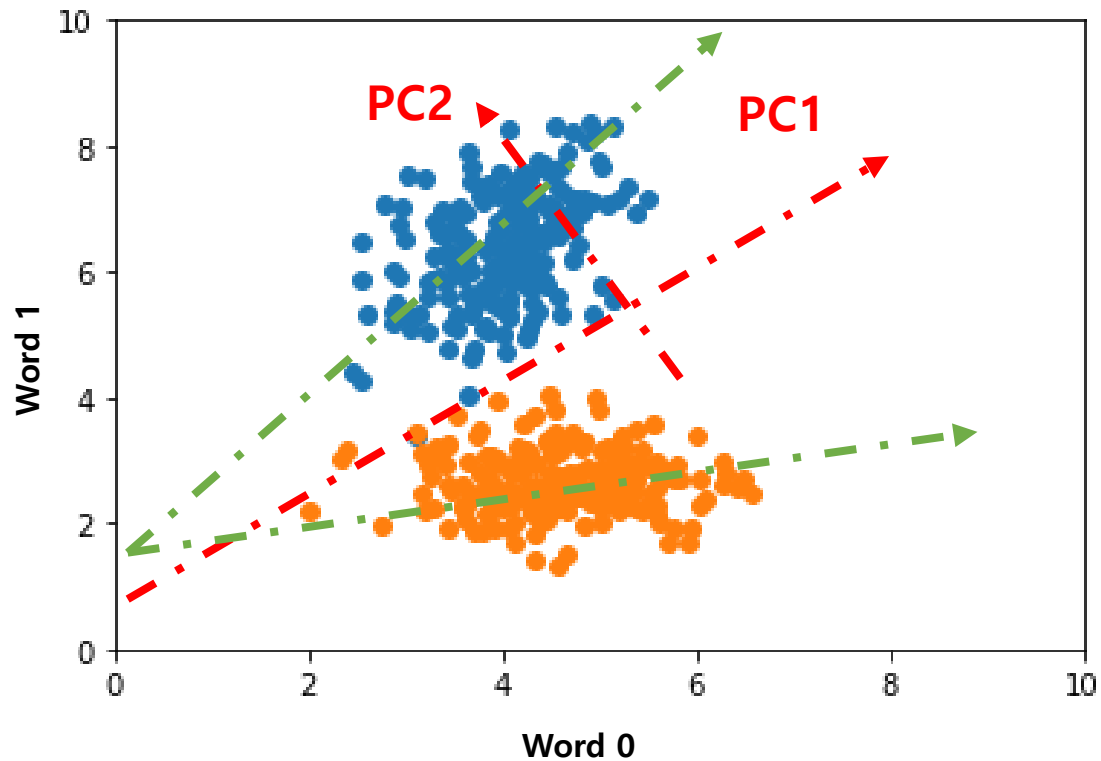




## Latent Semantic Analysis (LSA)

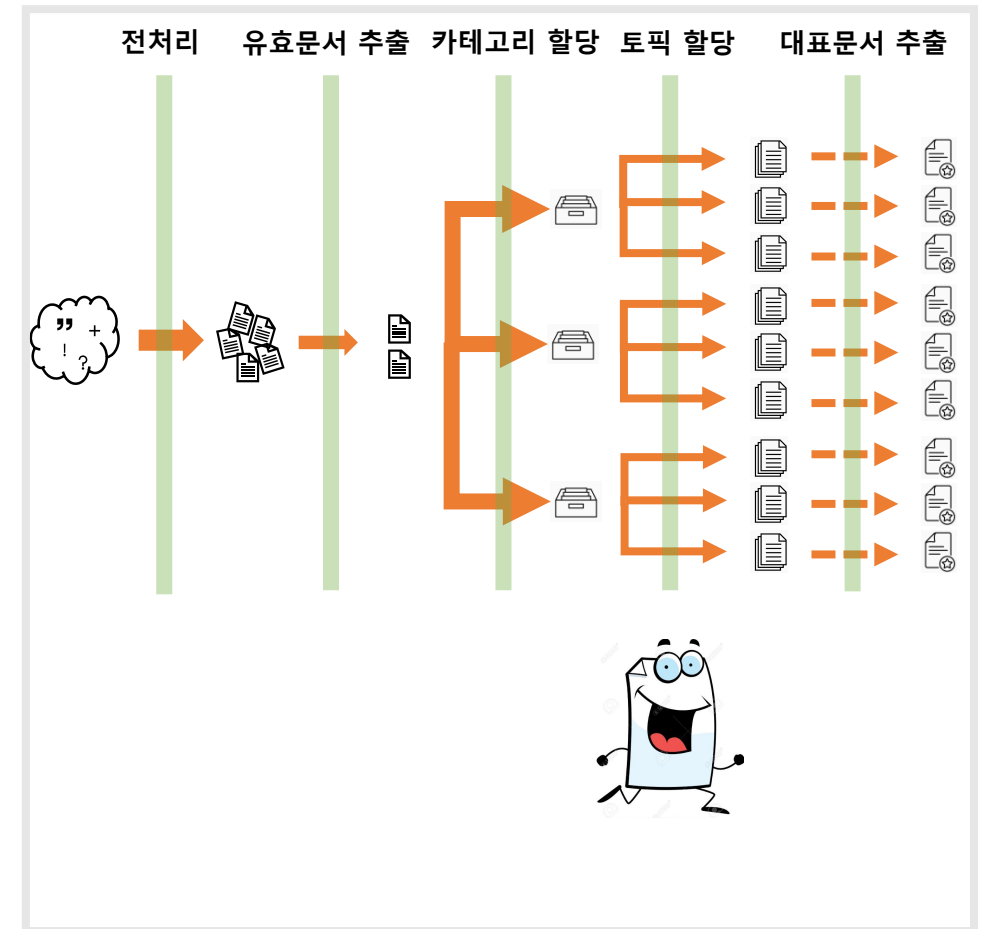
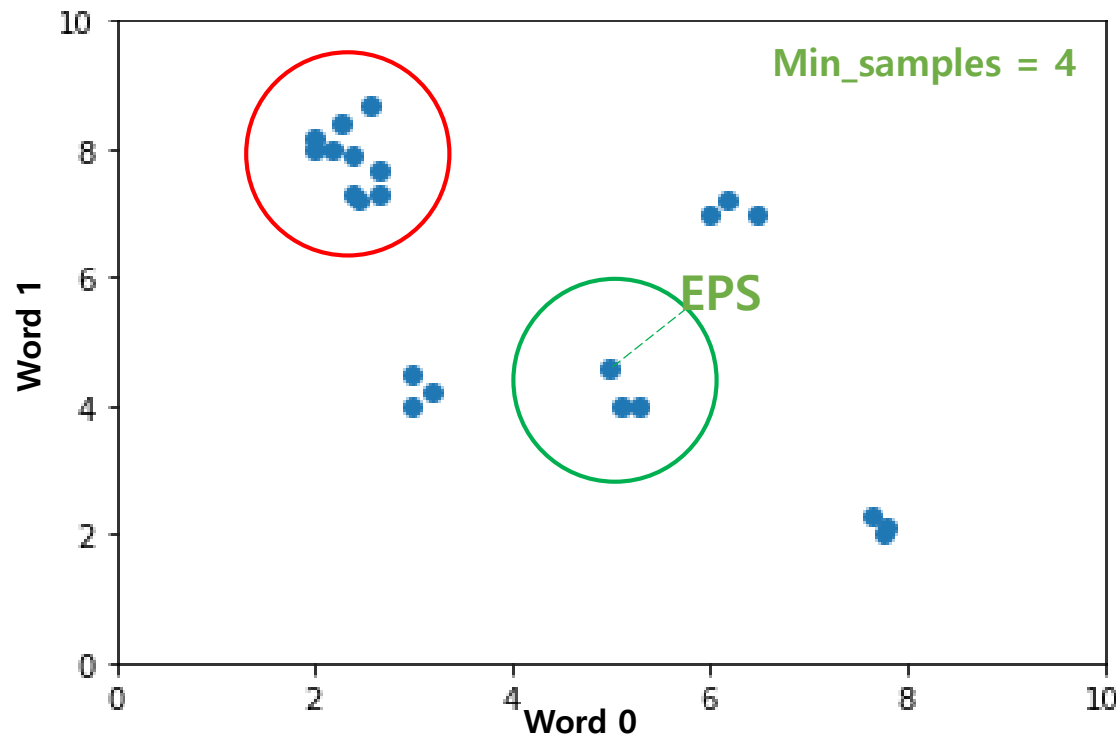


