

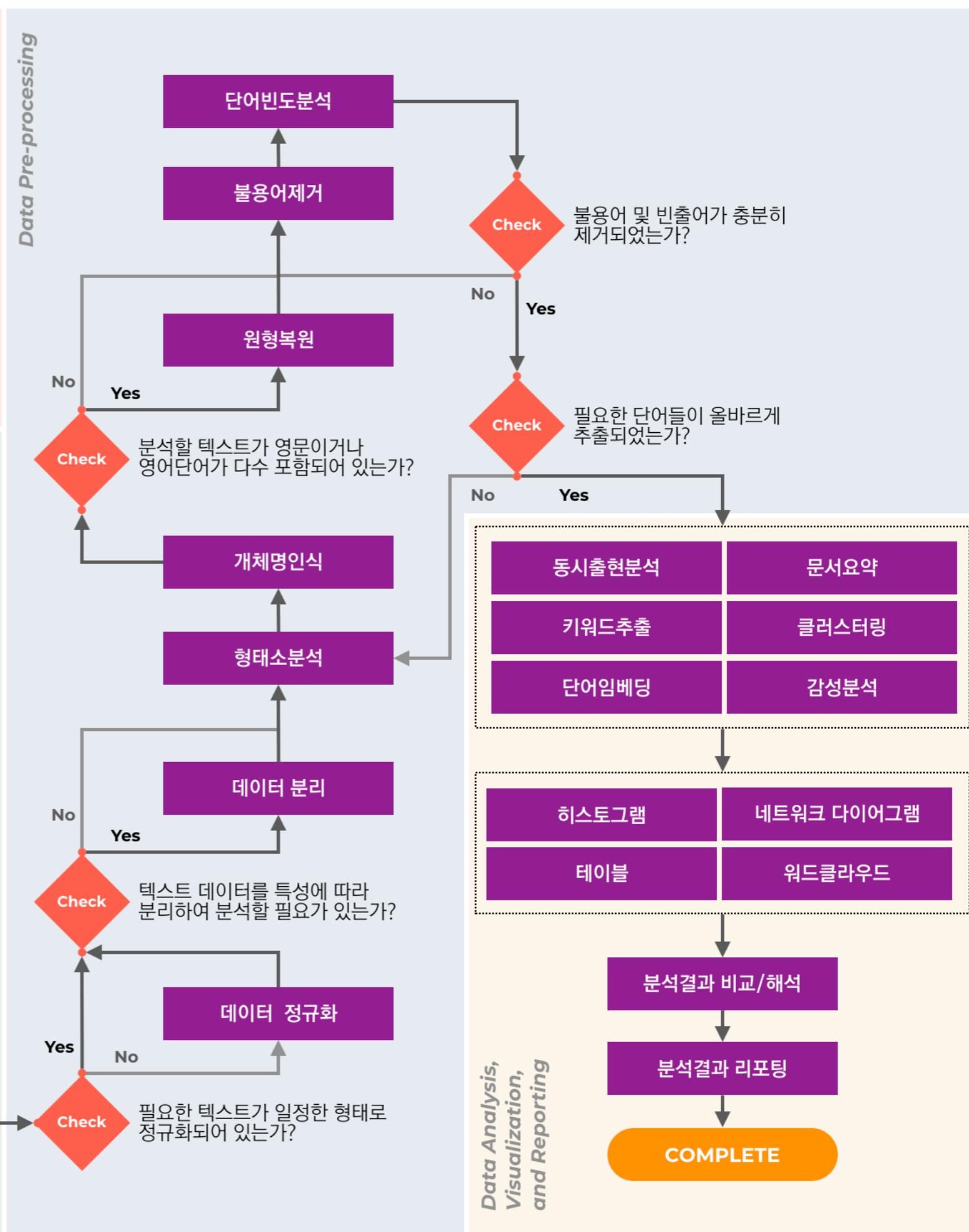
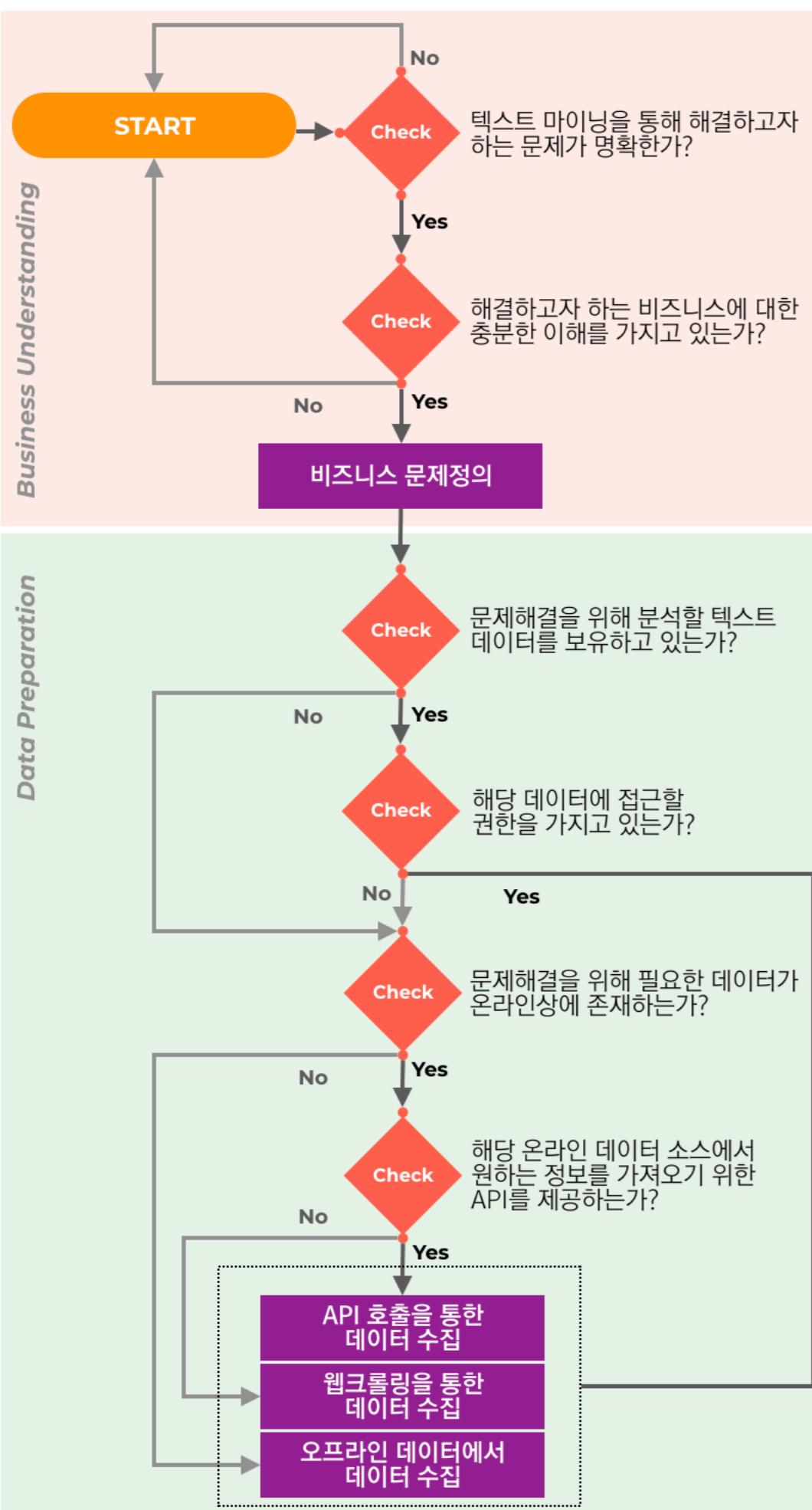
TEXT MINING for BEGINNER

Part 3. 텍스트 마이닝 절차 알아보기

전병진 FINGEREDMAN (fingeredman@gmail.com)

Part 3.

텍스트 마이닝 절차 알아보기



텍스트 데이터 수집

아무데서나 찾은 텍스트는
아무의미없는 정보를 포함한다

텍스트 데이터 수집

텍스트 데이터 수집 유형

- ▶ 수집대상 : 웹페이지, SNS, 댓글, 음성, 비디오 등 텍스트 형태로 변환 가능한 모든 데이터
- ▶ 저장유형 : Plain Text, PDF, Table, XML, JSON
- ▶ 수집방법 : Web Crawling, API 호출, DB Query, Online Survey

유형	수집방법	장/단점
오프라인 데이터	<ul style="list-style-type: none">온/오프라인 설문지음성 녹음, 비디오 촬영	<ul style="list-style-type: none">타겟팅 대상에 대한 데이터 수집 가능사람이 직접 수집해야함수집에 시간적, 공간적 제약이 큼
자체 시스템	<ul style="list-style-type: none">서비스 또는 사내 데이터베이스 활용사내 게시판, 유저 댓글, 업무 보고서, 내부문서	<ul style="list-style-type: none">기 수집된 데이터를 빠르게 활용 가능소속기관/업체/서비스 내부 관련자가 아니면 접근이 어려움정보 유출에 대한 위험이 큼
웹크롤링	<ul style="list-style-type: none">프로그래밍 언어를 활용해 웹페이지에 존재하는 대량의 정보를 반복 수집	<ul style="list-style-type: none">대량의 정보를 빠르게 수집할 수 있음데이터 수집과 함께 정규화된 데이터셋 구성 가능프로그래밍 언어 활용이 필요함개인정보 문제에 취약함
API 호출	<ul style="list-style-type: none">프로그래밍 언어를 활용해 서비스에서 정식으로 제공하는 데이터를 수집네이버 API, 카카오 API, Reddit API, News API, SNS (Twitter, Facebook, Instagram)	<ul style="list-style-type: none">바로 활용할 수 있는 양질의 데이터를 얻을 수 있음수집할 수 있는 소스가 제한적임프로그래밍 언어 활용이 필요함

텍스트 데이터 수집

텍스트 데이터를 제공하는 API

서비스 유형	서비스	제공 유형	비고
SNS	Facebook	<ul style="list-style-type: none">Graph APIPost, Comments	<ul style="list-style-type: none">https://developers.facebook.com 참고
	Instagram	<ul style="list-style-type: none">Media, Comments, Tags	<ul style="list-style-type: none">https://www.instagram.com/developer 참고
	Twitter	<ul style="list-style-type: none">Search APIStreaming API	<ul style="list-style-type: none">https://developer.twitter.com 참고
포털	네이버	<ul style="list-style-type: none">블로그 API뉴스 API백과사전 API웹문서 API	<ul style="list-style-type: none">https://developers.naver.com 참고본문만 제공
	다음(카카오)	<ul style="list-style-type: none">웹문서 검색 API블로그 검색 API카페 검색 API	<ul style="list-style-type: none">https://developers.kakao.com 참고본문만 제공
커뮤니티	Reddit	<ul style="list-style-type: none">Thread, Comment	<ul style="list-style-type: none">https://www.reddit.com/dev/api 참고
뉴스	News API	<ul style="list-style-type: none">Article Summery	<ul style="list-style-type: none">https://newsapi.org 참고
	American Broadcasting Company	<ul style="list-style-type: none">Resources API	<ul style="list-style-type: none">Resources API retrieves content produced by ABC businesses, including national and local news, entertainment videos, and more
	BBC	<ul style="list-style-type: none">Platform API	<ul style="list-style-type: none">Platform API power all the BBC's product areas
	New York Times	<ul style="list-style-type: none">Article Search APICommunity APIMovie Reviews APITimesTags APITop Stories API	<ul style="list-style-type: none">Headlines, abstracts and links to associated multimediaComments by NYTimes.com usersLinks to reviews and NYT Critics' Picks, and search movie reviewsTerms that match search queries, and filters by Times dictionariesLists of home page articles and associated images

텍스트 데이터 수집

타겟팅 유저 텍스트 수집을 위한 소스

소스	유형	플랫폼	주요토픽	회원수	사용연령대	성별
디시인사이드	커뮤니티	자체플랫폼	공통		10~30	공통
루리웹	커뮤니티	자체플랫폼	공통		20~30	공통
뽐뿌	커뮤니티	자체플랫폼	공통		10~30	공통
일베저장소	커뮤니티	자체플랫폼	공통		10~40	공통
스레딕	커뮤니티	자체플랫폼	공통		20~30	여성
도탁스	카페	다음	공통	511,049	-	공통
이토랜드	토렌트	자체플랫폼	공통		-	공통
네이트판	커뮤니티	자체플랫폼	고민, 이슈		10~30	공통
오늘의유머	커뮤니티	자체플랫폼	유머		10~30	공통
웃긴대학	커뮤니티	자체플랫폼	유머		10~30	공통
엽기혹은진실	카페	다음	유머	247,754	-	공통
유머나라	카페	다음	유머	114,626	-	공통
와이고수	커뮤니티	자체플랫폼	유머, 스포츠, 게임		10~40	남성
쭉빵카페	카페	다음	연예, 뷰티	1,731,956	20~30	여성
뉴빵카페	카페	다음	연예, 뷰티	1,101,596	20~30	여성
여성시대	카페	다음	연예, 뷰티	729,142	20~30	여성
파우더룸	카페	네이버	뷰티	1,856,696	20~30	여성
인스티즈	커뮤니티	커뮤니티	연예, 오락		10~30	여성
theqoo	커뮤니티	자체플랫폼	연예		10~20	여성
해연갤	커뮤니티	자체플랫폼	해외 연예		20~30	여성
가생이	커뮤니티	자체플랫폼	연예, 한류		20~40	-
베스티즈	커뮤니티	자체플랫폼	연예		20~30	여성
디젤매니아	카페	네이버	패션	882,132	20~30	남성
외방커뮤니티	커뮤니티	자체플랫폼	미용, 패션		20~30	여성

텍스트 데이터 수집

타겟팅 유저 텍스트 수집을 위한 소스

소스	유형	플랫폼	주요토픽	회원수	사용연령대	성별
레몬테라스	카페	네이버	육아, 인테리어	3,020,341	30~40	여성
맘스홀릭 베이비	카페	네이버	육아	2,684,457	20~30	여성
개드립	커뮤니티	자체플랫폼	유머, 게임		10~20	공통
인벤	커뮤니티	자체플랫폼	게임		10~20	남성
에펨코리아	커뮤니티	자체플랫폼	축구		10~40	남성
아이러브사커	카페	다음	축구	167,706	10~40	남성
MLB파크	커뮤니티	커뮤니티	야구		20~40	남성
이종격투기	카페	다음	격투기	1,023,757	10~30	남성
클리앙	커뮤니티	자체플랫폼	테크, 통신, 앱		20~40	남성
쿨엔조이	커뮤니티	자체플랫폼	테크, 하드웨어		20~40	남성
Seeko	커뮤니티	자체플랫폼	전자기기		30~40	남성
아사모 - 애플	카페	네이버	애플 아이폰	1,635,061	-	공통
중고나라	카페	네이버	중고거래	16,477,444	10~50	공통
중고카페 그린유즈	카페	네이버	중고거래	2,543,783	-	공통
보배드림	커뮤니티, 쇼핑몰	자체플랫폼	중고거래		30~50	공통
취업뽀개기	카페	다음	취업, 학생	1,399,394	20~30	공통
독취사 - 취업	카페	네이버	취업	2,393,699	20~30	공통
오르비	커뮤니티	자체플랫폼	수험생, 입시		10~20	공통
수만휘	카페	네이버	수험생	2,515,951	10~20	공통
82쿡	커뮤니티, 쇼핑몰		요리		20~50	여성
SLR클럽	커뮤니티	자체플랫폼	사진		30~50	공통
유랑 - 유럽여행	카페	네이버	여행	1,880,443	20~30	공통
네일동 - 일본여행	카페	네이버	여행	1,205,047	20~30	공통

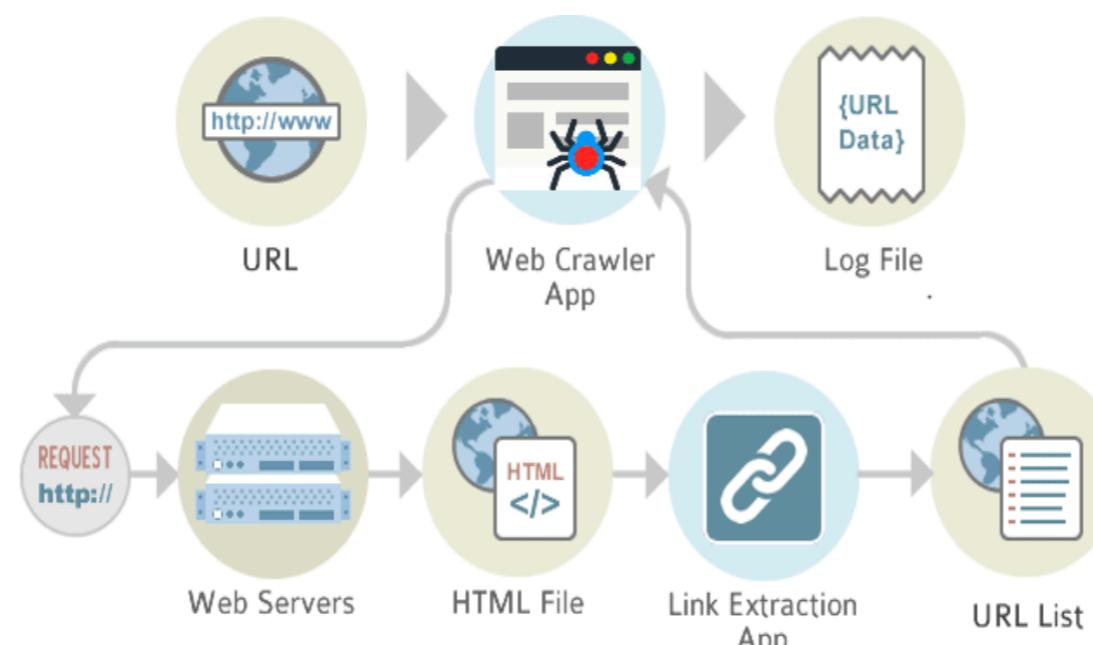
텍스트 데이터 수집

Crawling 이란?

- ▶ 정제되지 않은 웹페이지에서 필요한 데이터를 추출하는 행위
- ▶ 활용할 수 있는 데이터가 정리되어 올라가 있는 데이터(API, 파일형태)를 제외하고, 웹페이지에 게시된 자료를 가져오는 기술
- ▶ API를 통한 데이터 제공이 활성화 되고 있는 추세이나, 국내에서는 소극적인 API 제공으로 웹크롤링을 통한 수집이 반드시 필요함

Data Crawling & Scrapping 차이점

- ▶ Data Crawling : 웹페이지의 정보를 추출하는 방법(=spider, bot), 단순히 하이퍼링크를 돌아다니며 웹페이지를 다운로드
- ▶ Data Scrapping : 웹페이지에서 필요한 정보만을 추출하는 방법, 다운로드한 웹페이지에서 필요한 부분만을 추출하고 저장
- ▶ 단순히 데이터 수집을 넘어, 데이터를 수집하고 활용하기 위한 사전작업



*Source : Rodrigo, I Will Code A Small Web Crawler, <https://www.fiverr.com/rodrigo/code-a-small-web-crawler/>.