## Latent Dirichlet Allocation (LDA)



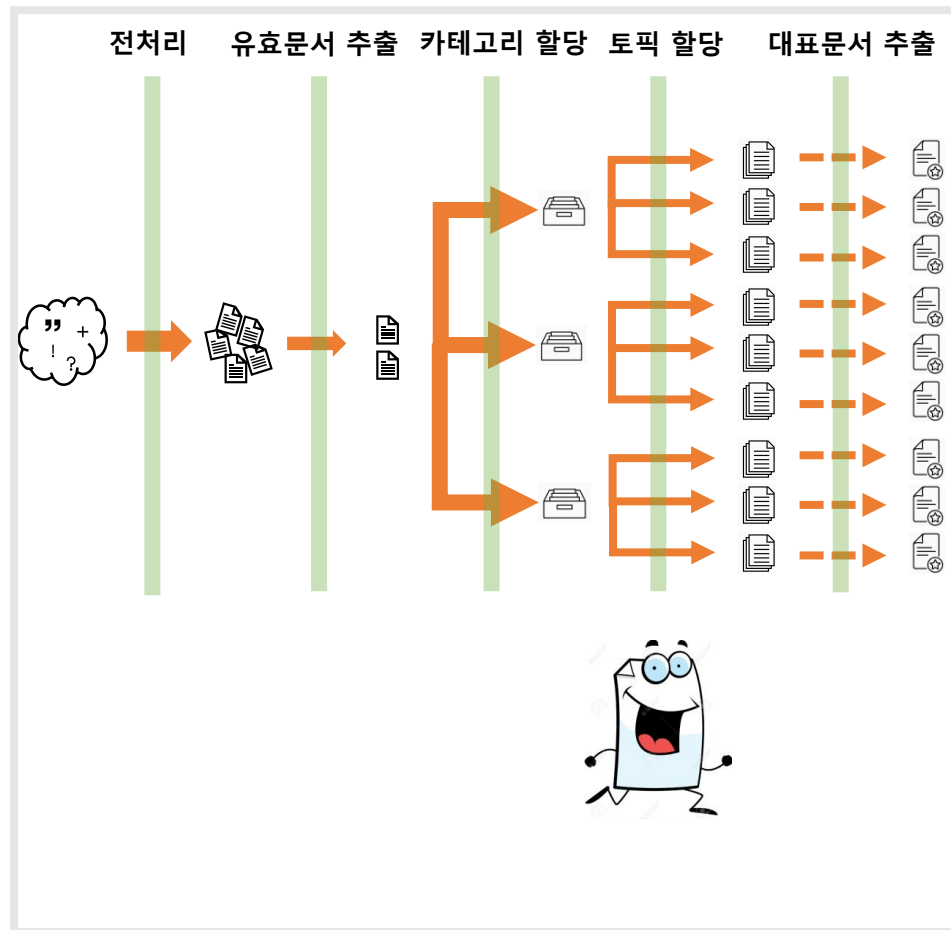
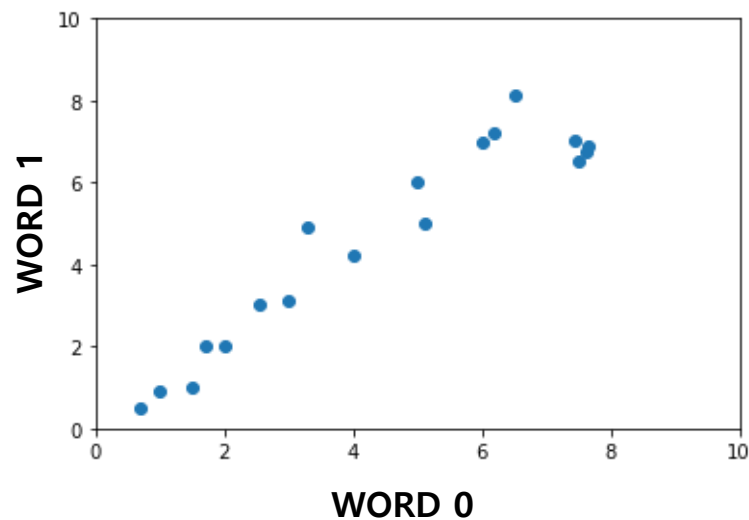
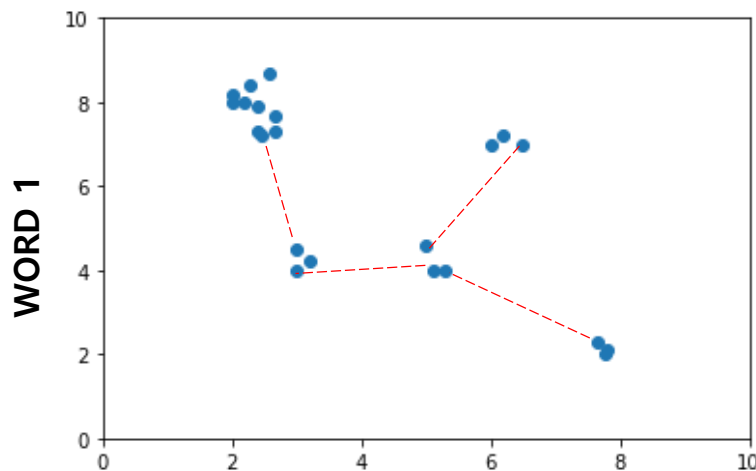
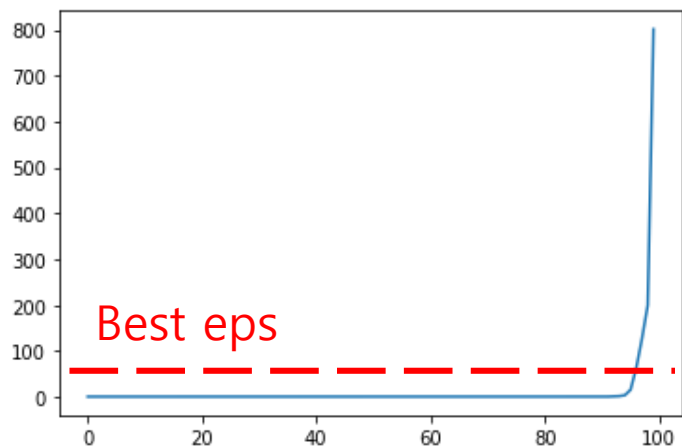
DBSCAN