크롤링의 도덕적 문제

데이터 무단수집과 저작권 침해

- ▶ 웹크롤링은 원래 검색엔진 등의 인터넷 사이트에서 데이터를 최신 상태로 유지하기 위해 사용
- ▶ 웹크롤링을 활용하여 타사 컨텐츠를 무단 활용하는 것은 불법행위에 해당함
- ▶ 웹크롤링을 통한 과도한 요청은 대상 서비스 서버 운영과 서비스 관리에 안 좋은 영향을 끼침
- ▶ 경쟁사간의 상도덕 문제 또는 개인 양심상의 문제

채용정보 무단복제 '사람인HR', 잡코리아에 120억 지급

양보다 질 중요한 취업포탈 업계…접근 쉬운 채용공고 속성 악용한 편취사례

이준영 기자 | 승인 2018.02.09 12:31 | 댓글 0

댓글부대 의혹 야놀자, 무단 DB 크롤링 의혹 여기어때

숙박 O2O 시장 논란 언제까지

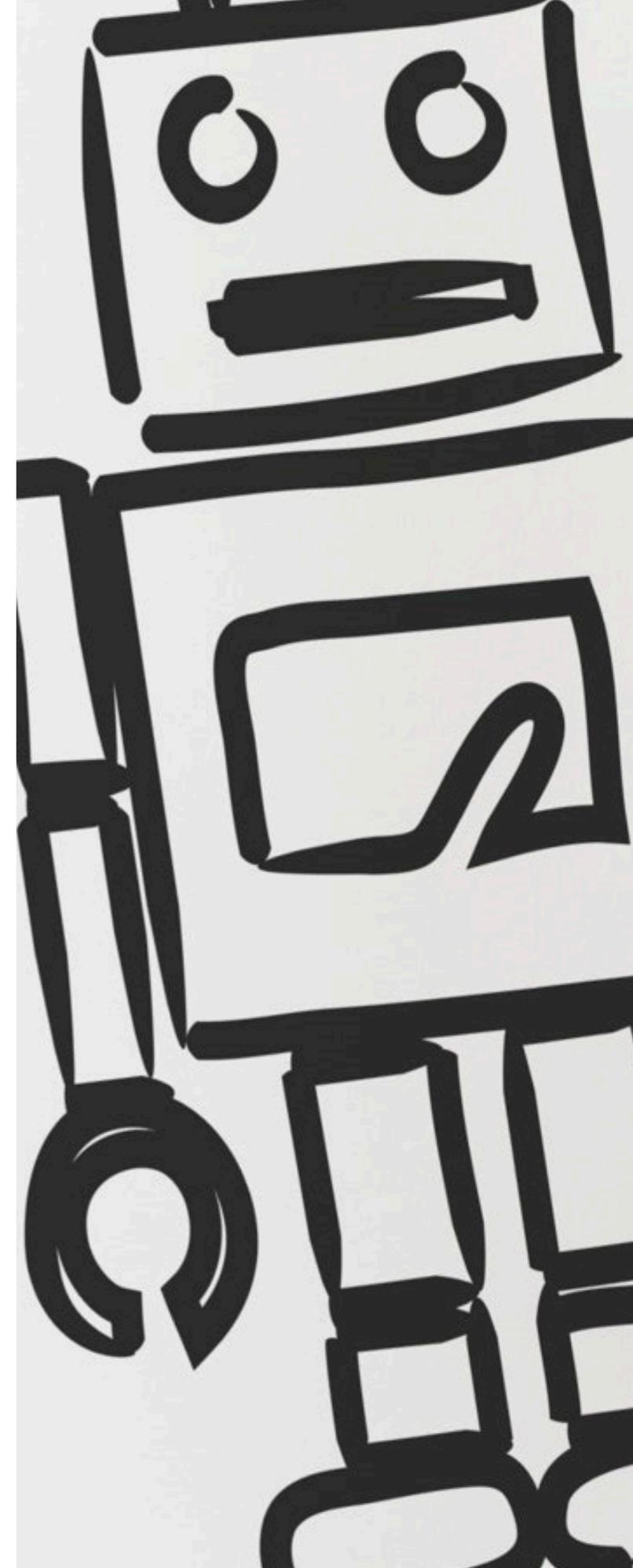
최진홍 기자 | rgdsz@econovill.com | 승인 2017.11.03 16:23:50



사진=각 사

제 및 게재 행위를 하지 않고 공정한 경쟁질서

모바일 시대가 도래하며 O2O 스타트업의 존재감이 날카로워지고 있지만 잡음 또한 높아지고 있다. 이들은 온라인 경쟁력을 키우면서도 오프라인 거점도 확보, 이를 통한 다양한 파생 서비스에 나선다는 목표도 세워놓고 있다. 그러나 숙박 O2O 업체들이 용인할 수 있는 수준을 넘어설 정도로 구설에 오르고 있는 것은 여간 심각한 문제가 아니다. 최근 나름 적절한 수위를 찾아간다는 평가가 나오고 있지만 배달의민족, 요기요, 배달통 등이 포진한 배달앱 업계도 마찬가지고 다방과 직방 등 부동산 O2O 시장도 사정이 비슷하다. 그 중에서 숙박 O2O 시장을 둘러싼 논란은 상상 이상이다.



*Source : 이준영(시장경제), 채용정보 무단복제 '사람인HR', 잡코리아에 120억 지급, 2018.2.9., <http://www.mconomynews.com/news/articleView.html?idxno=11088/>.

**Source : 최진홍(이코노미리뷰), 댓글부대 의혹 야놀자, 무단 DB 크롤링 의혹 여기어때, 2017.11.3., <http://www.econovill.com/news/articleView.html?idxno=325820/>.

크롤링의 도덕적 문제

Robots.txt

- ▶ 웹사이트에 배치된 텍스트 파일로, 크롤링 접근권한에 대해 명시해 놓은 문서
- ▶ 웹크롤링은 Robots.txt 파일에서 허용하는 항목에 대해서만 수집 가능하며 그 외의 수집에 대한 책임은 모두 본인에게 있음
- ▶ 수집이 허용되 있더라도 대상 서비스 운영에 피해를 주지 않는 선에서 필요한 만큼만 수집
- ▶ Robots.txt 파일이 없는 경우 서비스 관리자에 직접 허락을 구한 후 수집

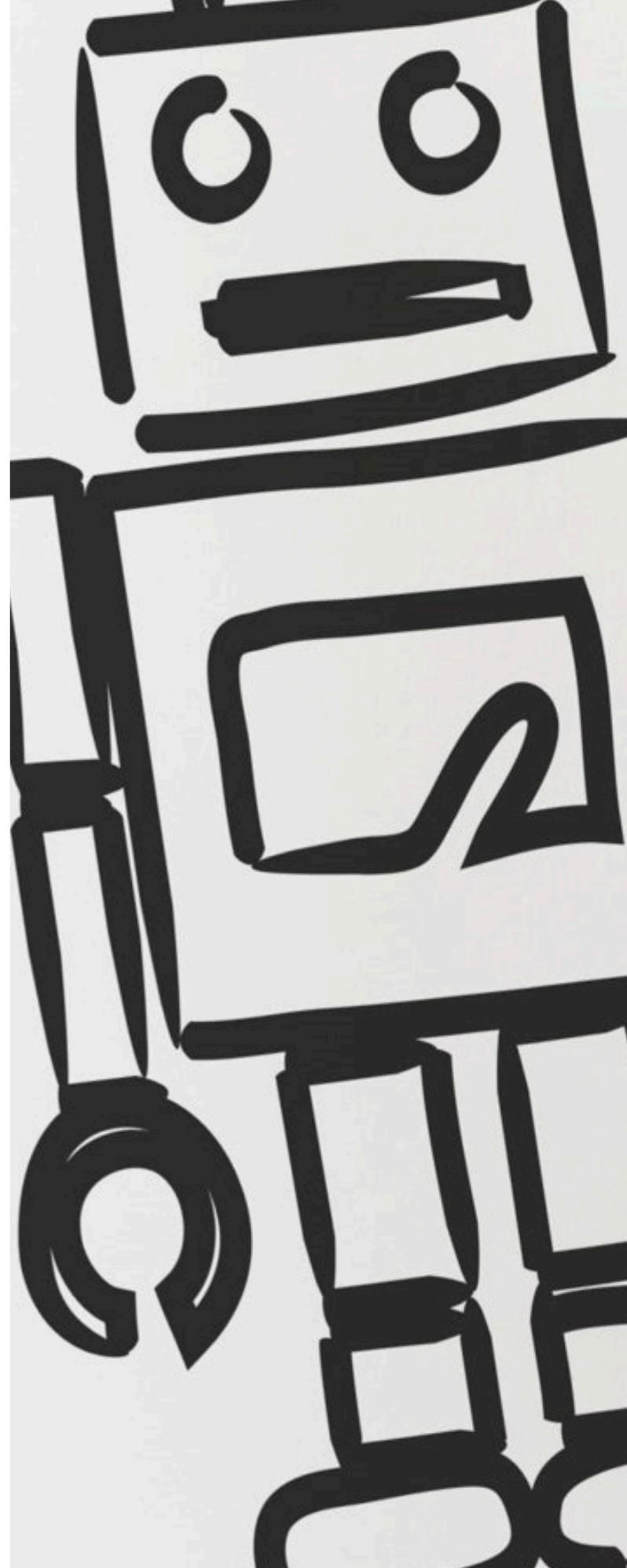


```
User-agent: Bingbot
Allow: /ajax/pagelet/generic.php/PagePostsSectionPagelet
Allow: /safetycheck/
```

```
User-agent: Googlebot
Allow: /ajax/pagelet/generic.php/PagePostsSectionPagelet
Allow: /safetycheck/
```



```
User-agent: *
Allow:/service/board/
Disallow:/service/group/
Disallow:/service/board/sold/
Disallow:/service/mypage/
Disallow:/service/message/
Disallow:/service/popup/
Disallow:/service/search/
Disallow:/service/cs/
```



Python 웹크롤링

Python을 활용하는 이유

- ▶ 가장 배우기 쉽고 읽기 쉬는 프로그래밍 언어
- ▶ 풍부한 외부 라이브러리
- ▶ 데이터 수집 뿐만아니라 데이터 전처리, 분석과의 친화성

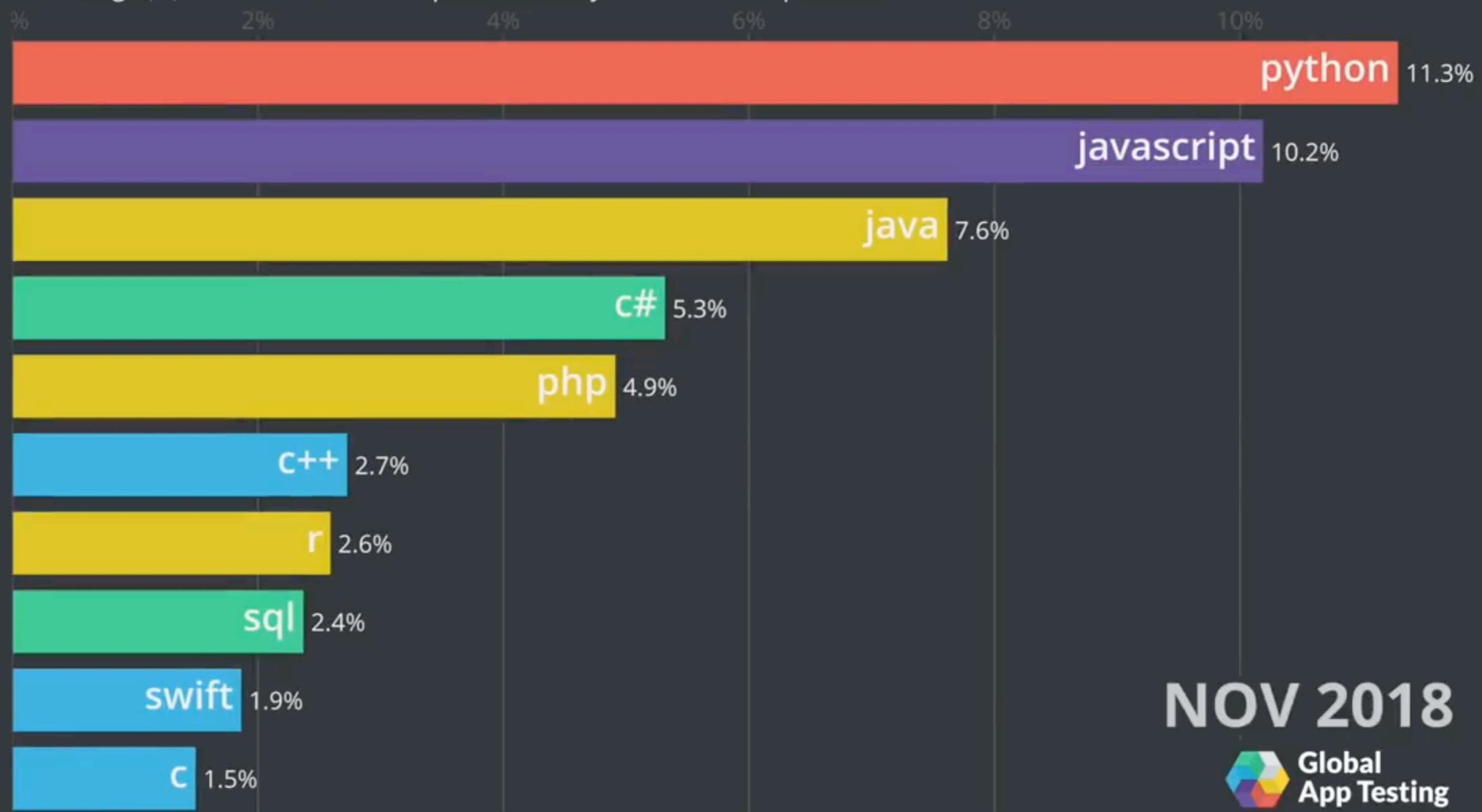
크롤링하는 방법 선택

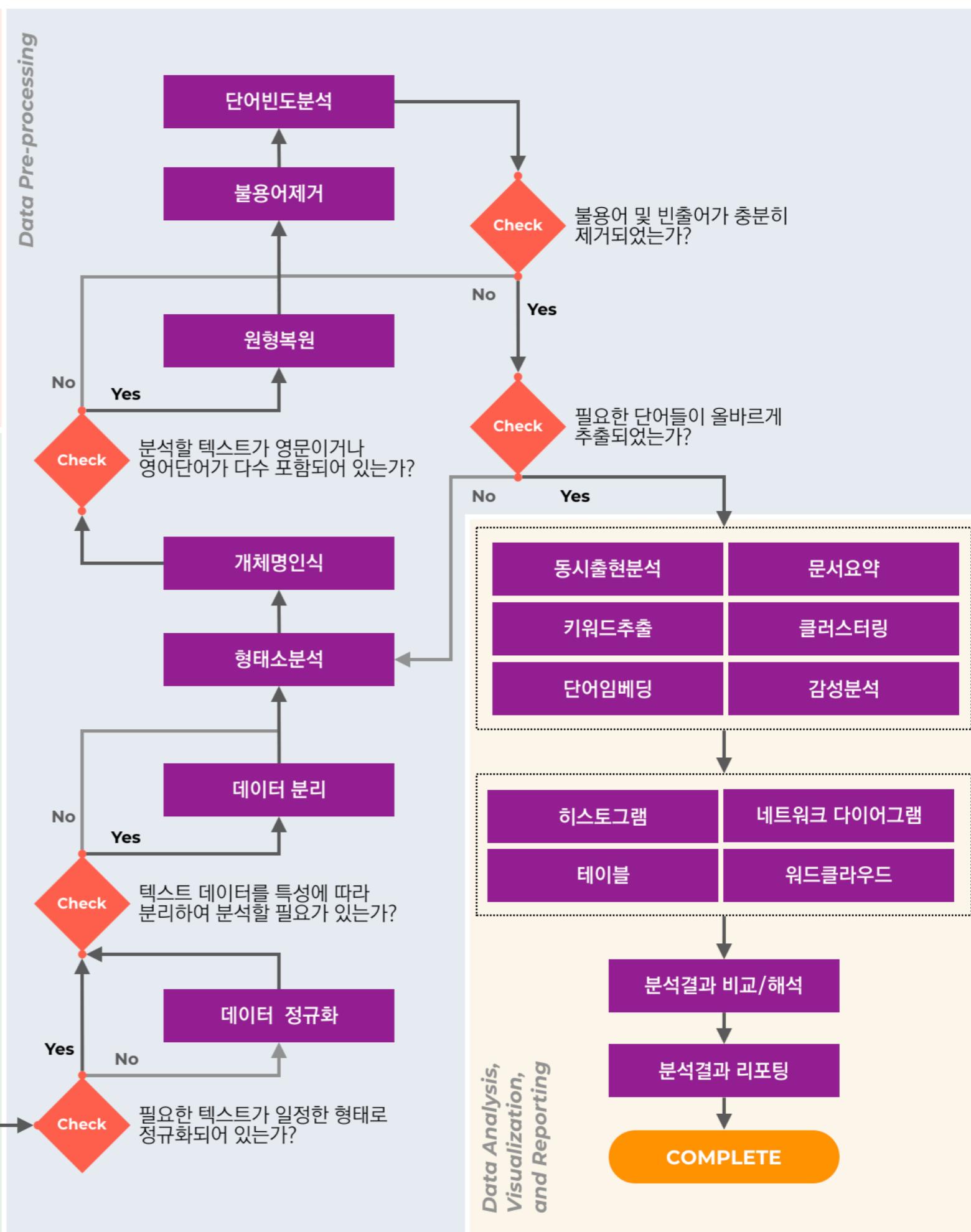
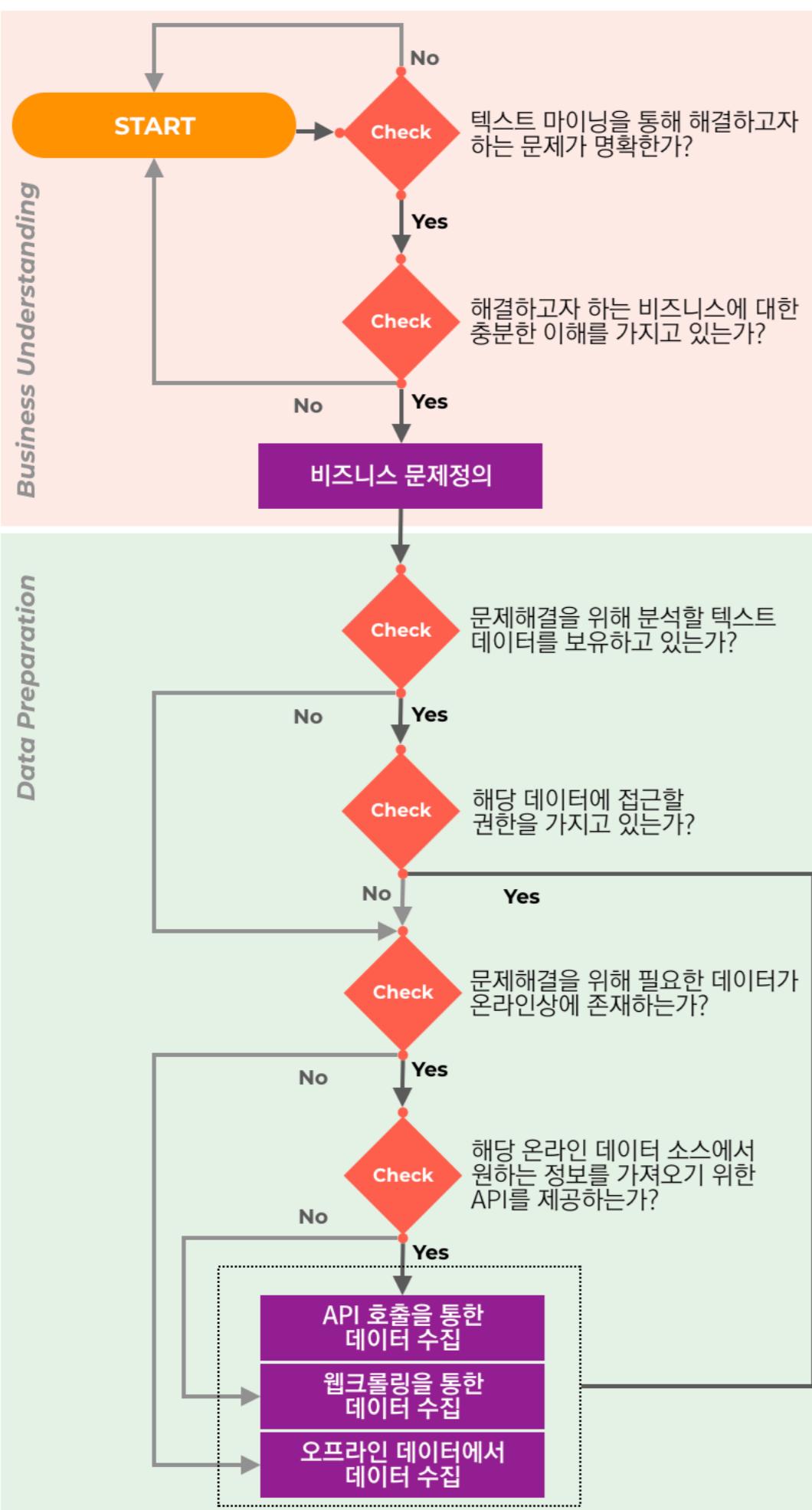
구분	라이브러리	역할 및 장점/단점
브라우저 운영	Requests (urllib)	장점 : Python에서 동작하는 가장의 작고 빠른 브라우저 거의 모든 플랫폼에서 구동 가능 단점 : 웹서버로부터 초기 HTML만 받아 추가적인 CSS/JavaScript 처리결과가 반영되지 않음
	Selenium	장점 : 웹브라우저를 통합된 방식으로 원격 컨트롤하는 라이브러리 (Chrome, Firefox, IE, PhantomJS 등) 단점 : 직접 브라우저를 사용하여 추가로 CSS/JavaScript 처리를 위한 리소스가 많이 필요함
파서 (Parser)	BeautifulSoup4	HTML로부터 원하는 위치/형식의 문자열을 획득



Most popular programming languages on Stack Overflow.

Percentage (%) of all Stack Overflow questions every month since September 2008.





텍스트 데이터 전처리

텍스트 전처리에 허투루하면
허투문 분석결과가 나온다

텍스트 마이닝 용어

① 문서 (Document)

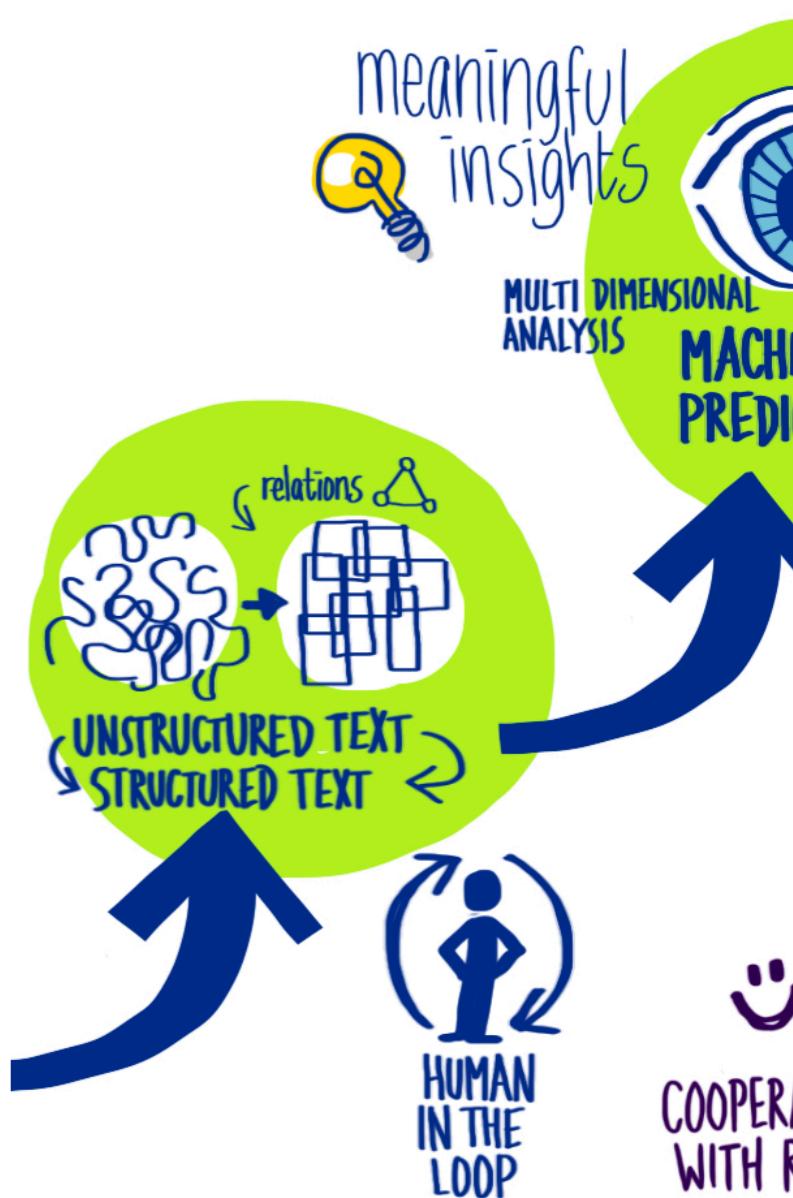
- ▶ 한 덩어리의 텍스트로서 짧은 문장에서부터 긴 문서 까지를 모두 포함하는 의미
- ▶ 문서의 집합을 말뭉치 (corpus)라고 함
- ▶ 문서의 레벨에 따라 말뭉치의 레벨이 바뀔 수 있음(문장, 문단, 페이지, 댓글 등)

② 어휘사전 (Lexicon)

- ▶ 어휘 (lexical)의 집합 또는 어휘에 대한 정의 혹은 설명을 가진 사전
- ▶ 특징 별 어휘사전이 나뉘어서 존재하기도 함(인물사전, 영어사전, 건물사전 등)

③ 불용어 (Stop-word)

- ▶ 텍스트 분석에 있어서, 또는 분석결과에 출현하더라도 아무런 의미가 없는 단어의 집합
- ▶ 정보전달 보다는 주로 기능적인 역할을 하는 단어에 해당함
 - 일반적으로 한국어는 외자 단어, 영어는 알파벳 두 자 단어를 불용어 취급
 - 한국어 예 : 그거, 여기, 이제, 은, 는, 이, ...
 - 영어 예 : a, an, the, of, the, ...
- ▶ 빈출어 (Common-word)
 - 너무 많이 출현하여 분석 결과에서 의미 또는 중요도가 떨어지는 단어의 집합
 - 예 : 기사, 기자, 제목, 사진, 네이버, 검색, 보다, 연기, 평점, 공감, 비공감



텍스트 마이닝 용어

④ 형태소 (Morpheme)

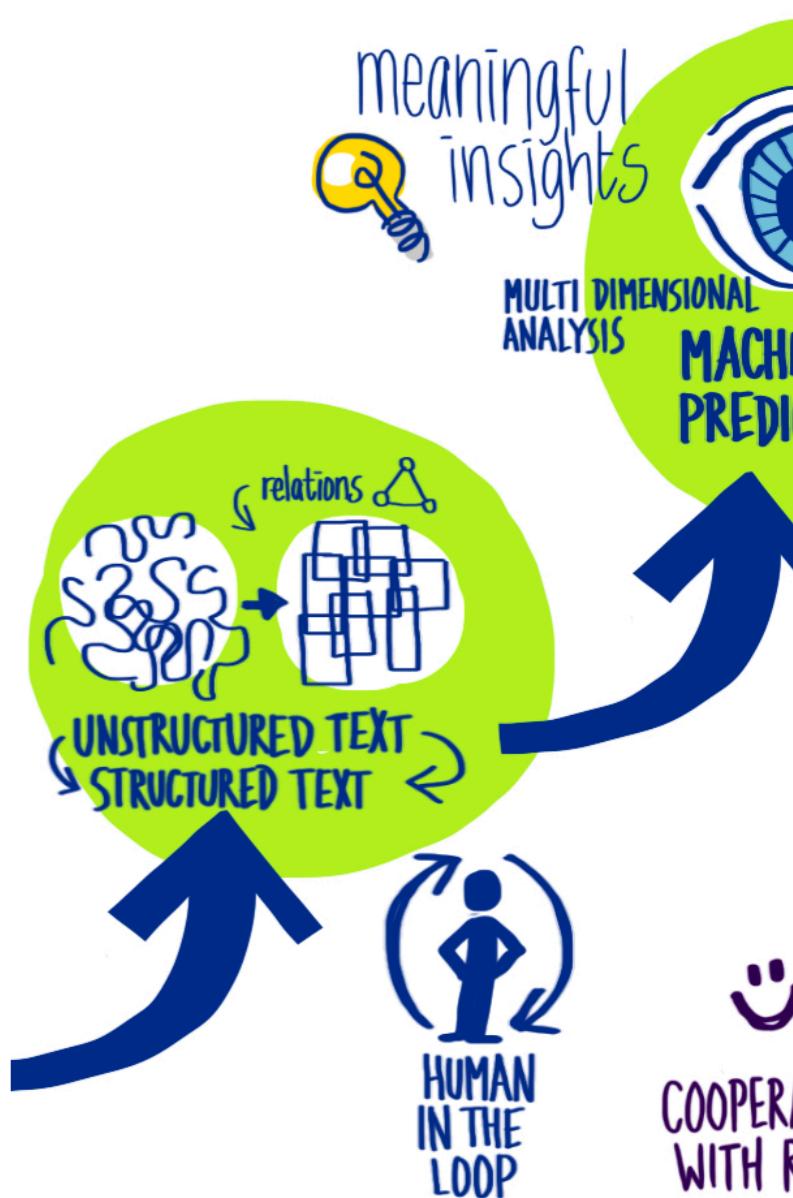
- ▶ 뜻을 가진 가장 작을 말의 단위
- ▶ 동사, 명사, 조사, 문장부호 등 보통 품사 (Part of Speech, POS) 단위를 의미함

⑤ 단어 주머니 (Bag of Words, BoW)

- ▶ 문서에 함께 사용된 단어의 집합
- ▶ 중복된 단어는 하나로 취급하며, 순서에 의미를 고려하지 않음
 - 예 : “아버지가 방에 들어가신다.” → [“아버지”, “방”, “들어가다”]

⑥ 토큰화 (Tokenization)

- ▶ 토큰 (token)
 - 유용한 의미적 단위로 함께 모여지는 일련의 문자열
 - 구분 기호 사이의 글자 시퀀스
- ▶ 문헌 단위의 문자열이 주어졌을 때 토큰들로 문자열을 분리하는 작업
- ▶ 구두점 등 불필요한 글자들을 제외하기도 함
- ▶ 영어는 언어학적 특성상 단어에 조사가 붙지 않아 한글보다 토큰화가 쉬움



문서 정규화

문서를 하나의 통일된 형식으로 정규화하는 과정

- ▶ 데이터를 수집하면서 정규화를 같이 진행하면 매우 효율적으로 할 수 있음
- ▶ 정규화 유형 : Table, XML, JSON

[Table]

employee	
name	age
James Kirk	40
Jean-Luc Picard	45
Wesley Crusher	27

1 name → age
2 James Kirk → 40
3 Jean-Luc Picard → 45
4 Wesley Crusher → 27

[XML]

```
<empinfo>
  <employees>
    <employee>
      <name>James Kirk</name>
      <age>40</age>
    </employee>
    <employee>
      <name>Jean-Luc Picard</name>
      <age>45</age>
    </employee>
    <employee>
      <name>Wesley Crusher</name>
      <age>27</age>
    </employee>
  </employees>
</empinfo>
```

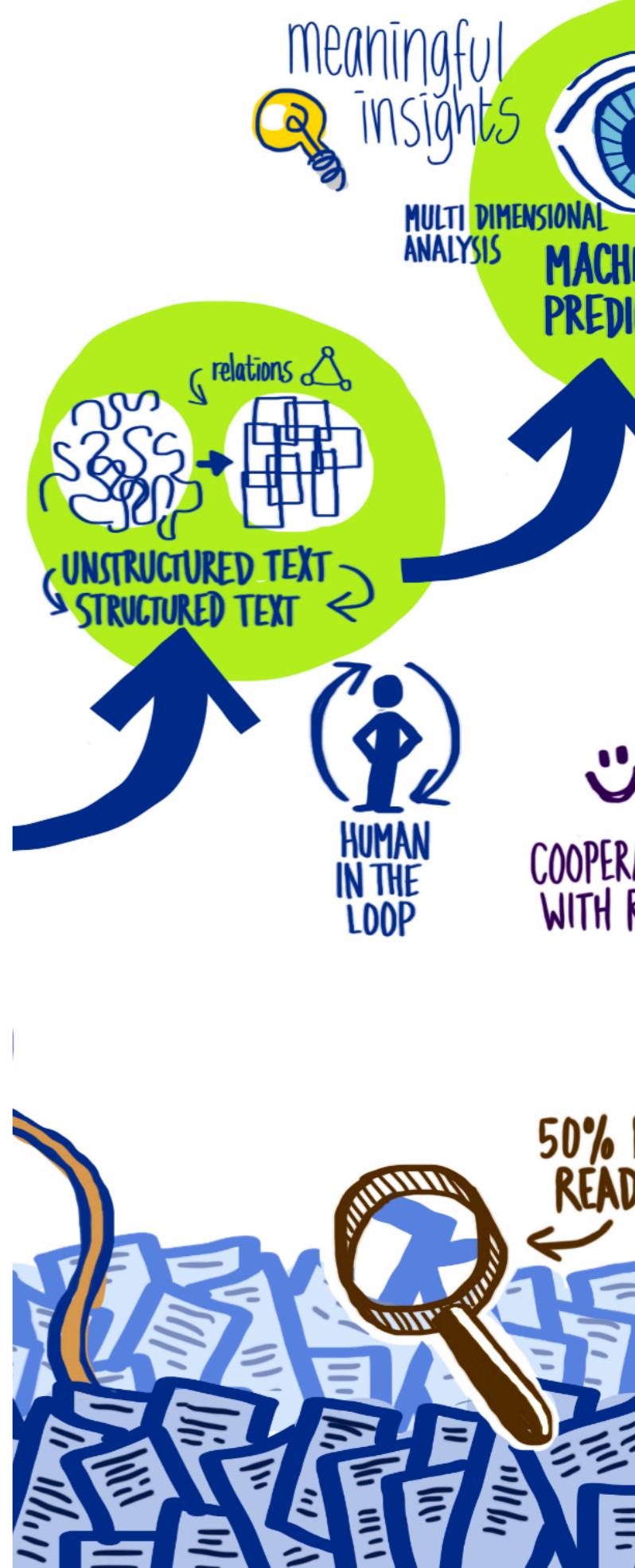
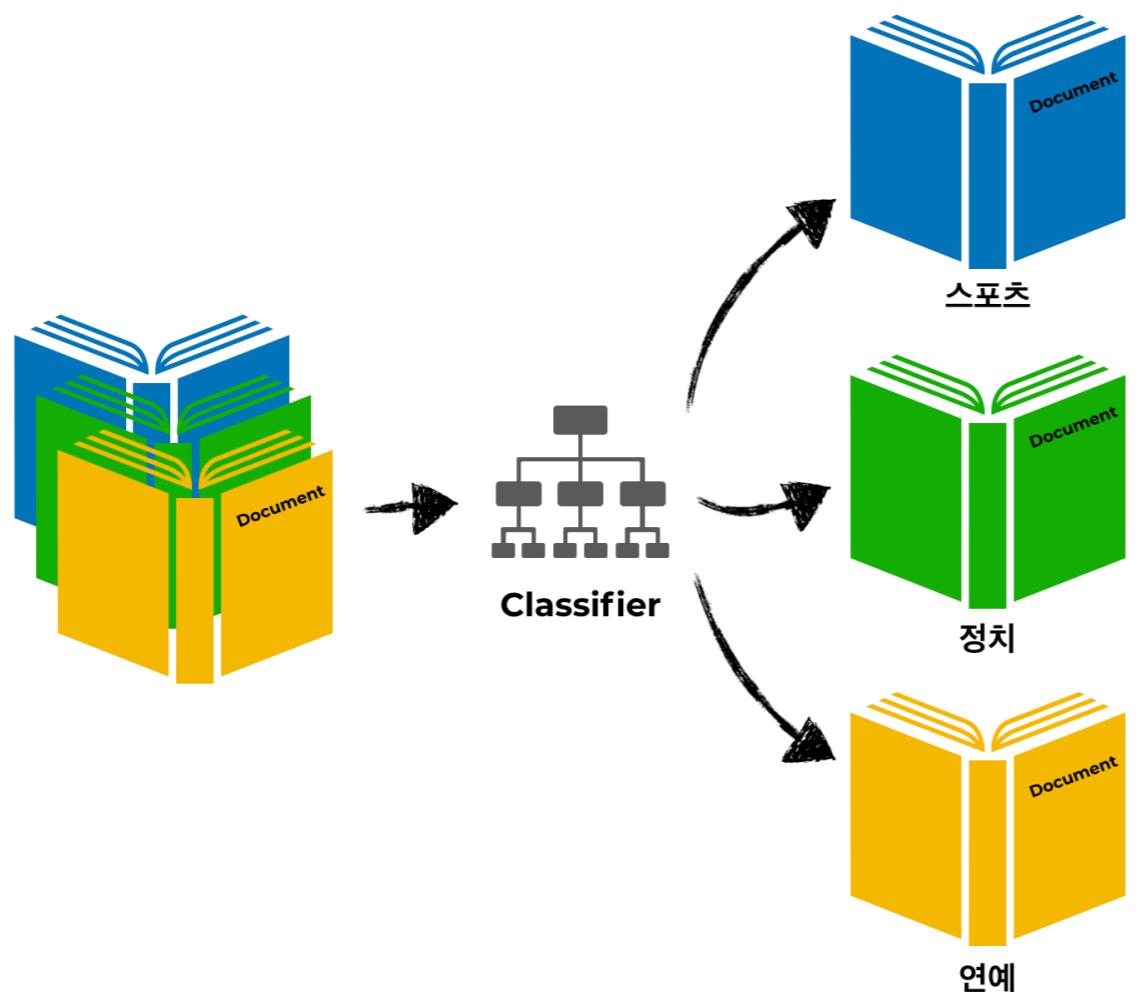
[JSON]

```
{ "empinfo" :
  {
    "employees": [
      {
        "name" : "James Kirk",
        "age" : 40,
      },
      {
        "name" : "Jean-Luc Picard",
        "age" : 45,
      },
      {
        "name" : "Wesley Crusher",
        "age" : 27,
      }
    ]
  }
}
```

문서 분리 (Seperation)

문서를 특정 기준에 의해 분리하거나 통합하는 과정

- ▶ 문서의 작성자, 날짜, 길이, 구분자, 감성 스코어, 랜덤 등을 기준으로 할 수 있음
- ▶ 모델 생성 : 모델 생성을 위한 훈련과 테스트용 데이터를 구분
- ▶ 분석단위 : 데이터에 포함된 특정 값을 기준으로 분리
- ▶ 문장단위 : 텍스트는 문장단위로 구분(언어를 이해하는 최소단위)



형태소 분석 (Part of Speech Tagging)

교착어, 굴절어, 그리고. 고립어

- ▶ 교착어 (agglutinative language) : 어근에 접사가 결합되어 각 단어의 기능을 나타내는 언어 (한국어, 일본어, 몽골어, ...)
- ▶ 굴절어 (inflectional language) : 단어 자체의 형태변화로 그 단어의 문법성을 나타내는 언어 (라틴어, 독일어, 러시아어, ...)
- ▶ 고립어 (isolating language) : 단어의 형태변화 없이 문법적 관계는 어순에 의해 정해지는 언어 (영어, 중국어, ...)

형태소 분석이란?

- ▶ 문장을 형태소 단위로 구분하고 품사를 구별하여 태깅하고 용언의 다양한 활용으로 인한 형태소 탈락현상을 복원하는 과정
- ▶ 분석기마다 형태소 구분 방식이 다르기 때문에 데이터에 맞는 분석기를 선택해야함
- ▶ 모든 언어의 자연어 처리 과정 중 가장 중요하고 기초적인 역할 수행
- ▶ 형태소 분석의 활용
 - 언어학적 측면 : 특정 언어현상의 생성과정을 설명하는 데 용이하게 쓰일 수 있음
 - 전산학적 측면 : 정보검색이나 자연어 처리 자동 처리시스템의 구문 분석의 전 단계 등의 용도로 쓰일 수 있음

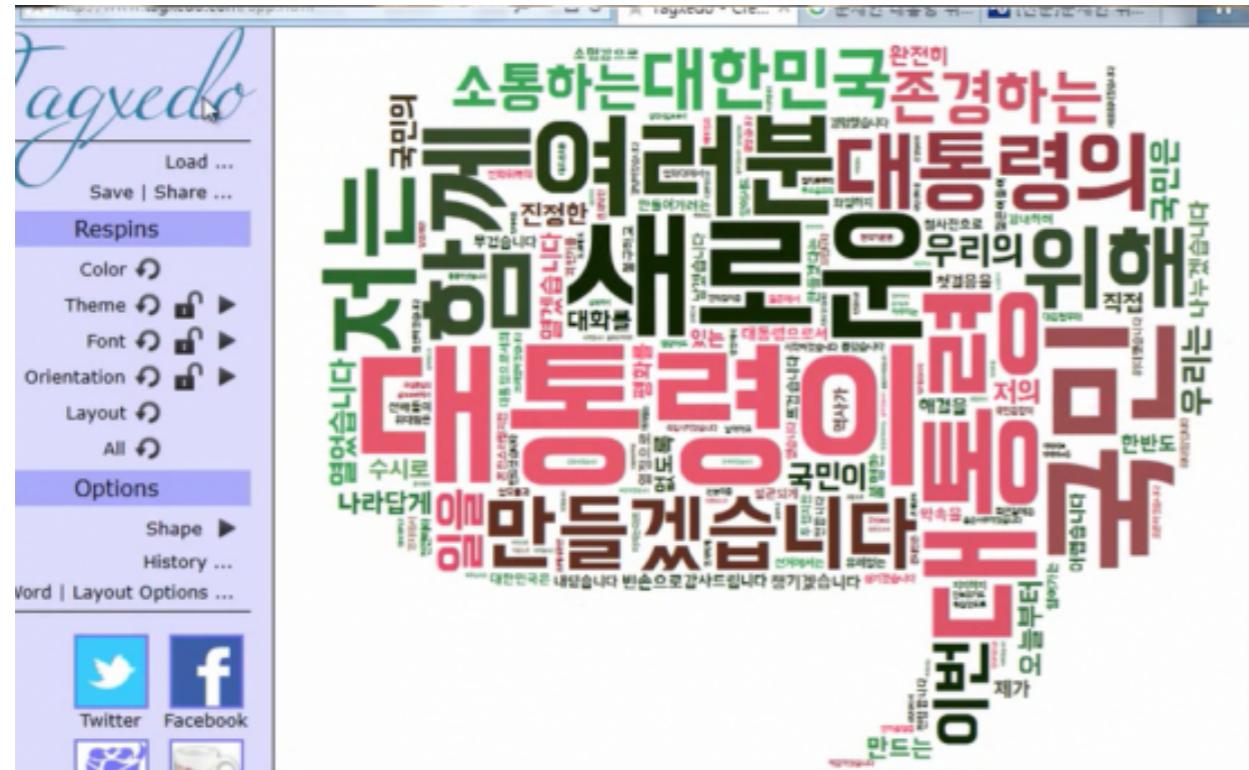
구분	내용
원문	<ul style="list-style-type: none">. 여러분 안녕하세요. 재미있는 텍스트 마이닝 수업입니다.
형태소 분석	<ul style="list-style-type: none">. 여러분/NP + 안녕/NNG + 하세요/EF + ./SF. 재미있/VA + 는/ETM + 텍스트/NNG + 마이닝/NNG + 수업/NNG + 입니다/EF + ./SF