- Epsilon, Min\_samples
- 군집수를 정할 필요가 없음
- 이상치를 제거





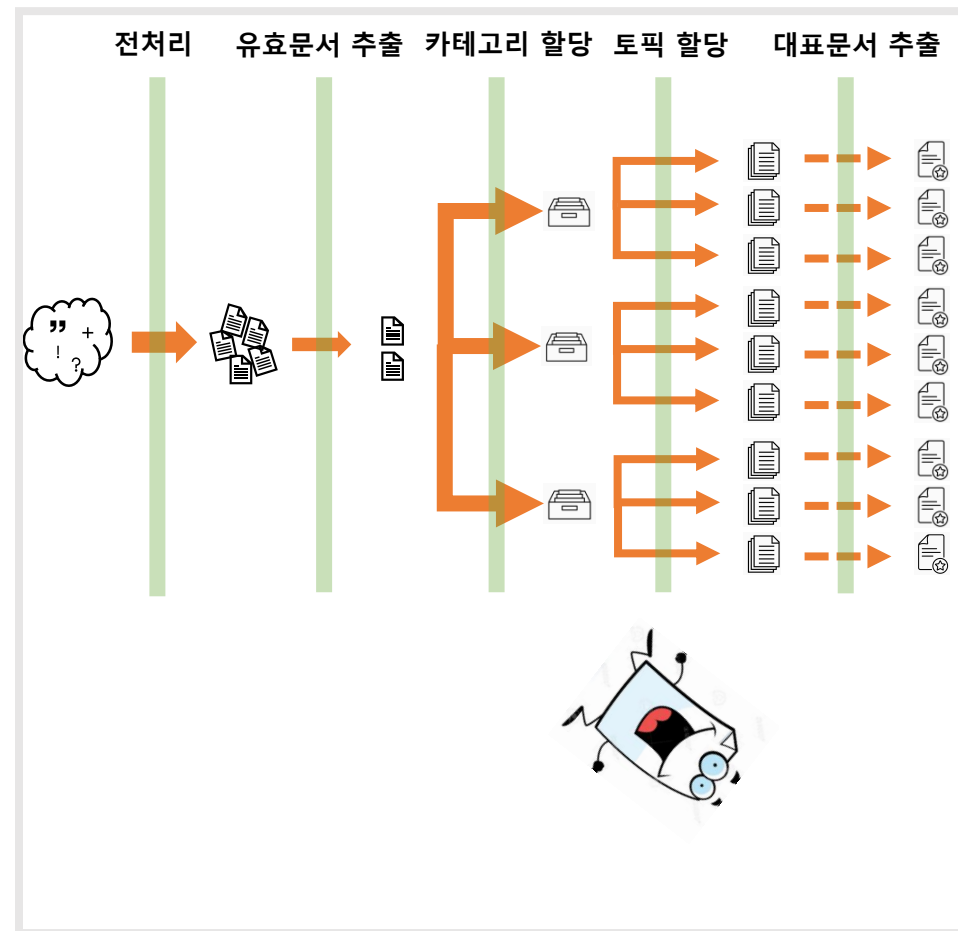
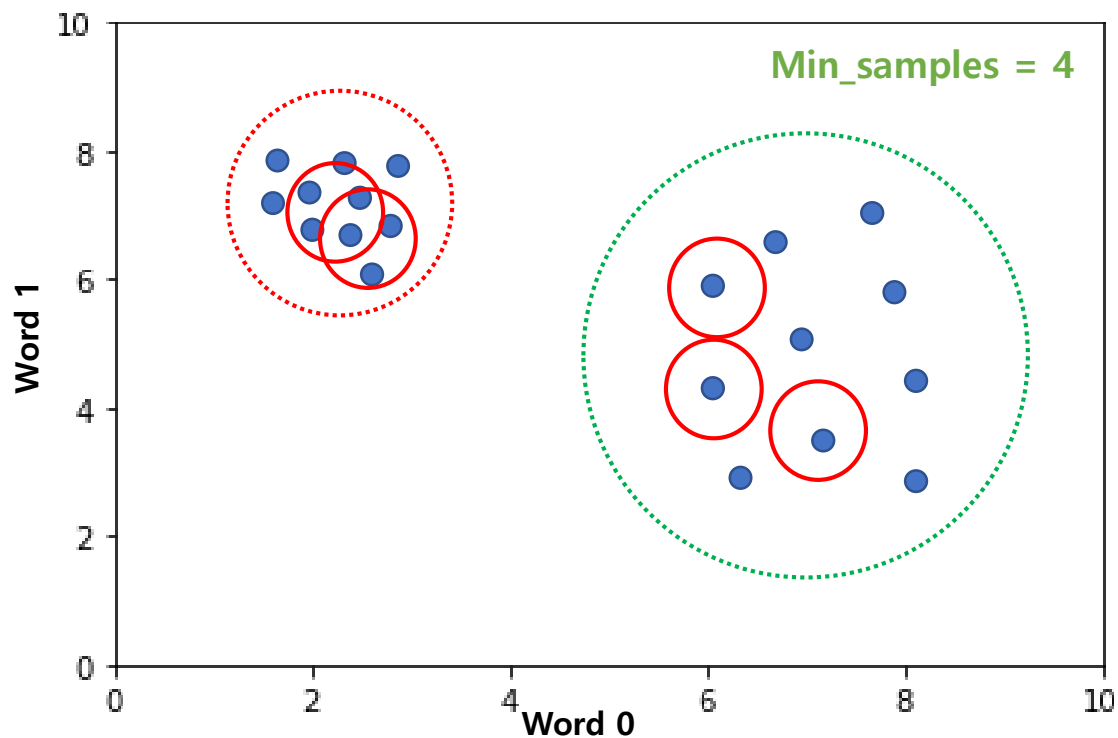
## DBSCAN





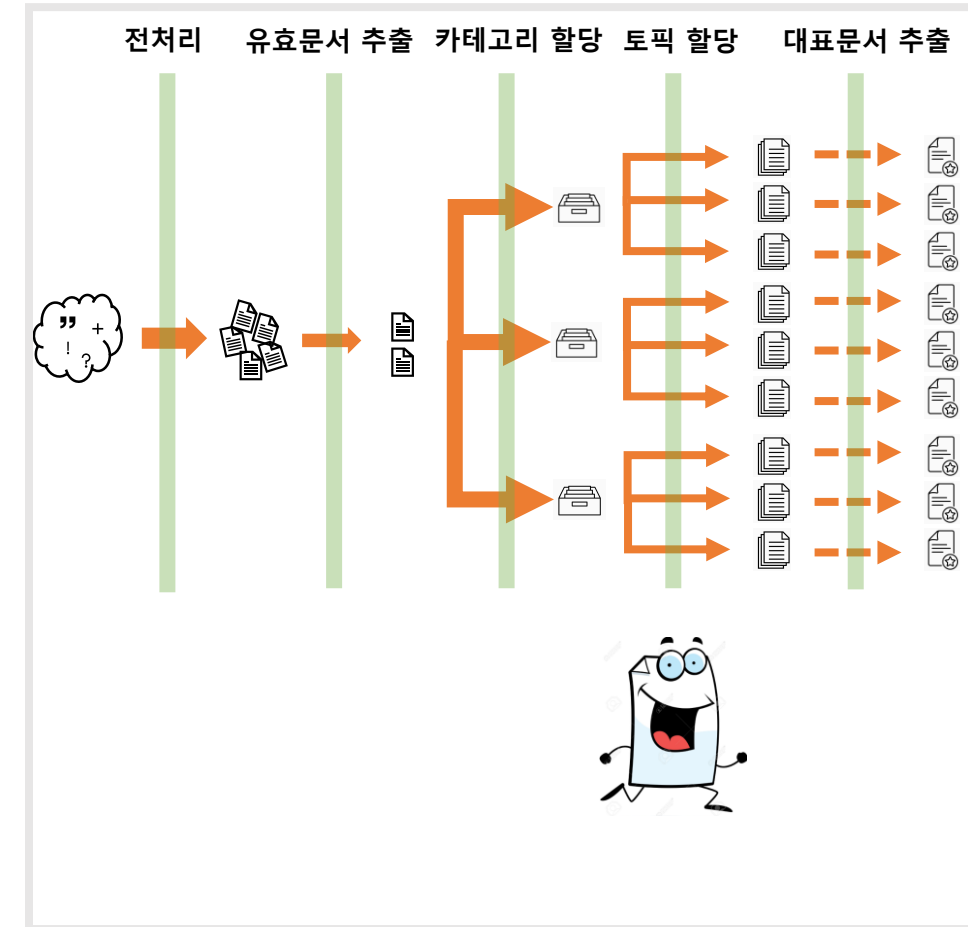
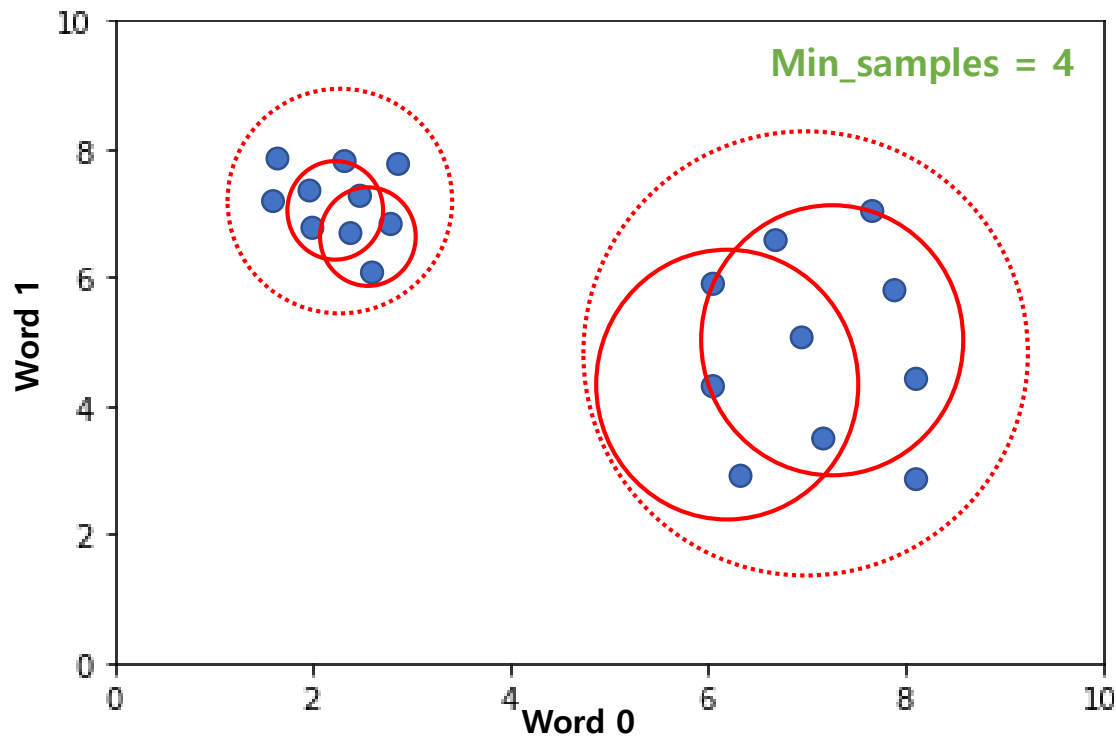
OPTICS