형태소 분석 (Part of Speech Tagging)

대분류	세종 품사 태그		KKMA 단일 태그 V 1.0					
	태그	설명	묶음 1	묶음 2	태그	설명	활용태그	저장사전
체언	NNG	일반 명사	N	NN	NNG	보통 명사	NNA	noun.dic
	NNP	고유 명사			NNP	고유 명사		
	NNB	의존 명사			NNB	일반 의존 명사		
	NNM	단위 의존 명사			NNM	단위 의존 명사		
	NR	수사		NR	NR	수사		
용언	NP	대명사			NP	NP	대명사	NP
	VV	동사	V	VX	VV	동사	VV	verb.dic
	VA	형용사			VA	VA	형용사	
	VX	보조 용언		VC	VXV	보조 농사	VX	
	VCP	긍정 지정사			VXA	보조 형용사	VX	
관형사	VCN	부정 지정사			VCP	긍정 지정사, 서술격 조사 '이다'	VCP	raw.dic
	MM	관형사	M	MD	MDT	일반 관형사	MD	
	MDN	수 관형사			MDN	수 관형사	MD	
부사	MAG	일반 부사	MA	MAG	MAG	일반 부사	MAG	simple.dic
	MAJ	접속 부사			MAC	접속 부사	MAC	
감탄사	IC	감탄사	I	IC	IC	감탄사	IC	simple.dic
조사	JKS	주격 조사	J	JK	JKS	주격 조사	JKS	
	JKC	보격 조사			JKC	보격 조사	JKC	
	JKG	관형격 조사			JKG	관형격 조사	JKG	
	JKO	목적격 조사			JKO	목적격 조사	JKO	
	JKB	부사격 조사			JKM	부사격 조사	JKM	
	JKV	호격 조사			JKI	호격 조사	JKI	
	JKQ	인용격 조사			JKQ	인용격 조사	JKQ	
	JX	보조사		JC	JX	보조사	JX	
	JC	접속 조사			JC	접속 조사	JC	
선어말 어미	EP	선어말 어미	EP	EPH	존칭 선어말 어미	EP	raw.dic	Symbol class
				EPT	시제 선어말 어미			
				EPP	공손 선어말 어미			

대분류	세종 품사 태그		KKMA 단일 태그 V 1.0						
	태그	설명	묶음 1	묶음 2	태그	설명	활용태그	저장사전	
어말 어미	EF	종결 어미	E	EF	EFN	평서형 종결 어미	EF	simple.dic	
	EFQ	의문형 종결 어미			EFO	명령형 종결 어미			
	EFA	청유형 종결 어미			EFI	감탄형 종결 어미			
	EFR	존칭형 종결 어미			ECE	대등 연결 어미	EC		
	ECD	의존적 연결 어미			ECS	보조적 연결 어미			
	ETN	명사형 전성 어미			ETM	관형형 전성 어미			
접두사	XPN	체언 접두사	XP	XPN	XPN	체언 접두사	XP	simple.dic	
	XPV	용언 접두사			XPV	용언 접두사			
접미사	XSN	명사 파생 접미사	XS	XSN	XSV	동사 파생 접미사	XS		
	XSV	동사 파생 접미사			XSA	형용사 파생 접미사			
	XSA	형용사 파생 접미사			XSM	부사 파생 접미사			
	XSM	부사 파생 접미사			XSO	기타 접미사			
	XSO	기타 접미사			XR	어근	XR		
부호	XR	어근	S	SF	SF	마침표물음표, 느낌표	SF	Symbol class	
	SP	쉼표, 가운뎃점, 콜론, 빛금			SP	쉼표, 가운뎃점, 콜론, 빛금			
	SS	따옴표, 괄호표, 줄표			SS	따옴표, 괄호표, 줄표			
	SE	줄임표			SE	줄임표			
	SO	불임표(물결, 숨김, 빠짐)			SO	불임표(물결, 숨김, 빠짐)			
	SW	기타기호 (논리수학기호, 화폐기호)			SW	기타기호 (논리수학기호, 화폐기호)			
	NF	명사추정범주			UN	명사추정범주	NNA		
분석 불능	NV	용언추정범주	U	UV	UV	용언추정범주	N/A	N/A	
	NA	분석불능범주			UE	분석불능범주	N/A		
	SL	외국어			OL	외국어	NNA		
한글 이외	SH	한자	O	OH	OH	한자	NNA	N/A	
	SN	숫자			ON	숫자	NR		

형태소 분석 (Part of Speech Tagging)



*Source : 이정훈, 텍스트의 시각화: 단어 구름 (태그 클라우드), 2016.12.29., <http://visualsoft.kr/tag-cloud/>.

**Source : NÉSTOR CORREA, Cómo implementar el Big Data en tu empresa, 2017., <http://bluelight.tistory.com/298>

***Source : 몬데이터, [mondata] 남북정상회담 판문점 선언 Text 키워드 분석, 2018.4.28., <https://www.youtube.com/watch?v=ba4EMdzSK-A/>.

개체명 인식 (Named Entity Recognition)

문장에서 하나의 개체로써 인식되어야하는 단어를 구별하는 과정

- ▶ 데이터에서 개체명을 구별하고 태깅함(지명, 사명, 인물명, 약자, 기관명 등)
- ▶ 사전 기반의 개체명 인식에서 개체명은 매일 새롭게 생겨나고 변형되므로, 개체명 사전을 유지하는 것이 매우 중요함
- ▶ 분석의 목적에 따라서 머신러닝 기반의 개체명 인식을 사용할 수 있으나 새로 생겨나거나 변형되는 단어에 취약함

#KoNLPy 형태소 분석

```
from konlpy.tag import Kkma
```

```
text = “호날두 한명이 주는 효과가 세리에 전체 인기도 영향을 미치다니.. 역시 개드립월클의 힘”
```

```
...
```

```
kkma = Kkma()
```

```
pos_result = kkma.pos(text)
```

```
...
```

Result :

```
[(호, NNG), (날, NNG), (두, MDN), (한명, NNG), (이, JKS), (줄,VV), (는, ETD), (효과, NNG), ... ,  
(세리, NNG), (에, JKM), (전체, NNG), (인기도, NNG), (영향, NNG), (을, JKO), (미, NNG), ... ,  
(역시, MAG), (개, NNG), (드립, UN), (월, NNM), (크, VA), (ㄹ, ETD), (의, NNG), (힘, NNG)]
```

개체명 인식 (Named Entity Recognition)

문장에서 하나의 개체로써 인식되어야하는 단어를 구별하는 과정

- ▶ 데이터에서 개체명을 구별하고 태깅함(지명, 사명, 인물명, 약자, 기관명 등)
- ▶ 사전 기반의 개체명 인식에서 개체명은 매일 새롭게 생겨나고 변형되므로, 개체명 사전을 유지하는 것이 매우 중요함
- ▶ 분석의 목적에 따라서 머신러닝 기반의 개체명 인식을 사용할 수 있으나 새로 생겨나거나 변형되는 단어에 취약함

#Eucalyptus 형태소 분석

```
from Eucalyptus.NerTagger import NerTagger
```

```
input_file, output_file = "output_pos.txt", "output_ner.txt"  
ner_tagger = euc.NerTagger(input_file, output_file)  
ner_tagger.tagging()
```

Result (output_ner.txt):

...

{result:

```
[(호날두, NNP, Person), (한명, NNG), (0|, JKS), (줄, VV), (는, ETD), (효과, NNG), (가, JKS),  
(세리에, NNP, Sports), (에, JKM), (전체, NNG), (인기도, NNG), (영향, NNG), (을, JKO), ... ,  
(역시, MAG), (개드립월클, NNG, Neologism), (의, NNG), (힘, NNG)]}
```

개체명 사전 (NER Corpus)

[단순 개체명 사전]

구분	의학	인물	고유명사	블록체인
1	불량 식품	사나	서울플랜트엔지니어링	블록체인
2	진행 암	쯔위	서울플리머	블럭체인
3	전진 피판	정연	서울피브이시상사	비트코인
4	유해 효과	나연	서울피브이씨	이더리움
5	무력증	황민현	서울피비씨	알트코인
6	유산소 운동	강다니엘	서울피앤씨	추격매수
7	산소 호흡	옹성우	서울하이테크	풀매수
8	공기 삼킴증	전병진	서울학연구	총알
9	분무제	진상형	고광엔지니어링	운전수
10	에어로졸	서지석	서울합금	고점
11	분무 주입법	배현진	서울합판	저점
12	대기 요법	현빈	서울합판목재상사	장투
13	정동 장애	진세연	서울합판상사	단타
14	정감성	남지현	서울해체산업	떡상
15	정동성	주상욱	서울행정신문사	떡락
16	들신경	김태희	서울행정학회	횡보
17	협력 병원	허맹호	서울화성	손절
18	친화력	유아인	서울화인테크	익절
19	친화 크로마토그래피	이승기	서울화학	반등
20	무섬유소원 혈증	한예슬	고광훈	패닉셀

[부가정보를 포함하는 개체명 사전]

구분	지역명	영문 지역명	구분
1	서울	Seoul	Metropolitan
2	종로	Jongno	district
3	중	Jung	district
4	용산	Yongsan	district
5	성동	Seongdong	district
6	광진	Gwangjin	district
7	동대문	Dongdaemun	district
8	중랑	Jungnang	district
9	성북	Seongbuk	district
10	강북	Gangbuk	district
11	도봉	Dobong	district
12	노원	Nowon	district
13	은평	Eunpyeong	district
14	서대문	Seodaemun	district
15	마포	Mapo	district
16	양천	Yangcheon	district
17	강서	Gangseo	district
18	구로	Guro	district
19	금천	Gumcheon	district
20	영등포	Yeongdeungpo	district

원형복원 & 불용어 처리

원형복원

- ▶ 변형된 단어의 원형을 복원하는 과정
- ▶ Stemming : 규칙 기반으로 단어의 변형된 형태를 제거
- ▶ Lemmatizing : 사전 기반으로 품사에 맞는 단어의 원형으로 변환

불용어 처리

- ▶ 분석에 불필요한 단어나 방해가 되는 단어를 제거하는 과정
- ▶ 주로 불용어 (Stop-word) 또는 최빈어 (Common-word)가 제거됨
- ▶ 하다, 이다, is, 기자, 뉴스

[Stemming & Lemmatizing 비교]

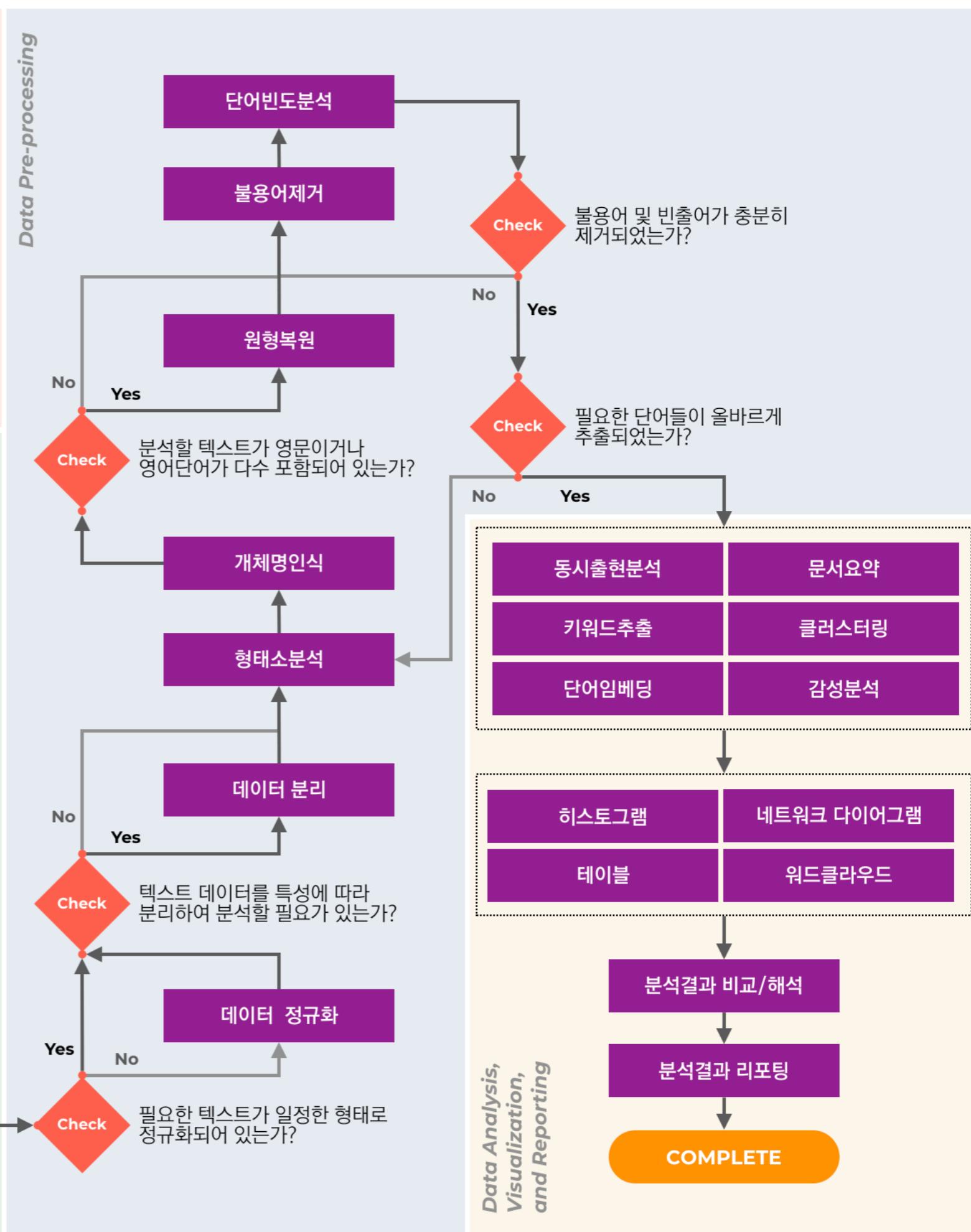
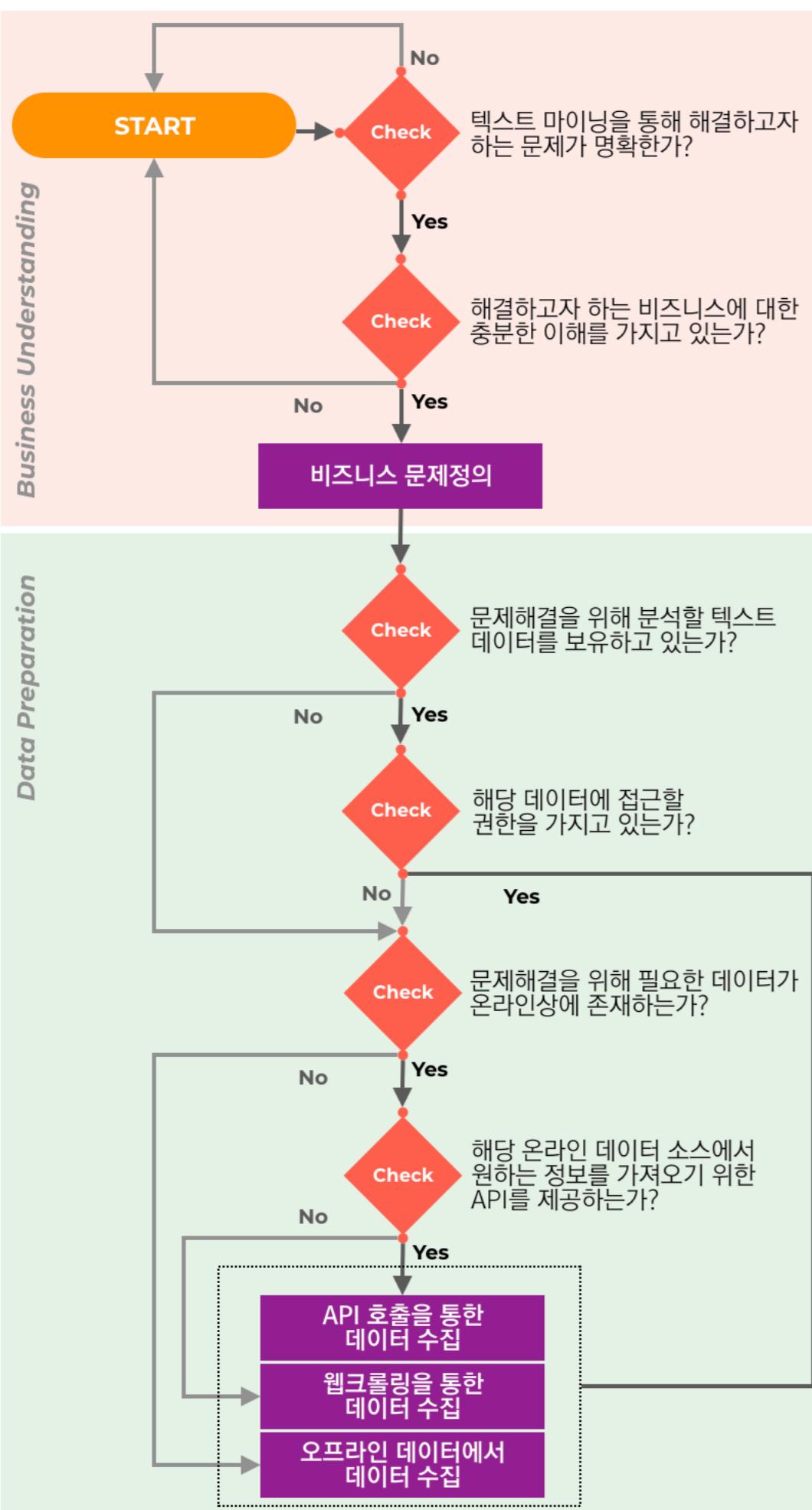
Word	Stemming	Lemmatizing
cooking(v)	cook	cook
cooking(n)	cook	cooking
cookbooks	cookbook	cookbook
believes	believ	believe
using	us	use

Note. 복원 결과는 Stemmer와 Lemmatizer의 종류에 따라서 다를 수 있음

과정	결과
개체명 인식	텍스트 마이닝/NNG/IT + 은/JX + 길/VA + 고/ECE + 지겹/VA + ㄴ/ETD + 작업/NNG + 이/VCP + ㅂ니다/EFN + .SF Text mining/NN/IT + is/VBZ + difficult/JJ + but/CC + very/RB + valuable/JJ + ./.
원형 복원	텍스트 마이닝/NNG/IT + 은/JX + 길다/VA + 고/ECE + 지겨운/VA + 작업/NNG + 이/VCP + ㅂ니다/EFN + .SF Text mining/NN/IT + be/VBZ + difficult/JJ + but/CC + very/RB + valuable/JJ + ./.
불용어 제거	텍스트 마이닝/NNG/IT + 길다/VA + 지겨운/VA + 작업/NNG Text mining/NN/IT + difficult/JJ + valuable/JJ

텍스트 데이터 분석

어렵고 복잡한 분석을 할수록
분석결과는 해석하기 어렵고
복잡해진다



단어 빈도분석 (Word Frequency)

단어의 빈도를 바탕으로 가중치를 계산하는 분석방법

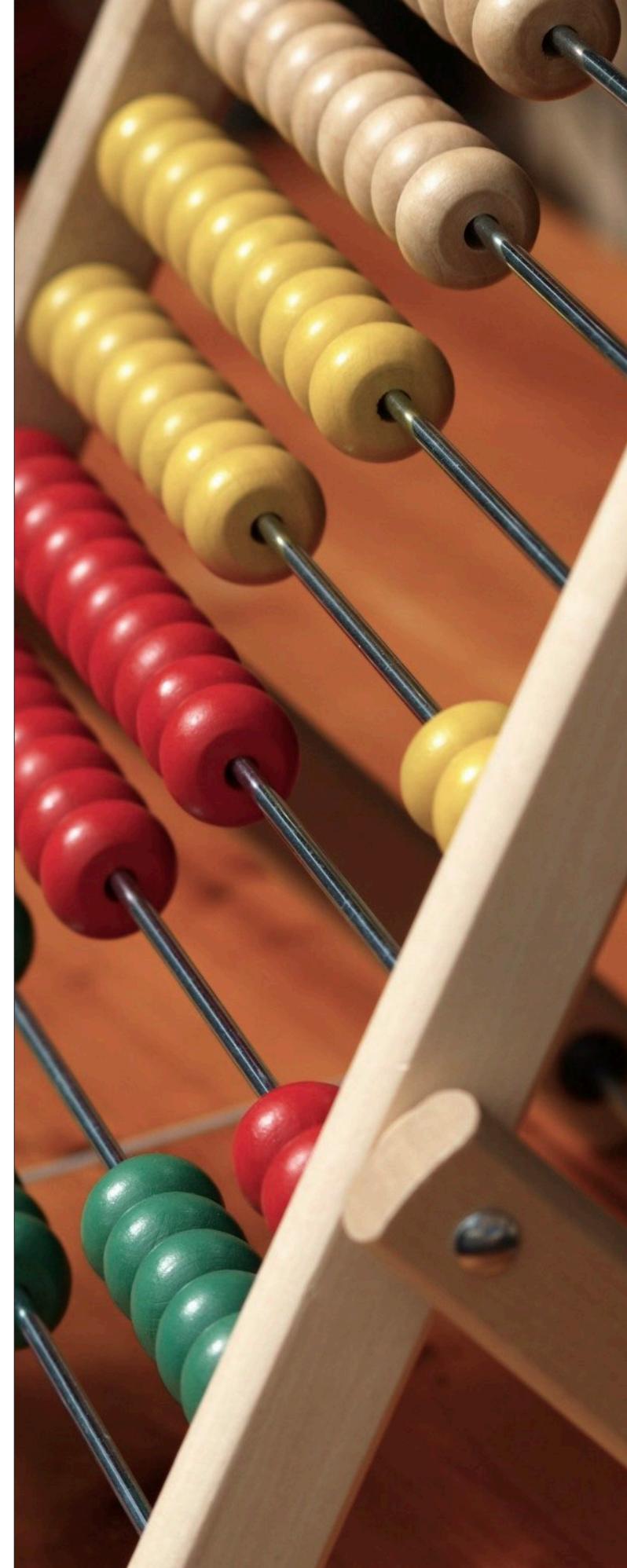
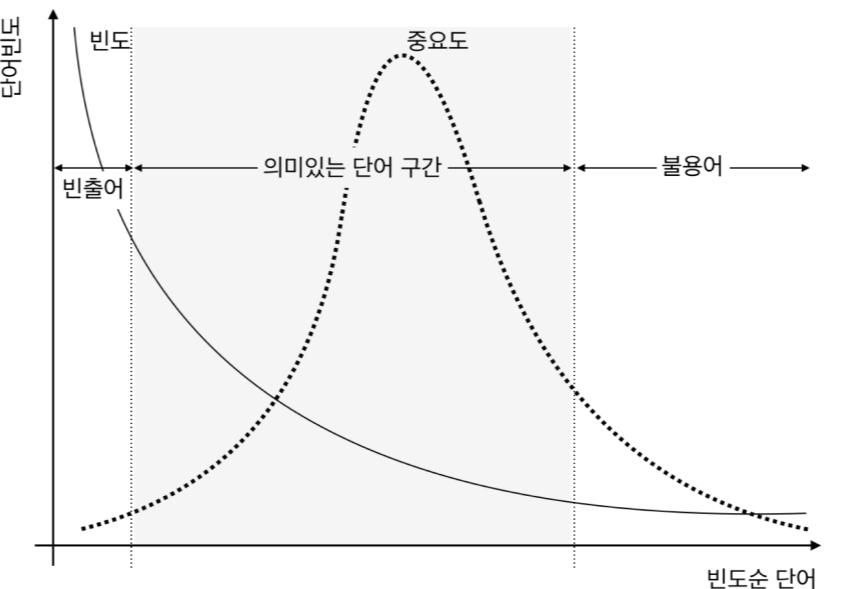
- ▶ 문서에 출현한 단어들의 출현 횟수를 기준으로 단순빈도, TF, TF-IDF 등을 계산
- ▶ 가장 간단한 텍스트 데이터 분석 방법이지만, 가장 빠르게 문서를 파악할 수 있으며 다른 분석방법을 수행하기 전에 반드시 한번이상 거쳐야 하는 과정
- ▶ 활용분야 : 문서요약, 내용 파악, 트렌드 분석, 불용어/빈출어 발견

단순 단어빈도 (Term Frequency, TF)

- ▶ 단어가 전체 문서에서 얼마나 흔하게 출현하는지 고려하는 방법
- ▶ 너무 희귀한 단어인 경우 또는 딱 한 번 나오는 단어는 의미를 부여하기 어려움
- ▶ 단어가 너무 흔한 경우 의미가 과도하게 부여될 가능성이 있음

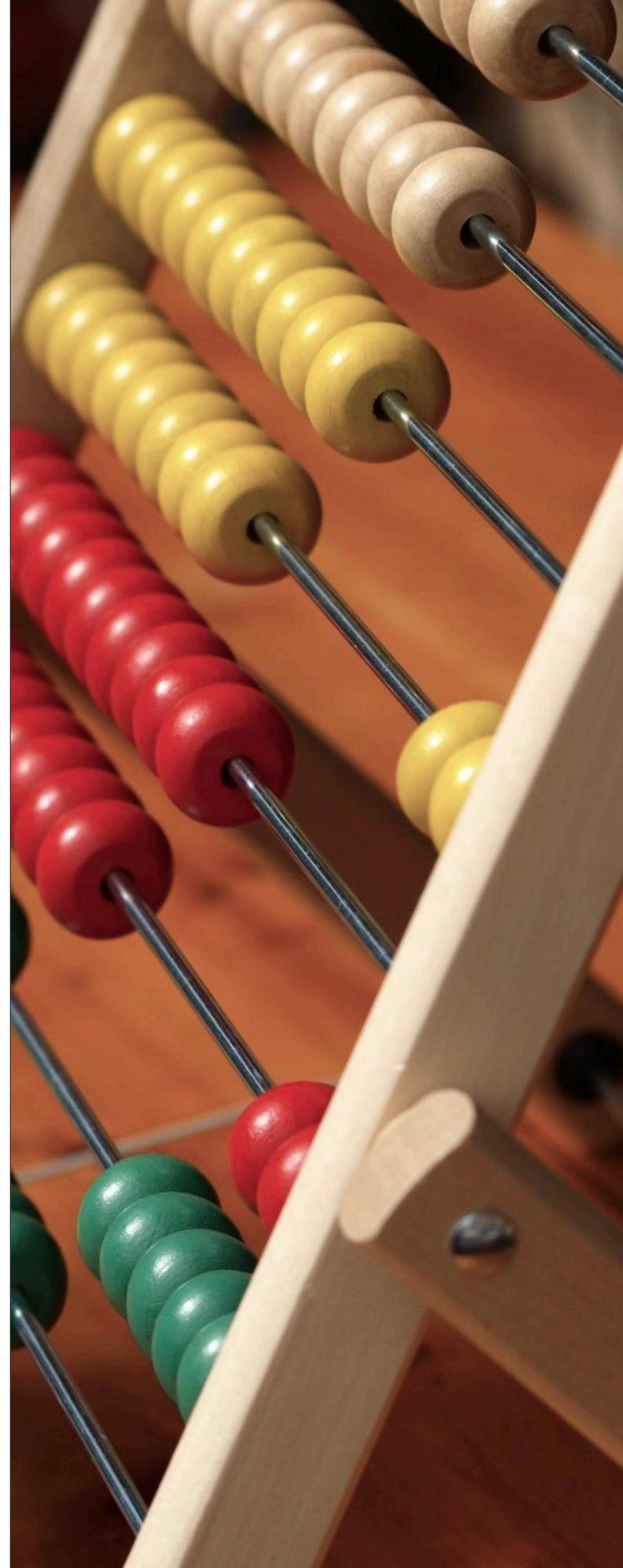
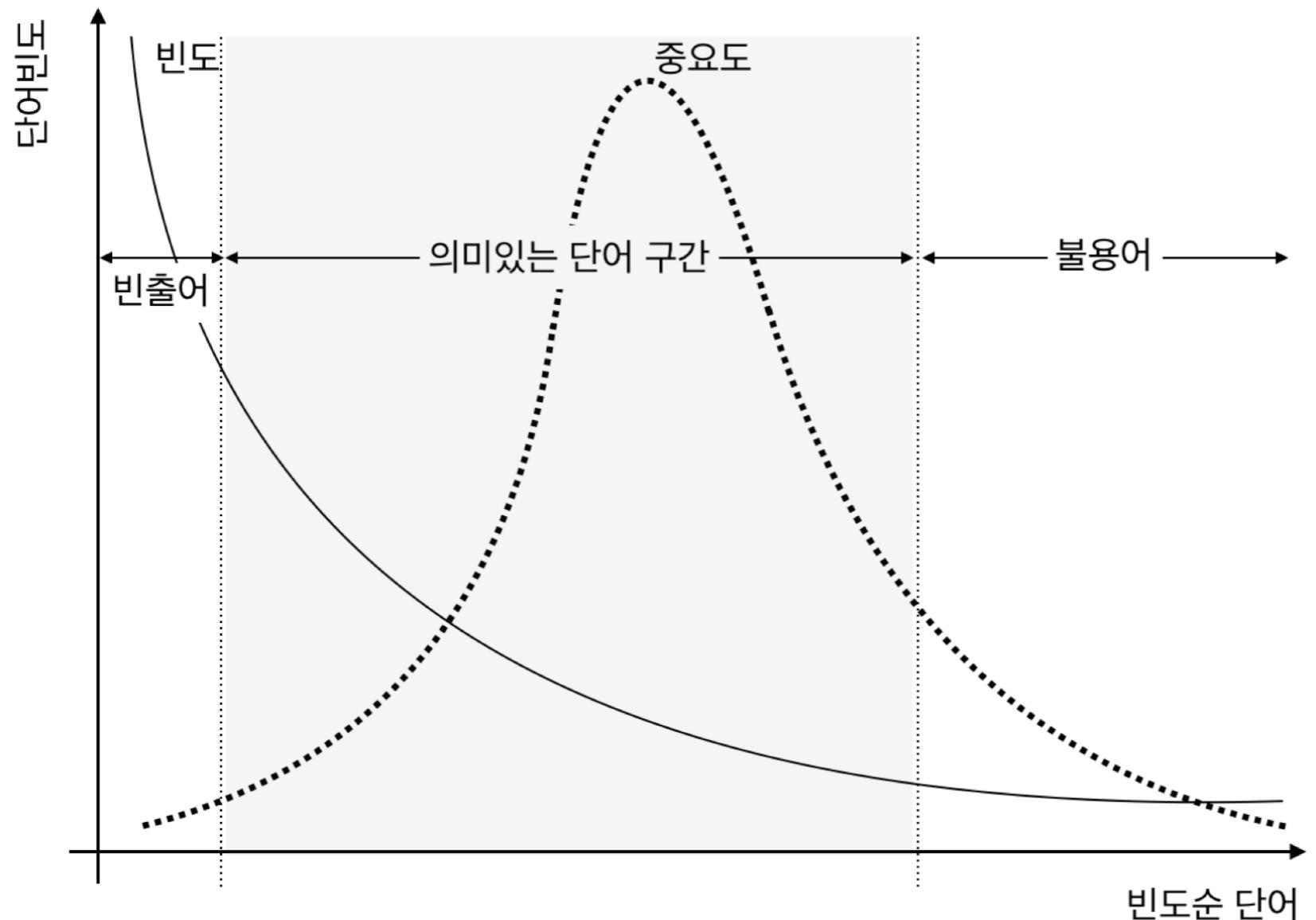
TermFrequency

$$= \text{count}(word | document)$$



단어 빈도분석

(Word Frequency)



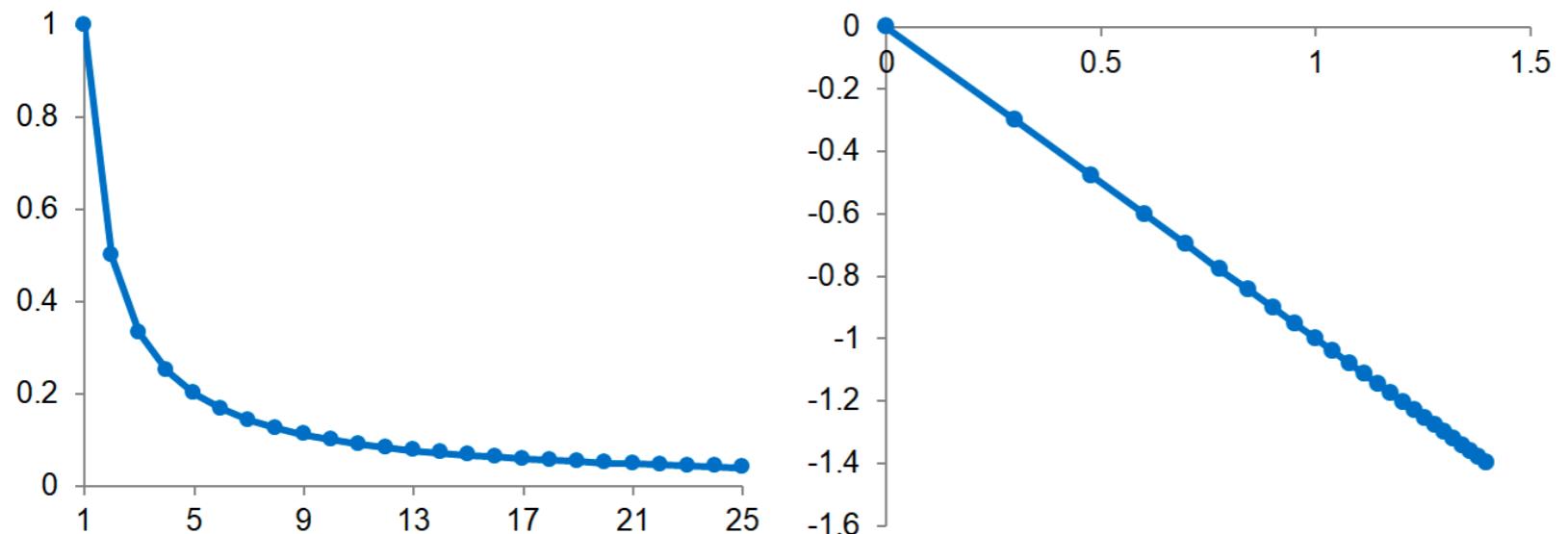
단어 빈도분석 (Word Frequency)

TF-IDF (TF-Inverse Document Frequency)

- ▶ 단어가 나온 문서의 수가 적을수록 단어가 문서에 중요할 가능성이 더 큼
- ▶ 단어의 희박성(sparseness)을 역문서빈도(IDF)로 측정
- ▶ 전체 문서 수가 고정된 체로 단어 t가 출현하는 문서가 많을수록 중요도가 감소

$$TF - IDF = Frequency * IDF$$

$$IDF = \log\left(\frac{N}{n_t} + 1\right)$$



- ▶ 지프의 법칙 (Zipf's law)

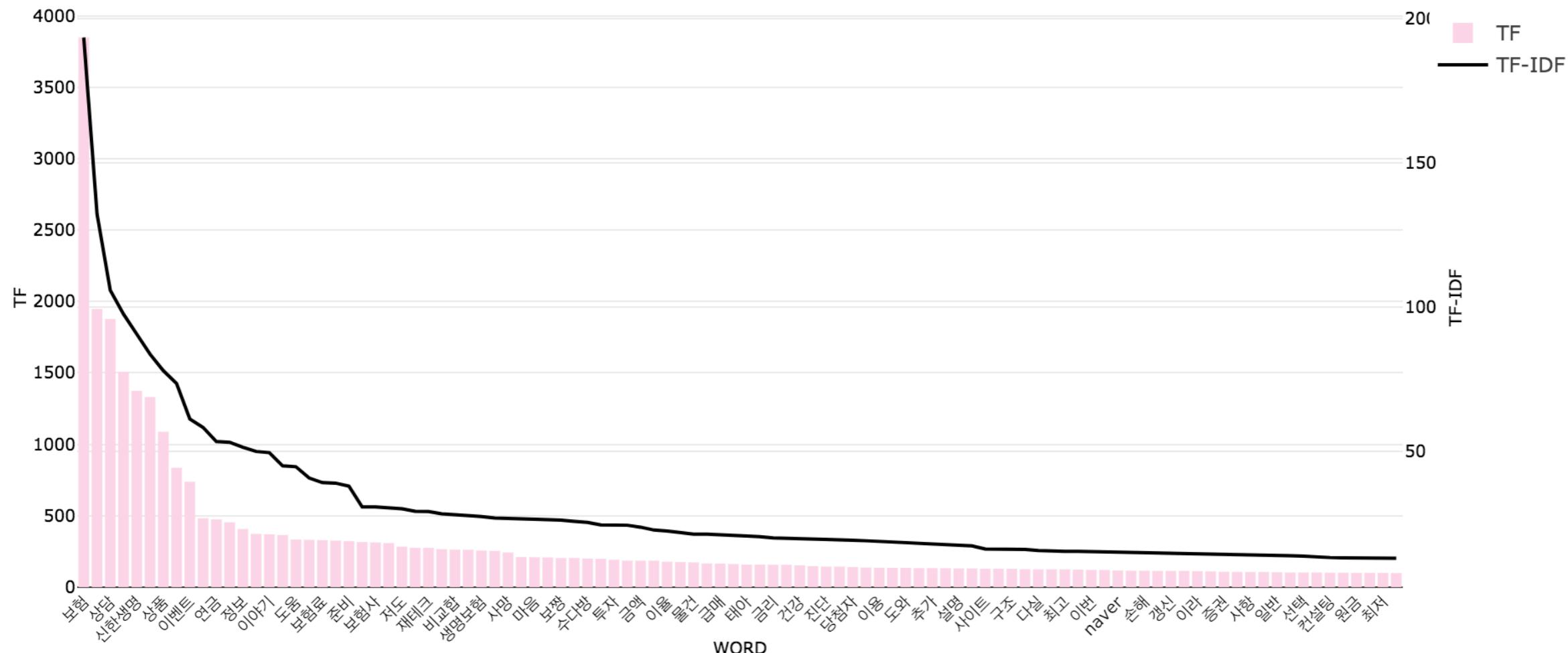
자연어 말뭉치 표현에 나타나는 단어들을 그 사용 빈도가 높은 순서대로 나열하였을 때, 모든 단어의 사용 빈도는 해당 단어의 순위에 반비례한다.

가장 사용 빈도가 높은 단어는 두 번째 단어보다 빈도가 약 두 배 높으며, 세 번째 단어보다는 빈도가 세 배 높다.



단어 빈도분석 (Word Frequency)

TF & TF-IDF Graph



동시출현분석

(Co-occurrence Analysis)

단어의 동시출현 빈도를 바탕으로 가중치를 계산하는 분석방법

- 단어들 사이의 동시 출현을 연관성으로 취급하여, 단어의 연관성을 파악하는 방법
- 상용 소셜 미디어 분석 솔루션에서 제공하는 가장 기본적인 분석 형태 (Social Matrix)
- 활용분야 : 브랜드 이미지 조사, 트렌드 분석, 여론조사, 마케팅 모니터링

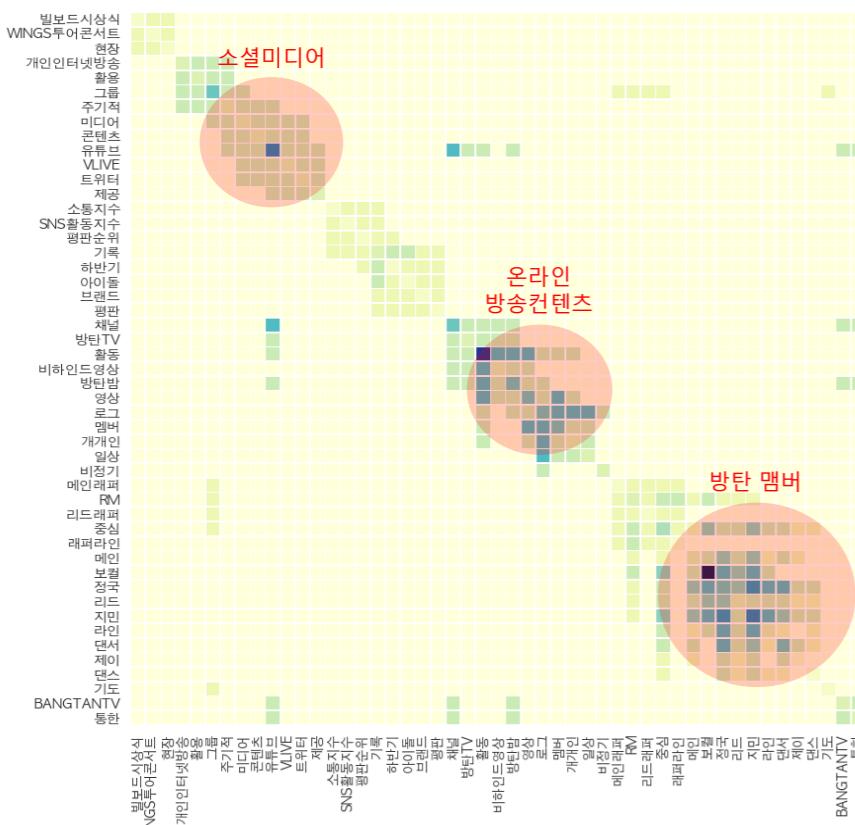


표 1 '아쿠르트 아줌마' 연관어 변화

아쿠르트 아줌마는 여전히 '아쿠르트'와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 아쿠르트 아줌마는 '배달하는' 역할에서 맛난 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화 중.