- Epsilon, 군집수를 정할 필요가 없음
- **부분마다 Epsilon을 다르게 설정**
- **공간의 밀도가 다른 군집들이 존재해도 군집화할 수 있다.**



OPTICS

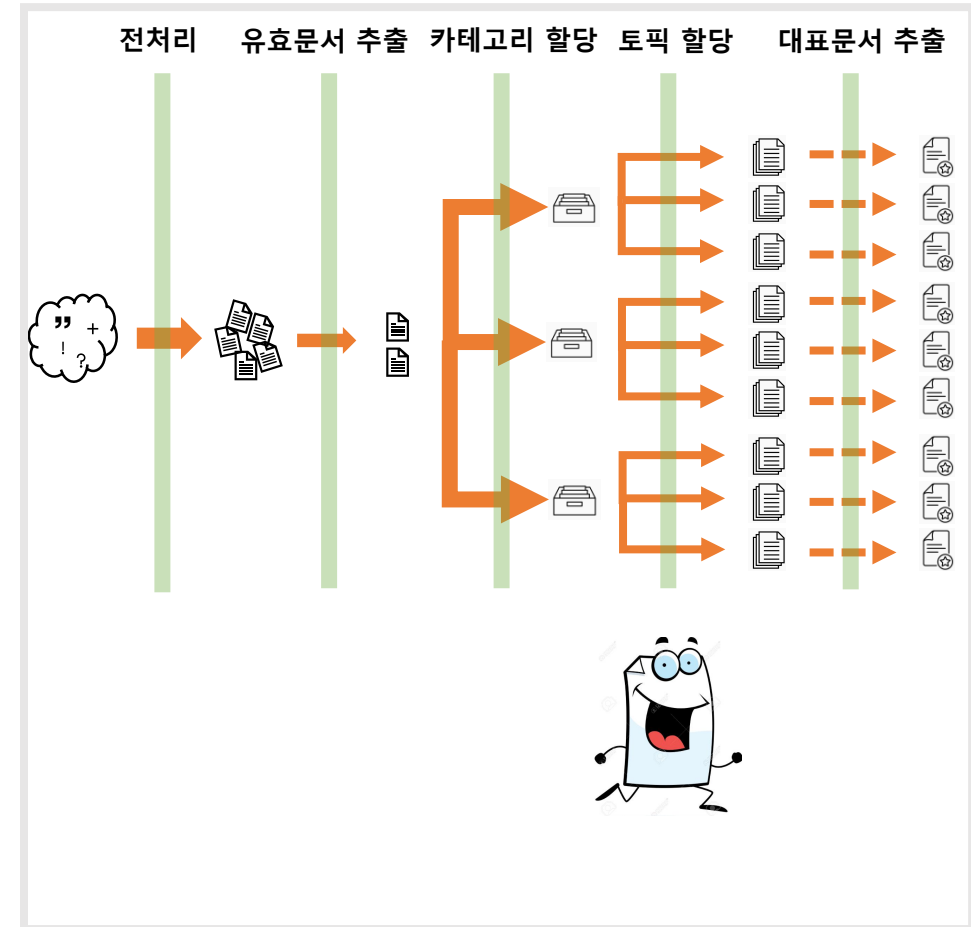
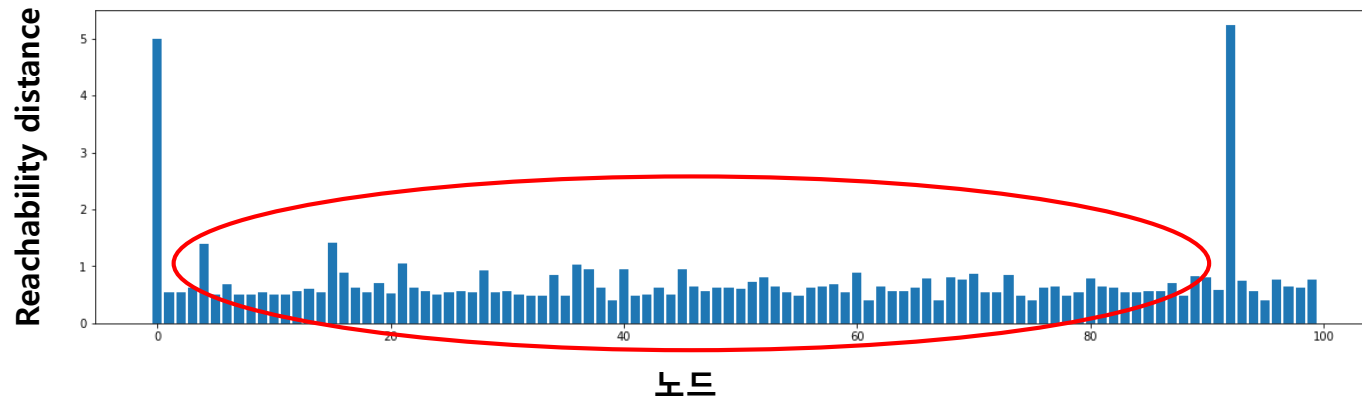
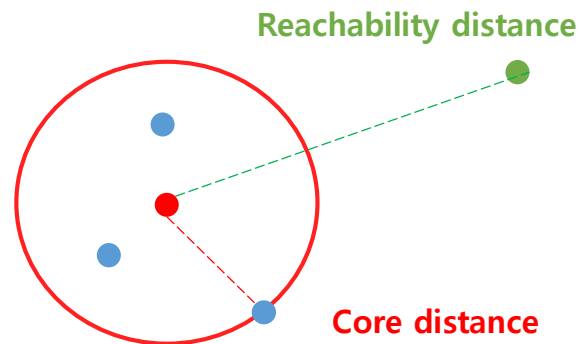
- Epsilon, 군집수를 정할 필요가 없음
- **부분마다 Epsilon을 다르게 설정**
- **공간의 밀도가 다른 군집들이 존재해도 군집화할 수 있다.**

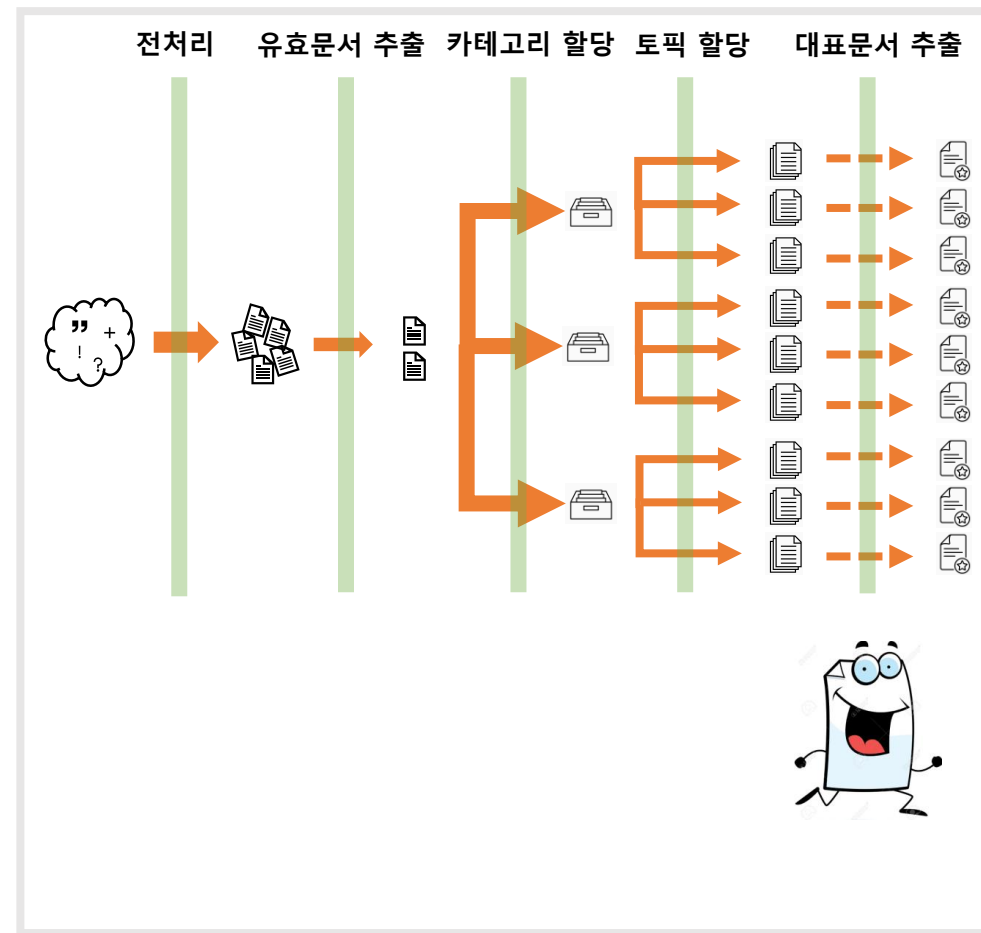
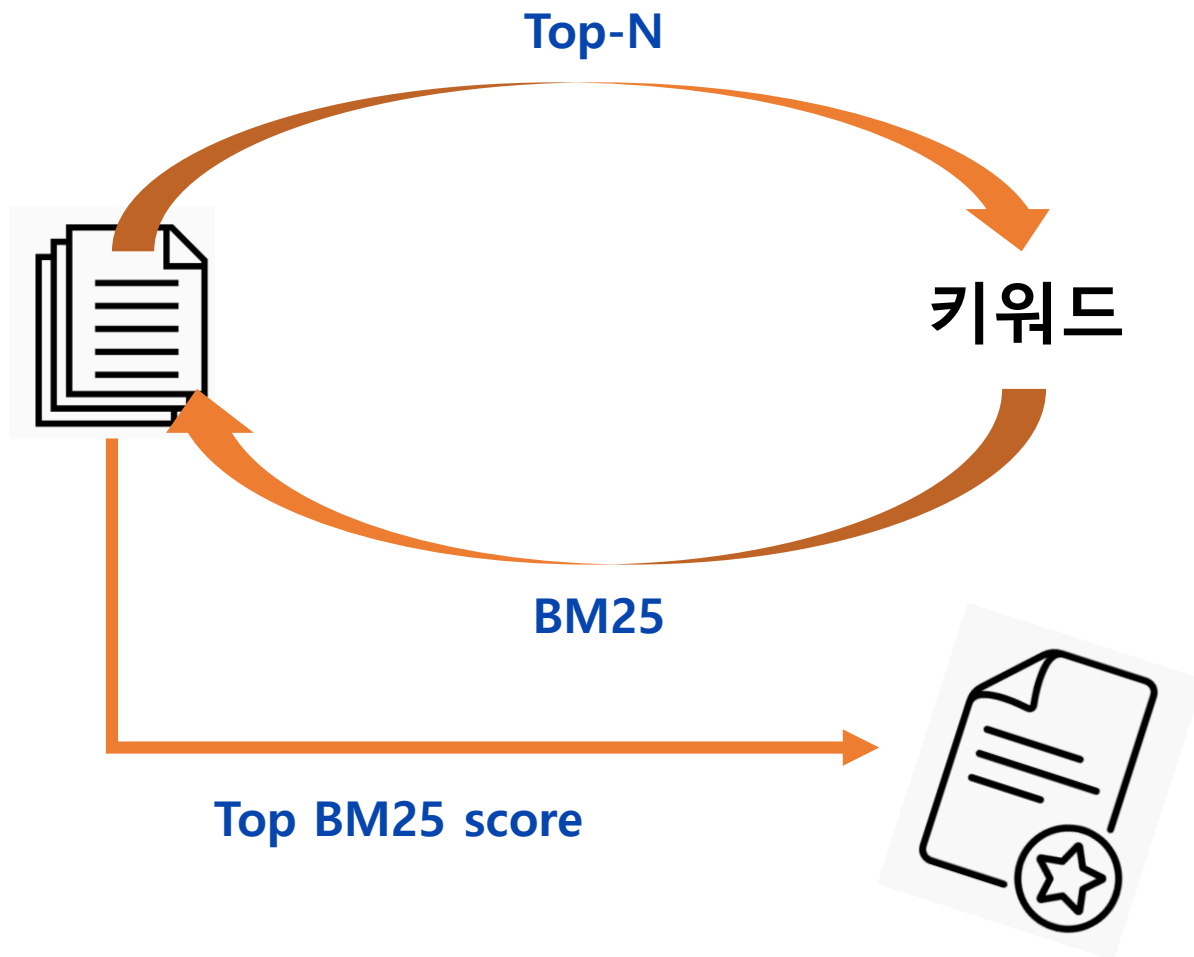