2013년		2014년		2015년		2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	아쿠르트	21.3%	1	아쿠르트	26.3%	1	아쿠르트	26.6%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%
11	아쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%
...				
29	팔다	1.4%	29	묻다	1.3%	29	배달하다	1.3%
29	구입하다	1.0%				29	구입하다	1.0%

■ 상승 키워드 ■ 하락 키워드 ■ 신규 키워드



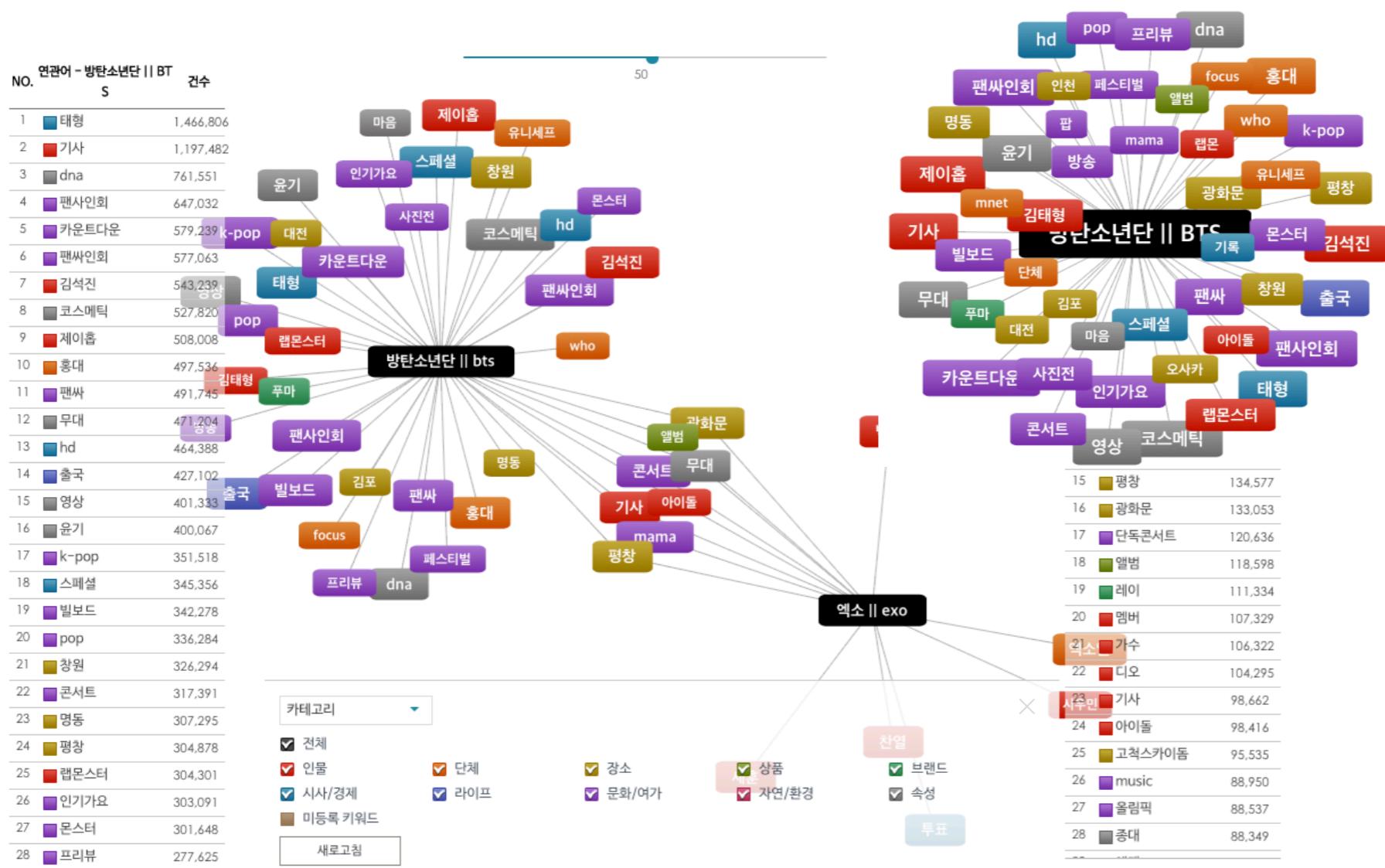
동시출현분석

(Co-occurrence Analysis)



연관어분석 (Co-word Analysis)

- ▶ 단어들 사이의 동시출현 빈도 중 빈번하게 사용되는 특정 단어를 기준으로 연관성을 파악하는 방법
 - ▶ 연관어(공기어, Co-word) : 같은 문맥안에서 함께 나타나 서로 밀접한 의미관계를 갖는 단어

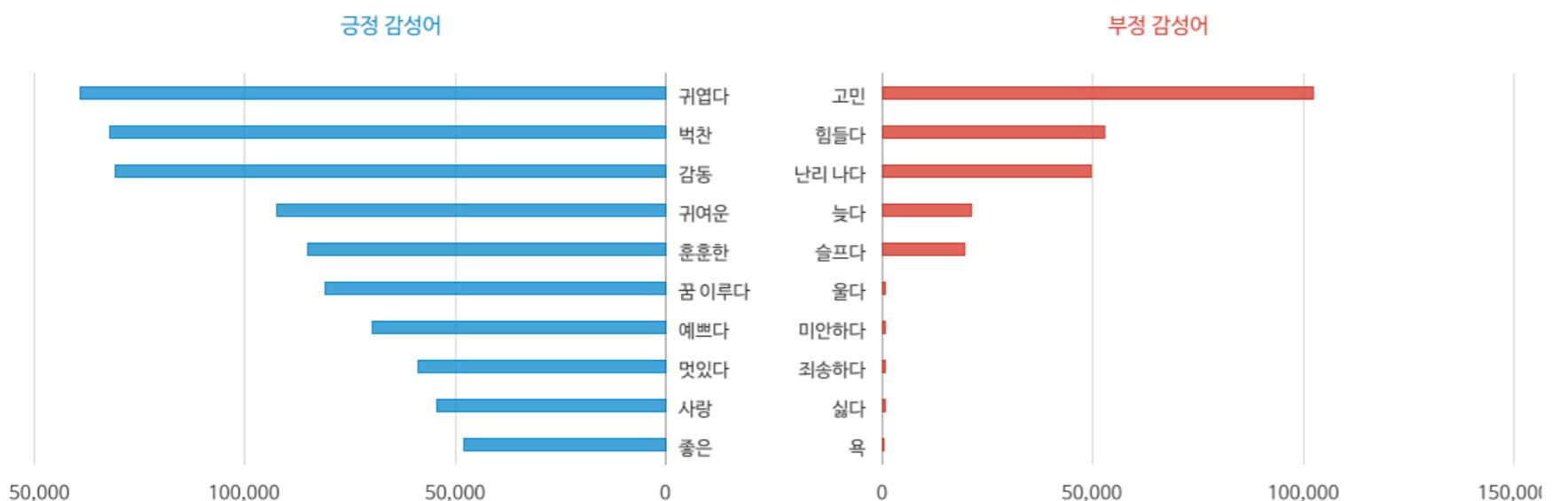


*Source : 소셜메트릭스, 2017.11.3., <http://www.socialmetrics.co.kr/>.

동시출현분석

(Co-occurrence Analysis)

감성 키워드 순위



기간별 연관어 순위: 방탄소년단 || BTS | 2017/10/03 ~ 2017/11/03

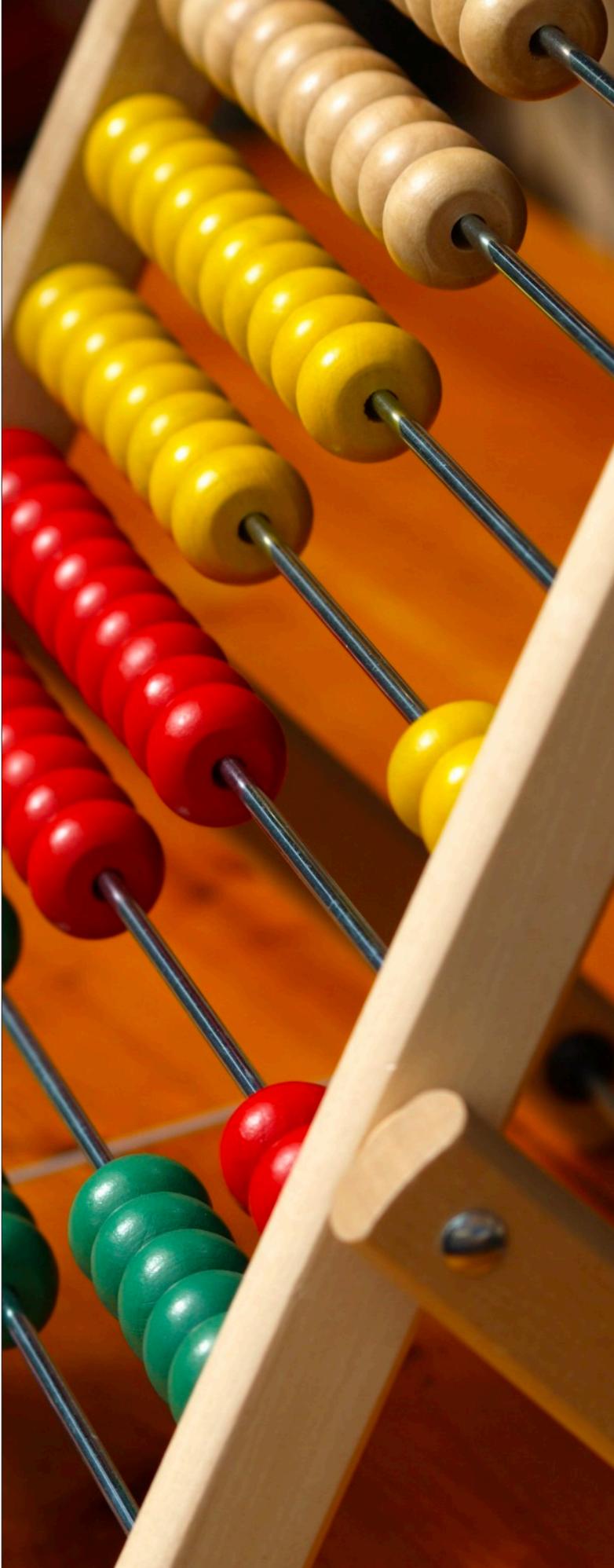
전체 트위터 블로그 커뮤니티 인스타그램 뉴스 확인

일별 주별 월별 분기별



카테고리 ▾

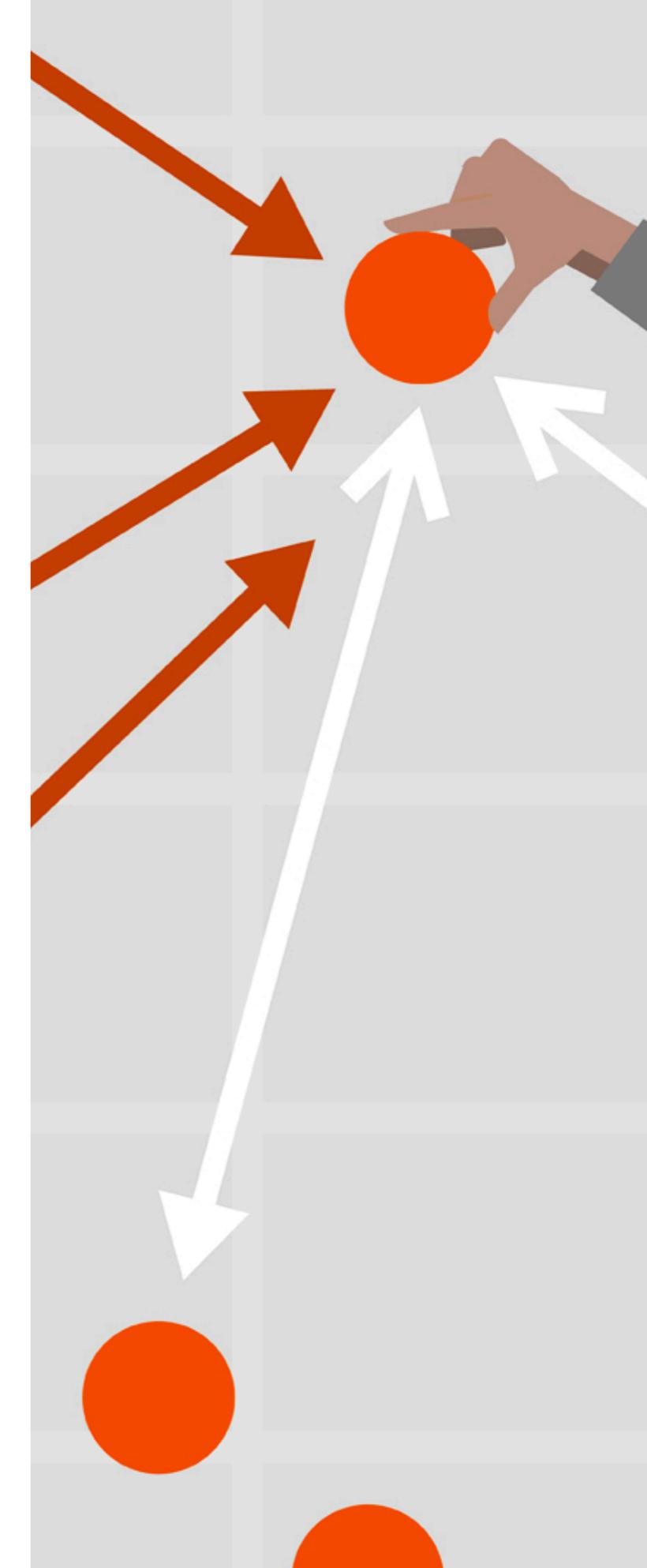
순위	2017/10/03~2017/10/07		2017/10/08~2017/10/14		2017/10/15~2017/10/21		2017/10/22~2017/10/28		2017/10/29~2017/11/03	
	연관어	건수								
1	방탄소년단	826,236	방탄소년단	987,650	방탄소년단	798,031	방탄소년단	549,312	방탄소년단	1,481,373
2	태형	449,263	태형	552,652	기사	260,201	김석진	147,539	평창	289,317
3	코스메틱	404,925	홍대	428,101	태형	217,125	태형	141,507	광화문	250,964
4	기사	260,803	기사	423,304	출국	205,573	hd	108,600	콘서트	214,887
5	명동	253,605	dna	362,319	푸마	187,739	무대	106,730	기사	205,638
6	팬사인회	207,776	카운트다운	355,310	김석진	165,320	타이페이	95,120	유니세프	205,279
7	팬싸인회	200,843	팬사인회	329,385	dvd	128,894	캡	92,580	캠페인	131,943
8	마음	198,856	영상	297,121	hd	127,530	제이홉	77,808	무대	119,974
9	who	186,415	팬싸인회	283,711	dna	124,366	콘서트	77,700	스페셜	119,423
10	dna	184,025	팬싸	222,826	제이홉	117,073	윤기	76,071	리허설	113,370
11	팬싸	175,731	출국	197,729	mama	107,045	대만	66,871	올림픽	109,404



토픽모델링 (Topic Modeling)

구조화되지 않은 방대한 문헌집단에서 주제를 찾아내기 위한 방법

- ▶ 뉴스, 블로그, 웹페이지, 기사 등의 형태로 온라인 상에서 방대한 양의 문서가 생성되고 저장되면서 사람들이 찾고자 하는 것을 발견하는 것이 어려워짐
- ▶ 방대한 양의 문서를 관리하고, 검색하고, 이해를 돋기 위한 새로운 도구로 텍스트 마이닝 기법이 주목받기 시작했으며, 그 중 문서 요약과 검색을 위해 토픽모델링 기법이 제안됨
- ▶ 맥락과 관련된 단어들을 이용하여 유사한 의미를 가진 단어들을 클러스터링하는 방식으로 주제를 추론하며, 같은 맥락에서 나타날 가능성이 있는 단어들을 그룹화함



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,³ two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

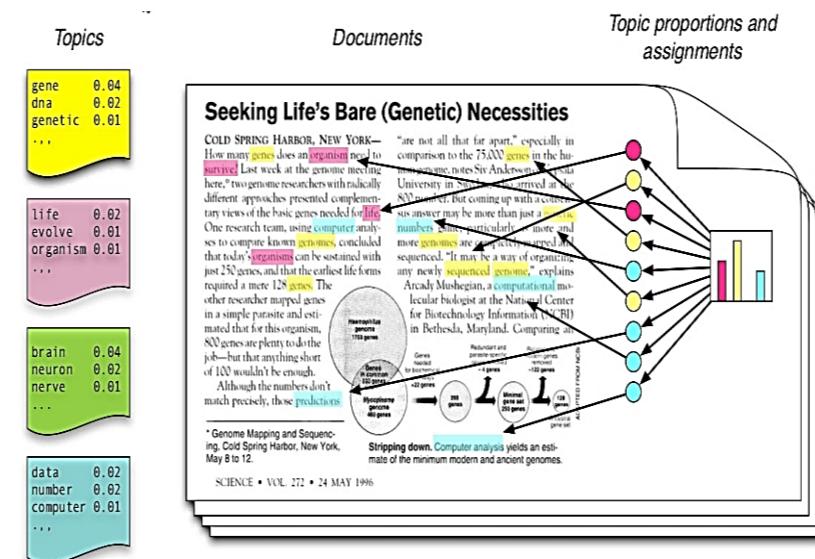
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

ancestor genome to a parasite genome, he found that the parasite genome contained 1703 genes, while the ancestor genome contained 469 genes. By removing redundant and parasite-specific genes, he was able to identify a minimal gene set of 250 genes.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

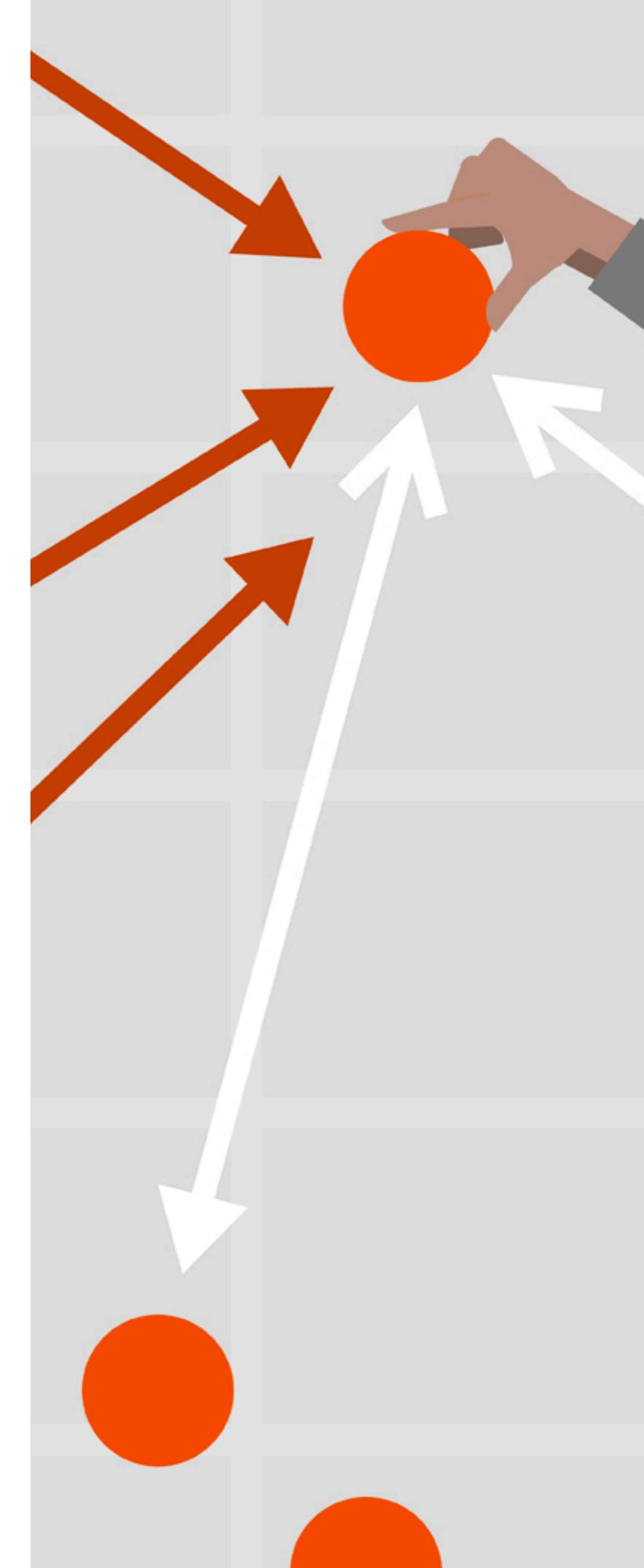
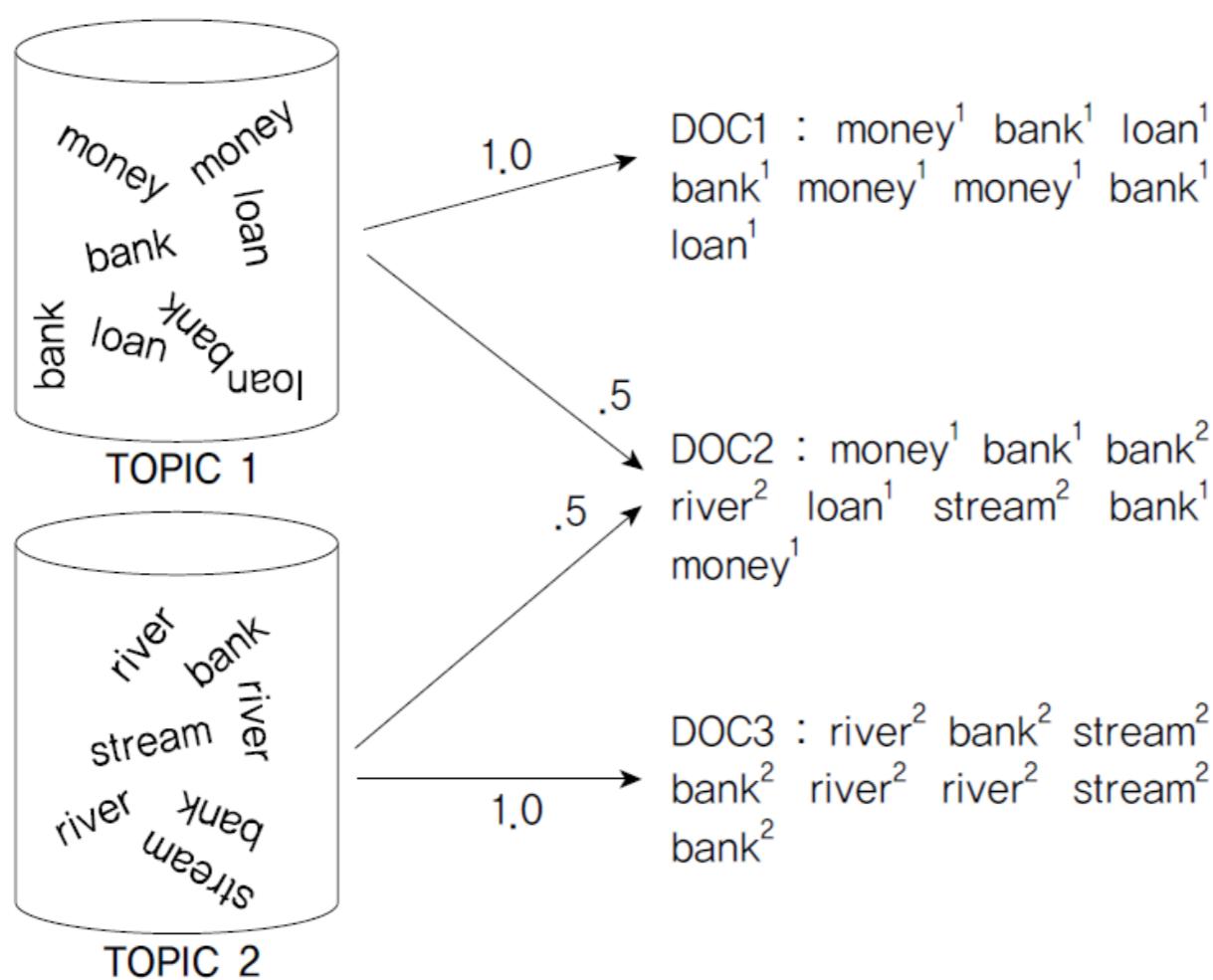
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



토픽모델링 (Topic Modeling)

토픽 모델링: 문헌 생성 모델

- ▶ 토픽 모델링의 문제점
 - 문헌 내의 용어 분포는 알 수 있지만 주제들 (Topic 1, Topic 2)의 용어분포를 사전에 알 수 없음
 - 직접 관찰할 수 있는 문헌 내 용어분포로부터 주제의 용어분포를 추정하는 과정이 필요
 - 잠재 디리클레 할당 (Latent Dirichlet Allocation, LDA) 기법 제안



토픽모델링 (Topic Modeling)

LDA 토픽 모델링

- ▶ 토픽 모델링 기법 중 텍스트 마이닝 분석에서 가장 많이 활용되고 있는 문헌 생성 모델 (generative probabilistic model)
- ▶ LDA (Latent Dirichlet Allocation)
 - 이미 관찰된 변수를 통해 각각의 확률을 계산하여 토픽을 생성하는 사후 추론방법
 - 특이값 분해 (singular value decomposition)를 활용한 LSI에서 발전됨
- ▶ LSI (Latent Semantic Index) : 용어-문헌 행렬의 차원을 축소하는 방법으로 문헌을 표현
- ▶ LDA 기법에서는 문헌 단위에서 각 주제들의 분포로 문헌을 표현

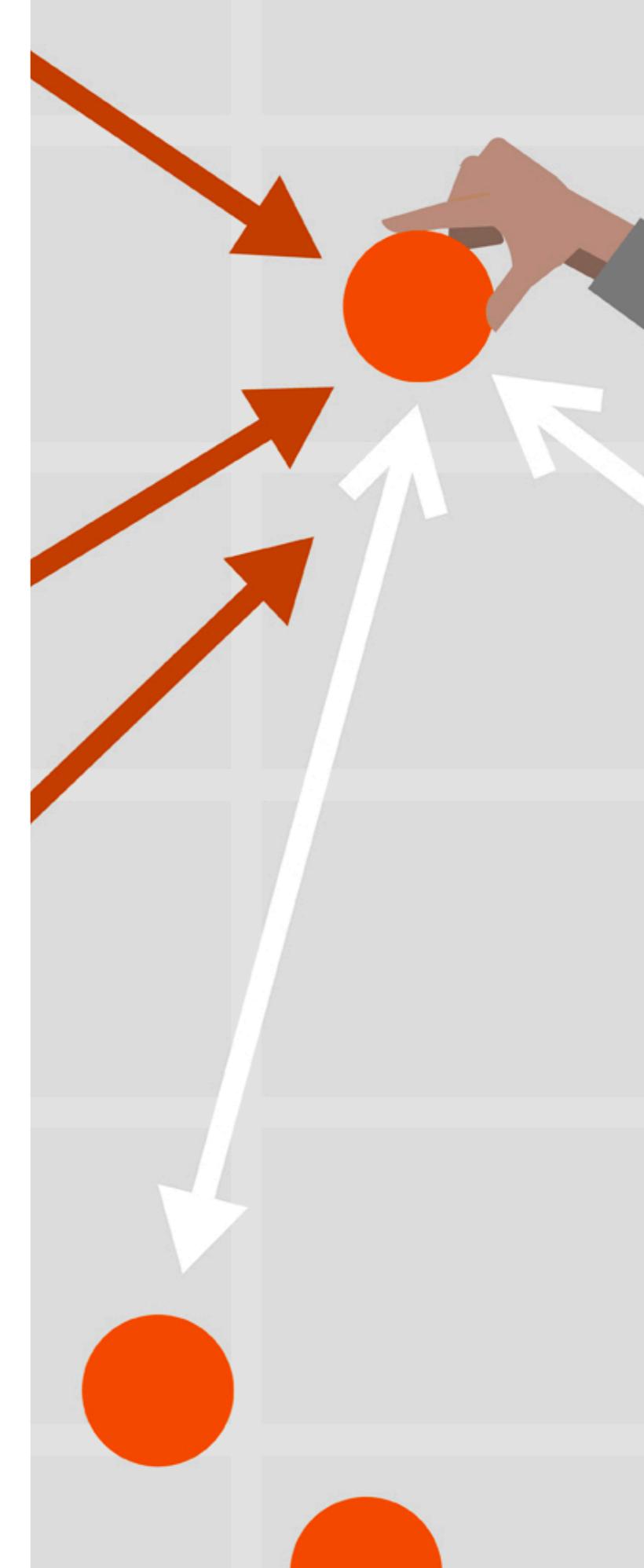
$$\text{LSI} \quad C = U \times \Sigma \times V^T$$

words \times docs words \times dims dims \times dims dims \times docs

$$\text{LDA} \quad C = \Phi \times \Theta$$

words \times docs words \times topics topics \times docs

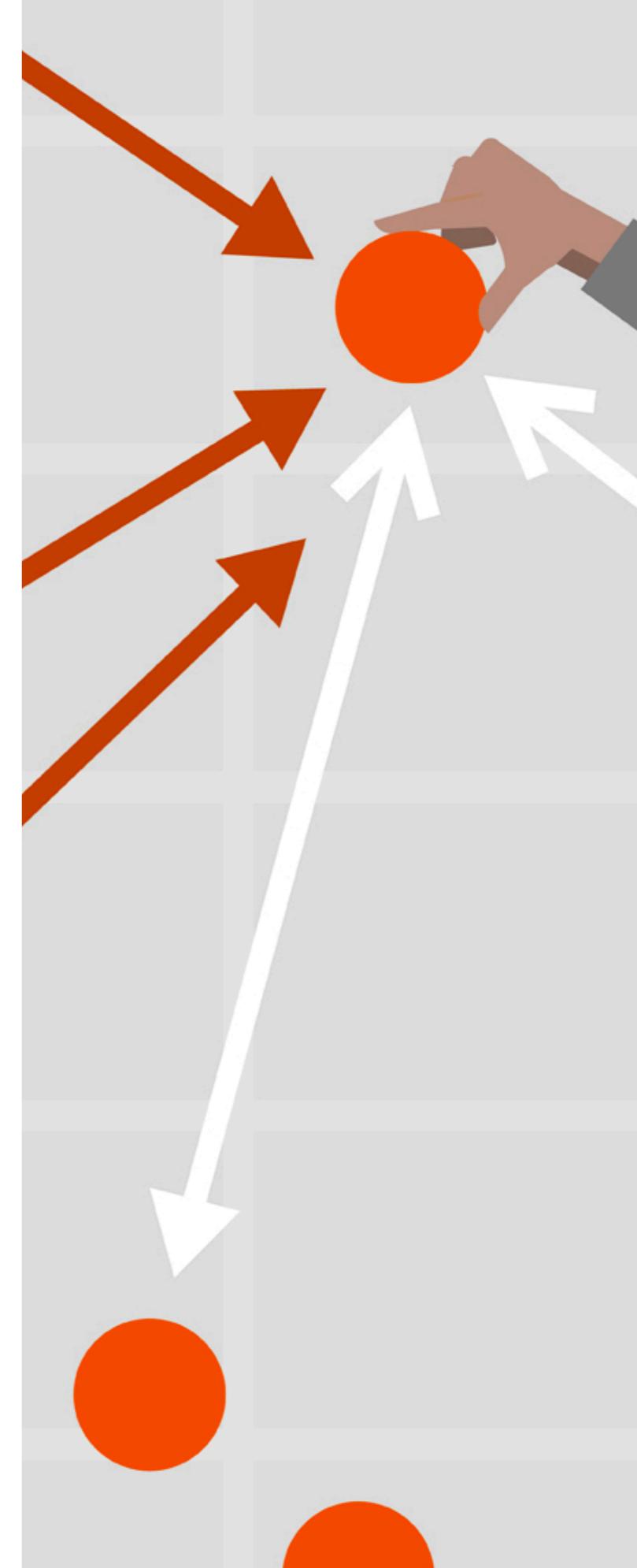
normalized co-occurrence matrix mixture component mixture weights



토픽모델링 (Topic Modeling)

LDA 토픽 모델링 수행과정

- ▶ 포아송분포로부터 임의의 문헌 길이 N 을 선택
 α 를 매개변수로 한 디리클레분포로부터 주제분포 θ 를 선택
(α : 문헌집단 내에서 정해지는 변수로 학습을 통해 추정됨)
 1. choose $N \sim Poisson(\xi)$
 - ▶ 어떤 문헌에 대해 주제 벡터인 θ 가 매개변수일 때, 앞에서부터 단어를 하나씩 채울 때마다 θ 로부터 하나의 주제($_z_n$)를 선택
 2. choose $\theta \sim Dir(\alpha)$
 - ▶ 다시 그 주제로부터 단어를 선택
 3. For each word in document:
 - (i) choose a topic $z_n \sim Multinomial(\theta)$
 - (ii) choose a word $w_n \sim Multinomial(\beta)$
- α : 문헌 내 주제분포 θ 를 추정하기 위한 매개변수
- β : 각 용어가 특정 주제에 할당될 사전 확률
- θ : 문헌 내에서 특정 주제가 할당될 사전 확률(prior probability)



감성분석 (Sentiment Analysis)

단어의 감성수준을 수치화하는 분석방법

- ▶ 문장이 의미하는 감성의 극성을 판별하거나 그 수준을 점수로 매기는 방법
- ▶ 텍스트 데이터를 계량 데이터로 바꾸는 가장 좋은 방법 중 하나
- ▶ 사전(말뭉치) 기반 감성분석과 머신러닝을 활용한 감성분석이 있음

[사전기반 감성분석]

A = I am not interested in class and have no fun.
B = Today class is very interesting and fun.



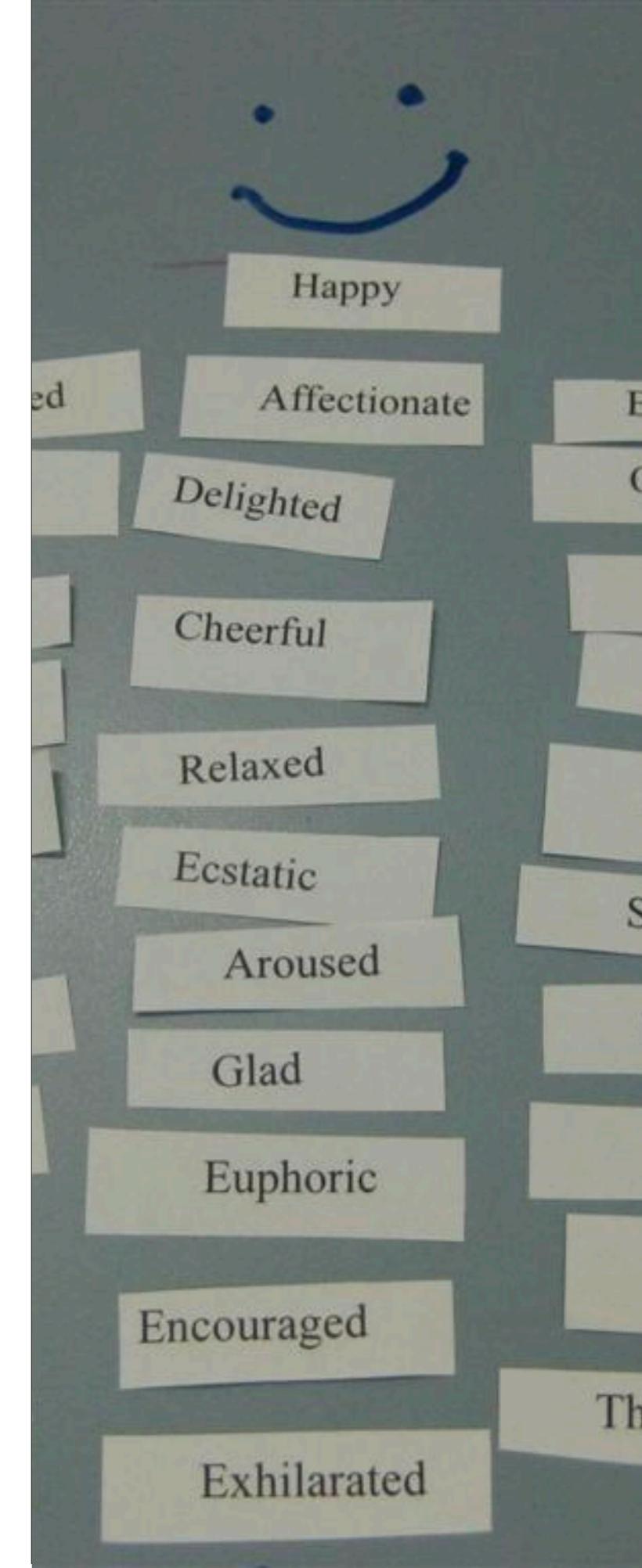
$A_{Senti} = [(i, 0), (be, 0), (not, x-1), (interest, 0.8), (in, 0), (class, 0), (and, 0), (have, 0), (no, x-1), (fun, 0.9)]$
 $B_{Senti} = [(Today, 0), (class, 0), (be, 0), (very, x2), (interesting, 0.8), (and, 0), (fun, 0.9)]$



$$A_{Score} = -1 \times 0.8 + -1 \times 0.9 = -1.7$$
$$B_{Score} = 2 \times 0.8 + 0.9 = 2.5$$

[감성사전 예시]

Word	Polarity	Weight
not	-	negation
no	-	negation
...
interest	+	0.8
fun	+	0.9
...
sorry	-	0.9
sad	-	0.8
...

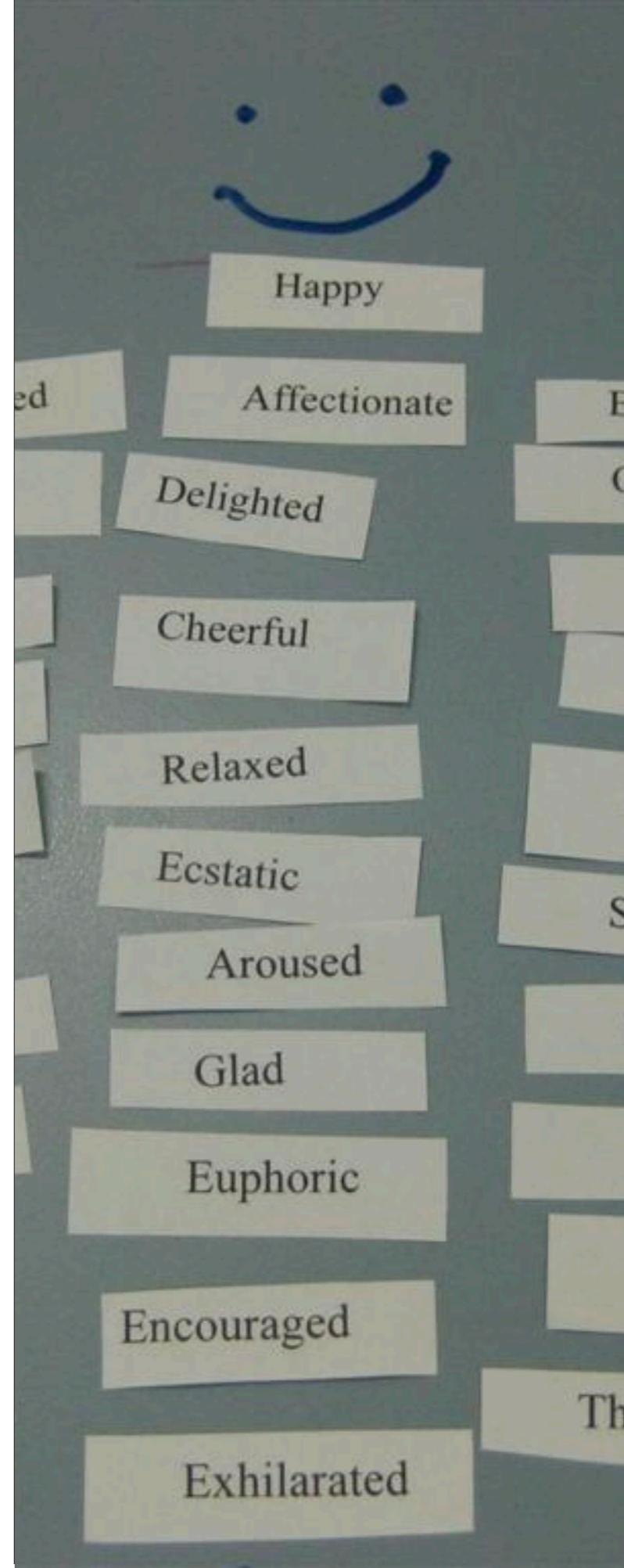


감성분석 (Sentiment Analysis)

상용/연구용 감성사전 종류

- ▶ Linguistic Inquiry and Word Count (LIWC) - <http://www.liwc.net/>
- ▶ MPQA Subjectivity Cues Lexicon - <http://www.cs.pitt.edu/>
- ▶ SentiWordNet - <http://sentiwordnet.isti.cnr.it/>
- ▶ KOSAC - <http://word.snu.ac.kr/kosac/icon.php>