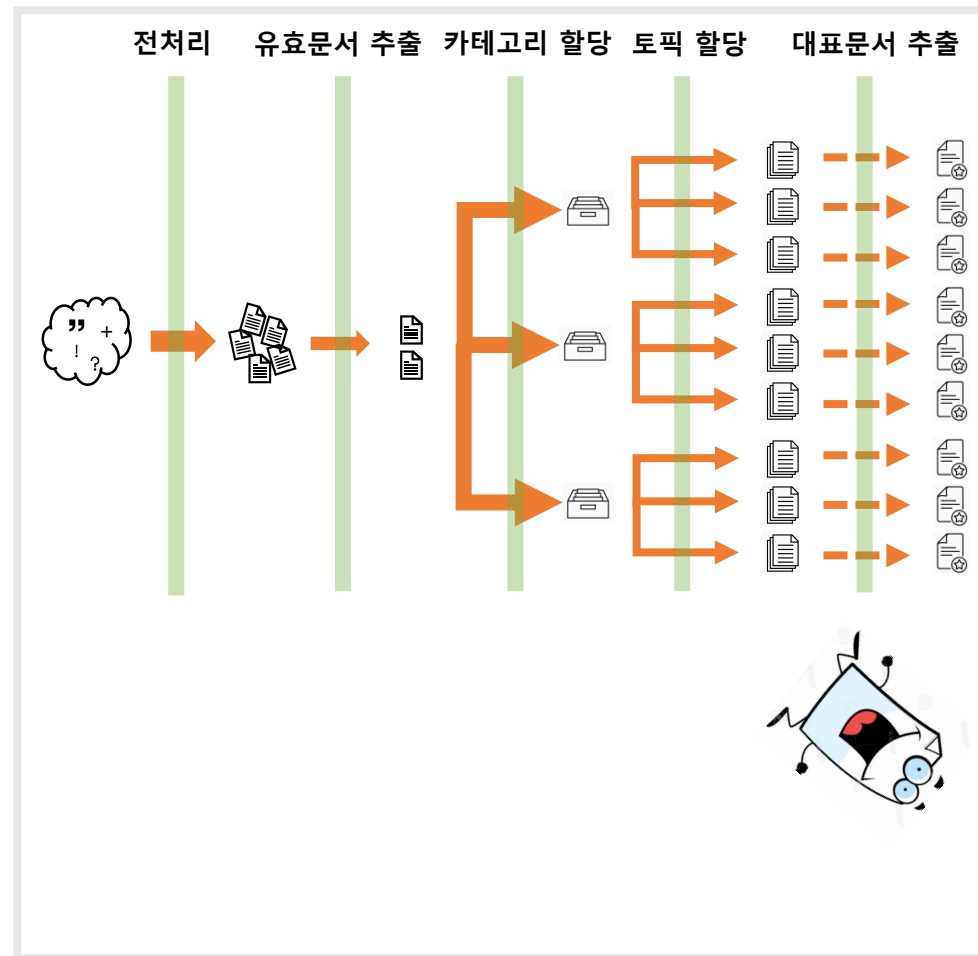
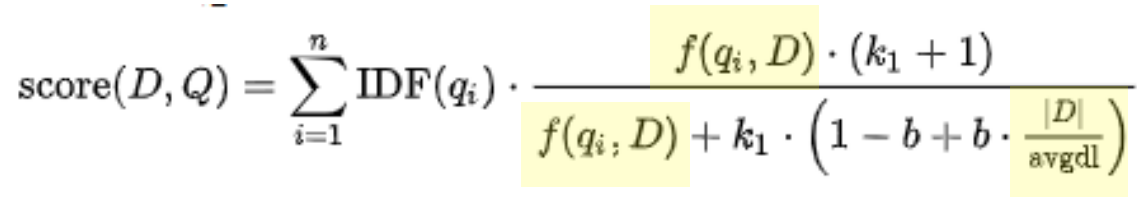


## OPTICS

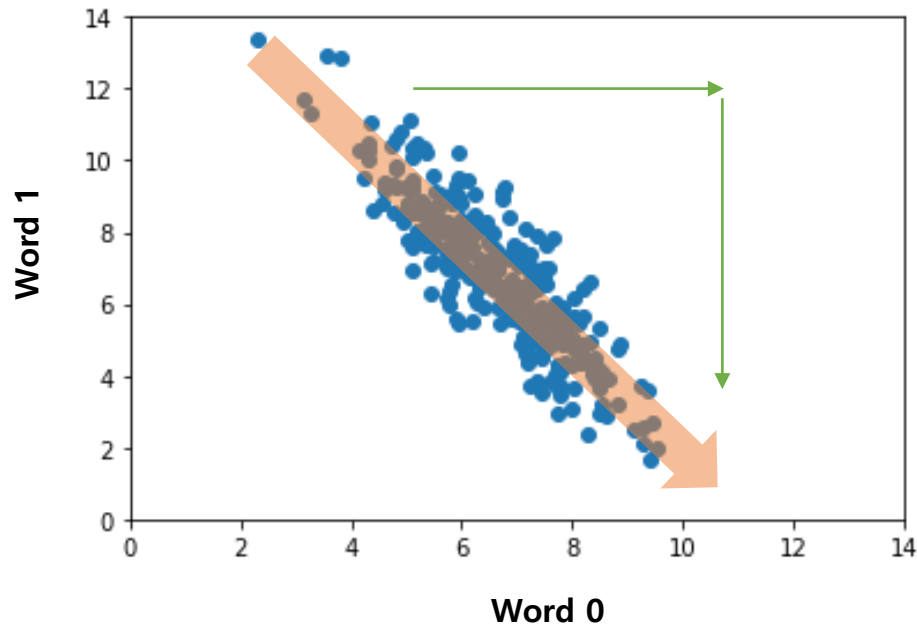
- Epsilon, 군집수를 정할 필요가 없음
- **부분마다 Epsilon을 다르게 설정**
- DBSCAN보다 외상치를 덜 엄격하게 체크함