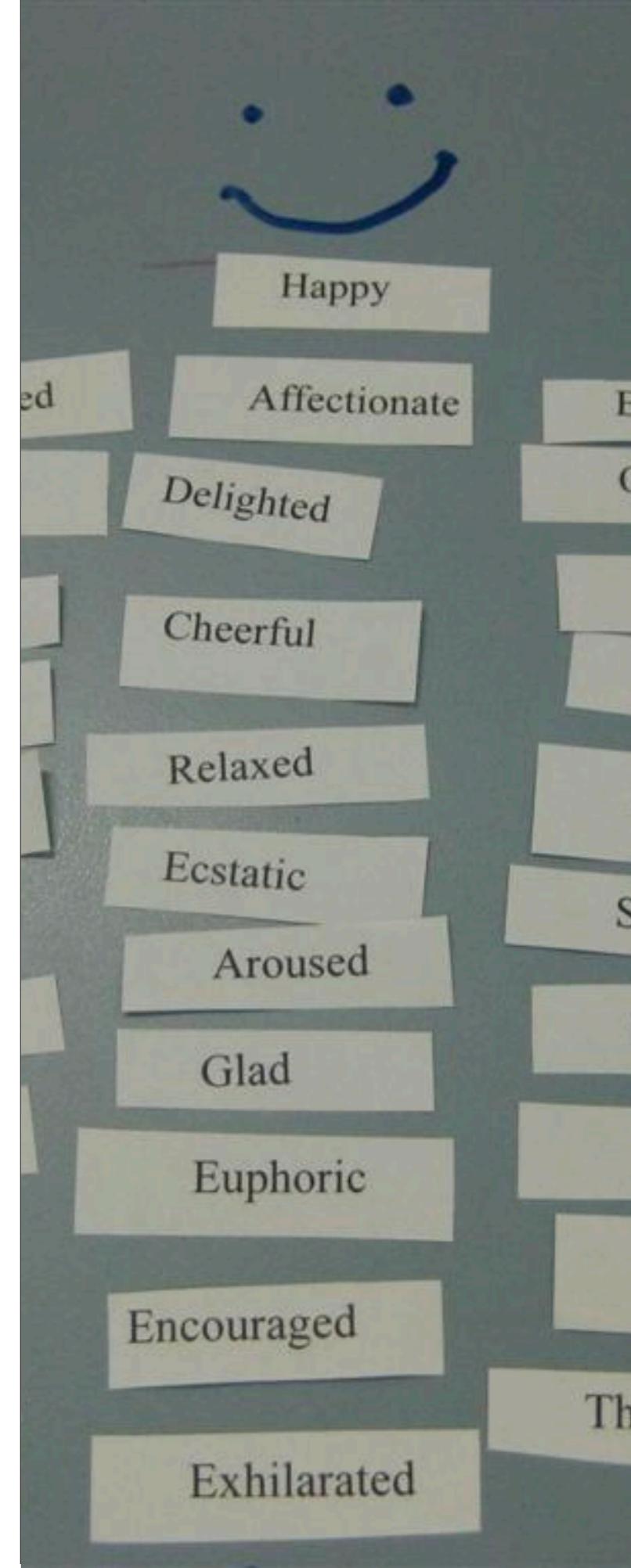
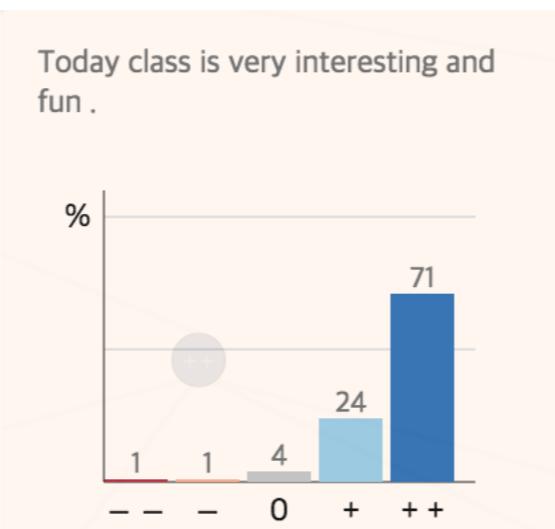
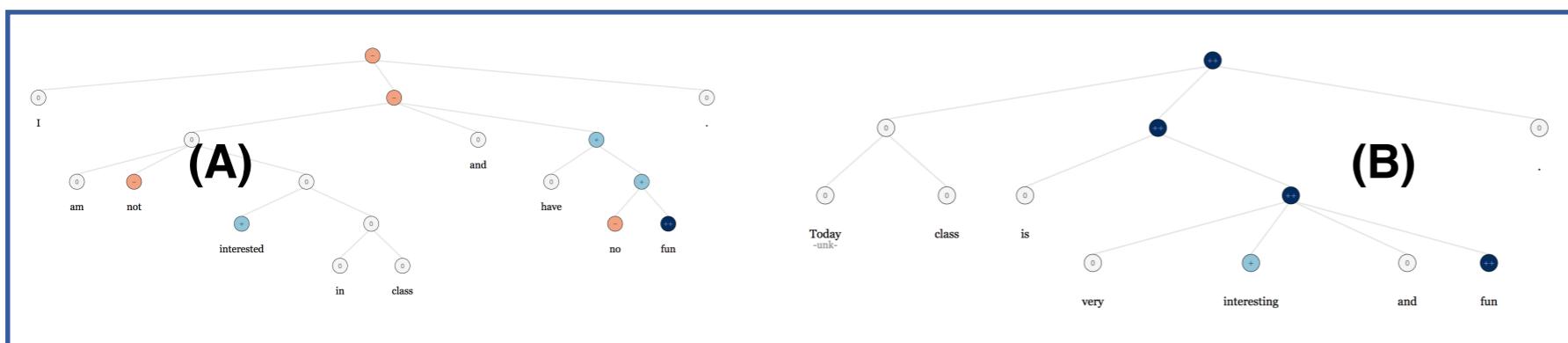
ngram	freq	COMP	NEG	NEUT	None	POS	max.value	max.prop
싸구려/NNG	1	0	1	0	0	0	NEG	1
싸구려/NNG;로/JKB	1	0	1	0	0	0	NEG	1
싸구려/NNG;로/JKB;둔갑/NNG	1	0	1	0	0	0	NEG	1
싸늘/XR	1	0	1	0	0	0	NEG	1
싸늘/XR;하/XSA	1	0	1	0	0	0	NEG	1
싸움/NNG	1	0	1	0	0	0	NEG	1
싸움/NNG;을/JKO	1	0	1	0	0	0	NEG	1
싸움/NNG;을/JKO;일으키/VV	1	0	1	0	0	0	NEG	1
써먹/VV	1	0	1	0	0	0	NEG	1
써먹/VV;지/EC	1	0	1	0	0	0	NEG	1
써먹/VV;지/EC;못하/VX	1	0	1	0	0	0	NEG	1
기대/NNG;되/XSV	2	0	0	0	0	1	POS	1
기대/NNG;를/JKO	2	0	0	0	0	1	POS	1
기대/NNG;하/XSV	2	0	0	0	0	1	POS	1
기량/NNG	2	0	0	0	0	1	POS	1
기뻐하/VV	2	0	0	0	0	1	POS	1
기회/NNG;를/JKO;주/VV	2	0	0	0	0	1	POS	1
길/VA	2	0	0	0	0	1	POS	1
꼭/MAG;필요/NNG	2	0	0	0	0	1	POS	1
꼭/MAG;필요/NNG;하/XSA	2	0	0	0	0	1	POS	1
꼽/VV	2	0	0	0	0	1	POS	1
꼽히/VV	2	0	0	0	0	1	POS	1
꽃/NNG;을/JKO	2	0	0	0	0	1	POS	1
꽃/NNG;을/JKO;피우/VV	2	0	0	0	0	1	POS	1



감성분석 (Sentiment Analysis)

[머신러닝(딥러닝) 기반 감성분석]

A = I am not interested in class and have no fun.
B = Today class is very interesting and fun.



텍스트 데이터 분석

시각화까지가
진정한 데이터분석 과정이다

테이블 (Table)

분석결과를 테이블 형태로 구분하여 표현하는 방법

<표 6> 불행요인 세부 토픽 모델링 결과

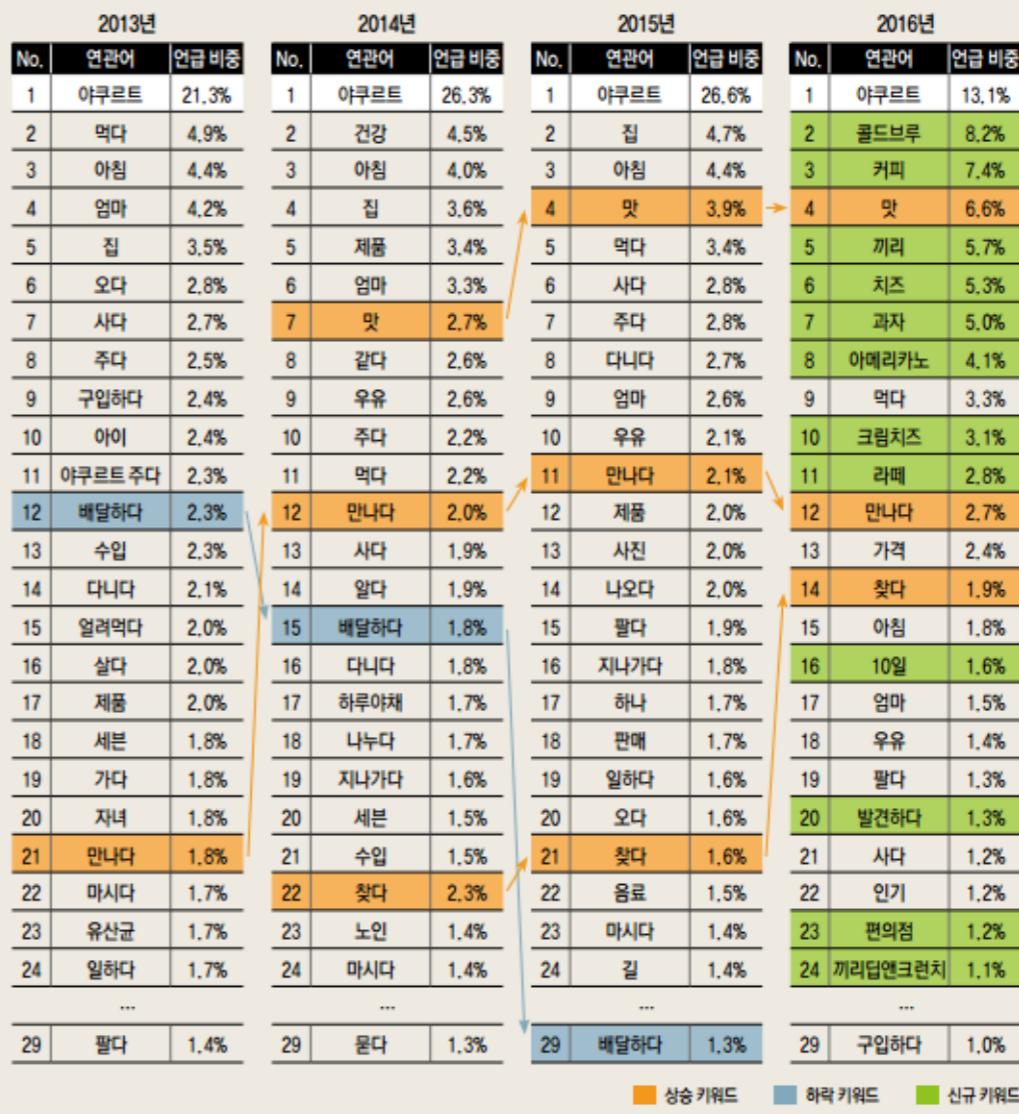
	토 퍽	키 워 드
1	가정 불화	불행, 사랑, 가족, 집, 아버지, 가정, 부모
2	가난	분배, 돈, 소득, 빈곤, 경제, 가난
3	자녀 문제	학교, 위험, 아이, 행동, 상황
4	부정적 인생관	불행, 사람, 인생, 마음, 성공, 공통점
5	인간관계 문제	불행, 자신, 관계, 마음, 생각, 환경, 상황
6	직업 불만족	불행, 사람, 생각, 인생, 직업, 친구
7	건강 문제	불행, 건강, 수명, 질병, 생명, 병, 사고
8	미 취업	오늘, 운세, 불행, 건강, 취업, 뱠띠, 금전
9	부정적 마음가짐	불행, 사람, 마음, 생각, 이기심, 자만심, 피해의식
10	-	예수, 교회, 신앙, 설교, 말씀, 축복

Table 10. Top Seller Characteristics of Rescator

#	Top key words	Interpretation
5	shop, wmz, icq, webmoney, price, dump,	Product: CCs, dumps (valid, verified);
6	валид (valid), чекер (checker), карты (cards), баланс (balance), карт (cards)	Payment: wmz, webmoney, bitcoin, lesspay;
8	shop, good, CCs, bases, update, cards, bitcoin, webmoney, validity, lesspay	Contact: shop, register, deposit, e-mail, icq, jabber
11	dollars, dumps, deposit, payment, sell, online, verified	
16	e-mail, shop, register, icq, account, jabber,	

표1 '아쿠르트 아줌마' 연관어 변화

아쿠르트 아줌마는 여전히 '아쿠르트'와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 아쿠르트 아줌마는 '배달하는' 역할에서 맛난 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화 중.



*Source : 양승준, 이보연, & 김희웅. (2016). 토픽모델링 기반 행복과 불행 이슈 분석 및 행복 증진 방안 연구. 지식경영연구, 17(2), 165-185.

**Source : Li et al., (2016). Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. Journal of Management Information Systems, 33(4), 1059-1086.

***Source : 백경혜(DBR), “매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다”, 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.

워드클라우드 (Wordcloud)

단어의 빈도를 반영해 그 분포를 시각화하는 과정

- ▶ 단어의 크기를 단어의 빈도 수에 비례하도록 아름답게 표현하는 방법
 - ▶ 일반적인 워드클라우드는 빈도 외에 다른 정보를 제공하지 않은나, 단어의 배치에 따라 더 많은 정보를 제공하기도 함



*Source : 몬데이터, [mondata] 남북정상회담 판문점 선언 Text 키워드 분석, 2018.4.28., <https://www.youtube.com/watch?v=ba4EMdzSK-A>.

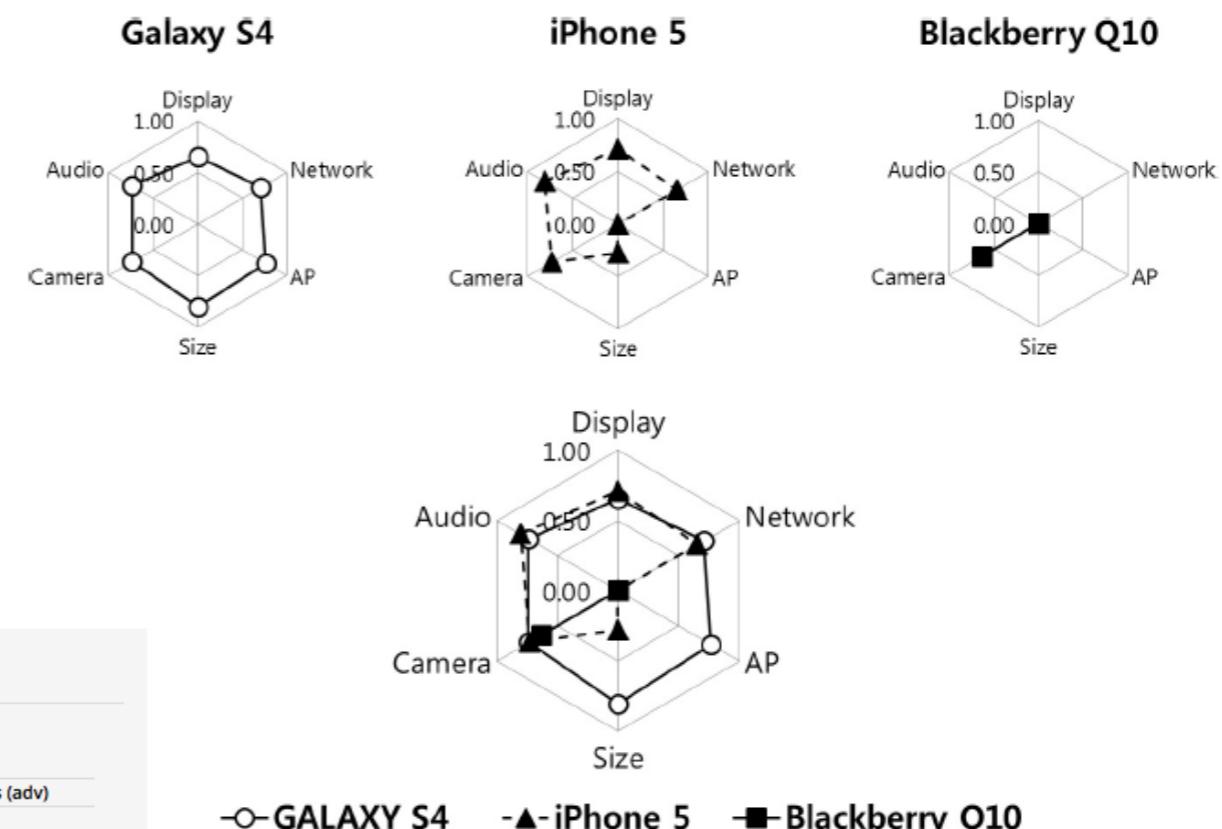
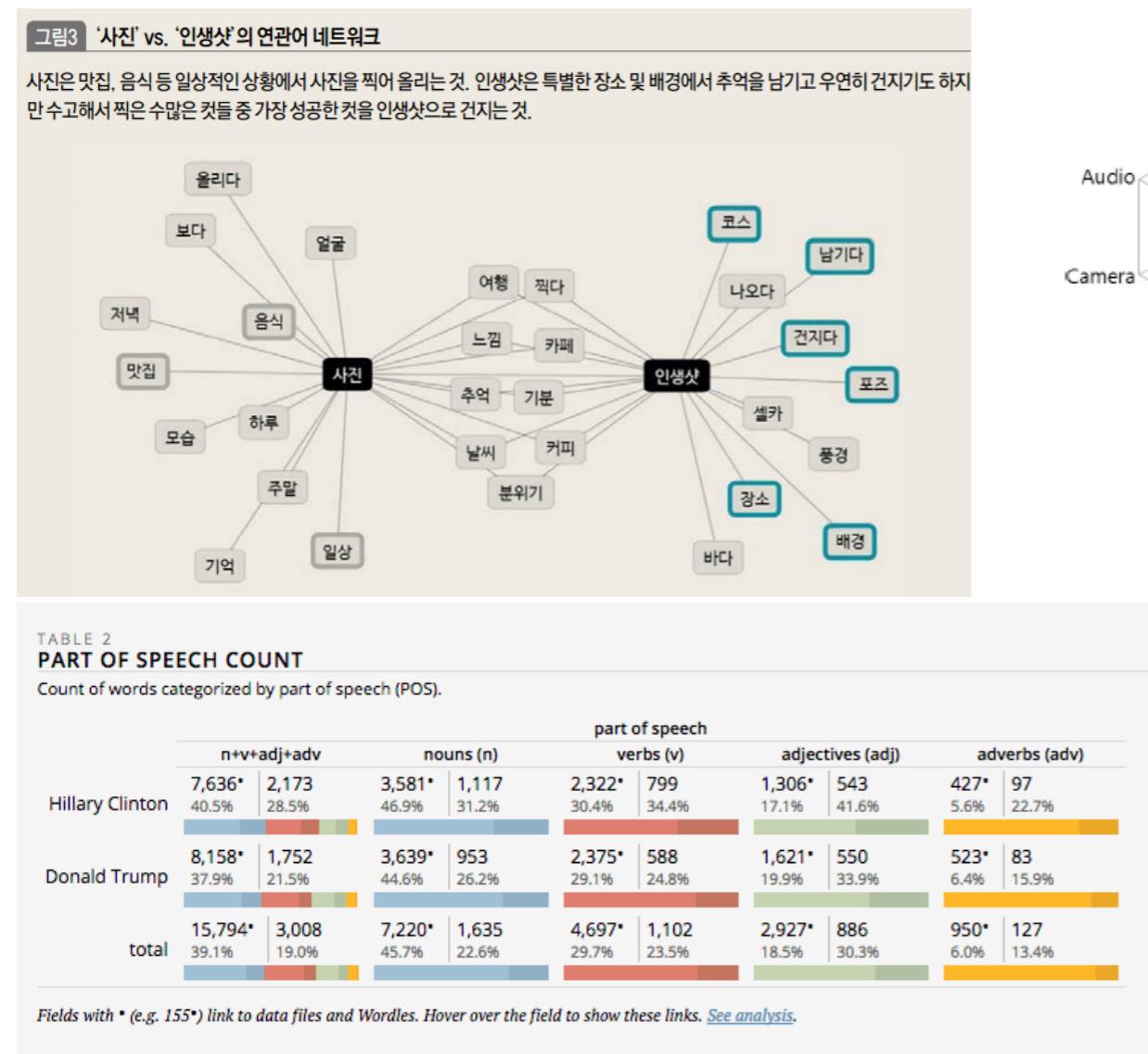
**Source : NÉSTOR CORREA, Cómo implementar el Big Data en tu empresa, 2017., <http://bluelight.tistory.com/298/>

***Source : 전병진, 신한은행 파이썬으로 시작하는 데이터분석: 텍스트 마이닝 기초, 2018.12.12.

그래프 & 네트워크 (Graph & Network)

단어 사이의 관계 강도를 시각화하는 과정

- ▶ 그래프 : 문서 또는 단어의 정량화된 특징을 도표로 표현하는 방법
- ▶ 네트워크 : 단어를 노드, 단어들 사이의 관계를 엣지로 취급하여 네트워크를 표현하는 방법



*Source : 백혜(DBR), “매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다”, 2017.1., http://dbr.donga.com/article/view/1203/article_no/7935/.

**Source : 최홍규(슬로우뉴스), 2016 미국 대선을 보여주는 텍스트 마이닝 분석방법들, 2017.1.9., <http://slownews.kr/60919/>.

***Source : Kim et al. (2014). Analysis on smartphone related twitter reviews by using opinion mining techniques. In Advanced Approaches to Intelligent Information and Database Systems (pp. 205-212).

E.O.D