Reachability distance






SVD



LDA

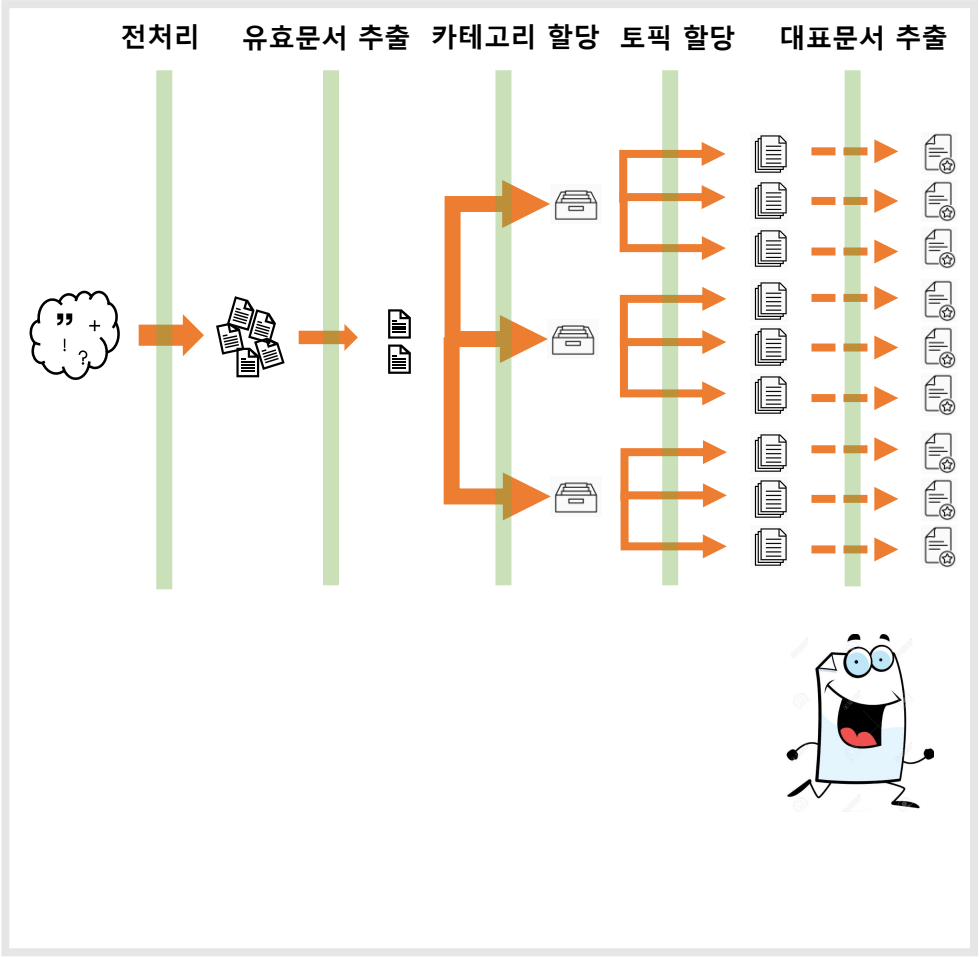


Topic 0



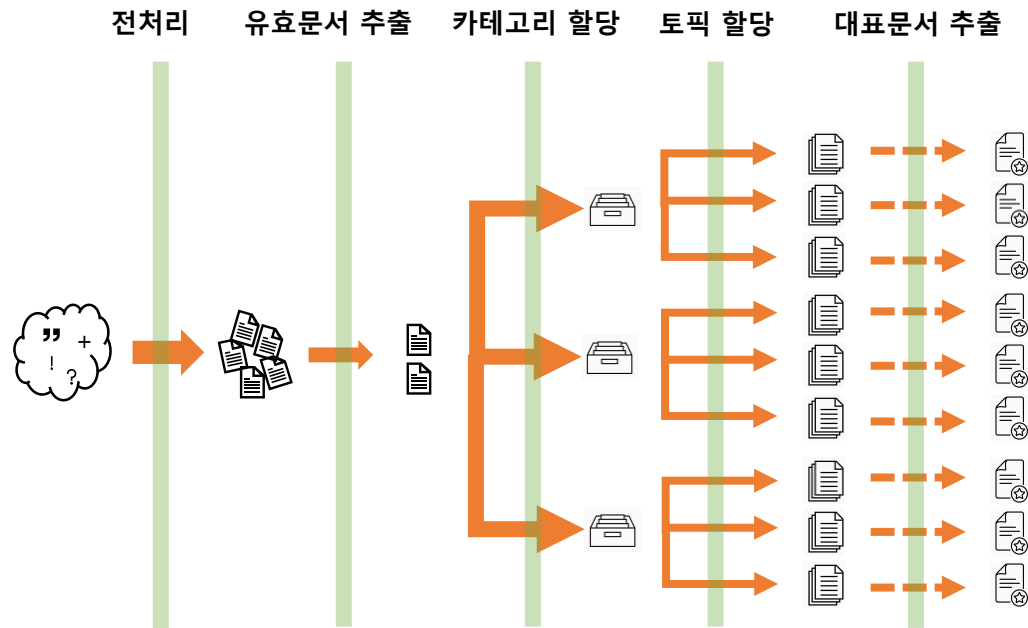
Topic 1

단어	Topic 0	Topic1
매수	0.9	0.1
로봇	0.2	0.8





- 유효 문서 추출
- 유의어 처리
- 카테고리 할당
- 성능 검증



+

- 감정분석
- SWOT





- [1] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander , "OPTICS: Ordering Points To Identify the Clustering Structure", *ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA*, 1999
- [2] Nadia Rahmah and Imas Sukaesih Sitanggang , "Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra", *2016 IOP Conf. Ser.: Earth Environ. Sci. 31 012012*, 2016
- [3] Robertson, Stephen; Zaragoza, Hugo, *The Probabilistic Relevance Framework: BM25 and Beyond (PDF)*. NOW Publishers, Inc. ISBN 978-1-60198-308-4, 2009





감사합니다



- **Kmeans**

- $O(n * k * d * i)$ 
  - $n$  : rows
  - $d$  : cols
  - $k$  : clusters
  - $i$  : iterations

- **LSA**

- $O(\min(m * n^2, n * m^2))$ 
  - $m$  : rows
  - $n$  : cols

- **LDA**

- $O(n * d^2) \sim O(n^3)$ 
  - $n$  : rows
  - $d$  : cols

- **DBSCAN**

- $O(n * \log n) \sim O(n^2)$

- **OPTICS**

- $O(n * \log n) \sim O(n^2)$
- 일반적으로 DBSCAN 보다 1.6배 느림

- **Community detection**

- NP-Class
- $O(n^2)$  이상