

TEXT MINING for Beginner

by FINGEREDMAN (fingeredman@gmail.com)

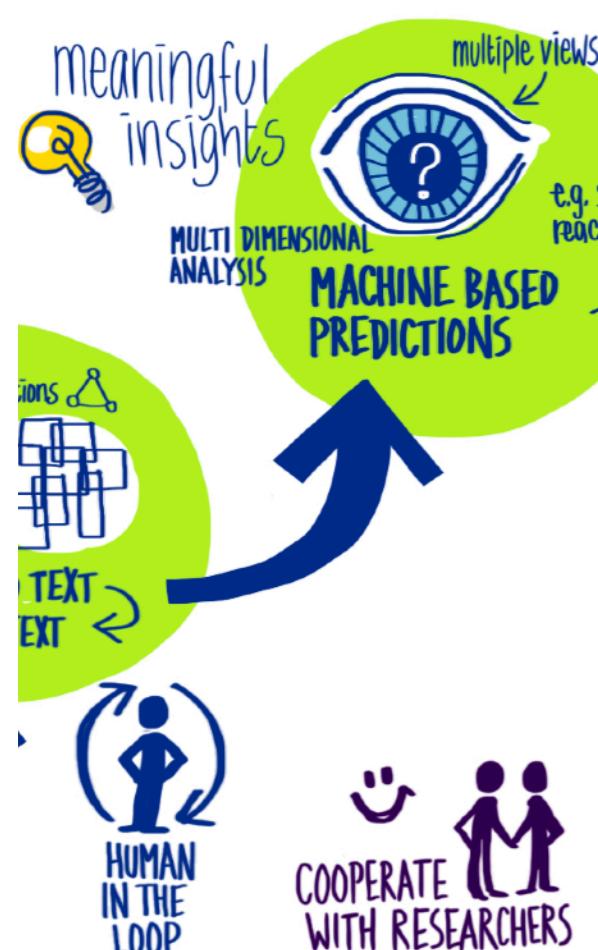
WEEK 03

텍스트 마이닝 프로세스: 수집

HOW?

아무데서나 찾는 텍스트는
아무 의미없는 정보를 포함한다

텍스트 데이터 수집소스



오프라인 데이터

- 수집방법 : 온/오프라인 설문지, 녹음, 촬영
- 정교한 타겟팅을 통해 원하는 대상 데이터 수집이 가능함
- 사람이 직접 또는 전문업체를 통해 비용을 많이 지불하고 수집해야함
→ 데이터 수집에 시간적, 공간적 제약이 큼

온라인 데이터

- 수집방법 : 웹페이지에 존재하는 모든 데이터
- 대량의 정보를 반복적으로 빠르게 수집할 수 있음
- 수집 대상 웹페이지의 API 호출이나 프로그래밍 언어 활용이 필요함
- 개인정보와 저작권 문제에 취약함

시스템 데이터

- 수집방법 : 게시판, 댓글, 보고서 등 시스템 상에 존재하는 모든 저장된 문서
- 기 수집된 데이터를 바로 활용하여 데이터 수집에 시간적 제약이 없음
- 시스템 관련 부서가 아니거나 관련부서 승인 없이 접근이 어려움
- 내부 데이터를 활용하므로 정보 유출 및 보안관리에 대한 위험성이 큼



비정형 데이터 수집소스

유형별 텍스트 데이터 수집 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

비정형 데이터 수집소스

유형별 텍스트 데이터 수집 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
외방커뮤니티	커뮤니티	자체플랫폼	미용, 패션		20~30	여성

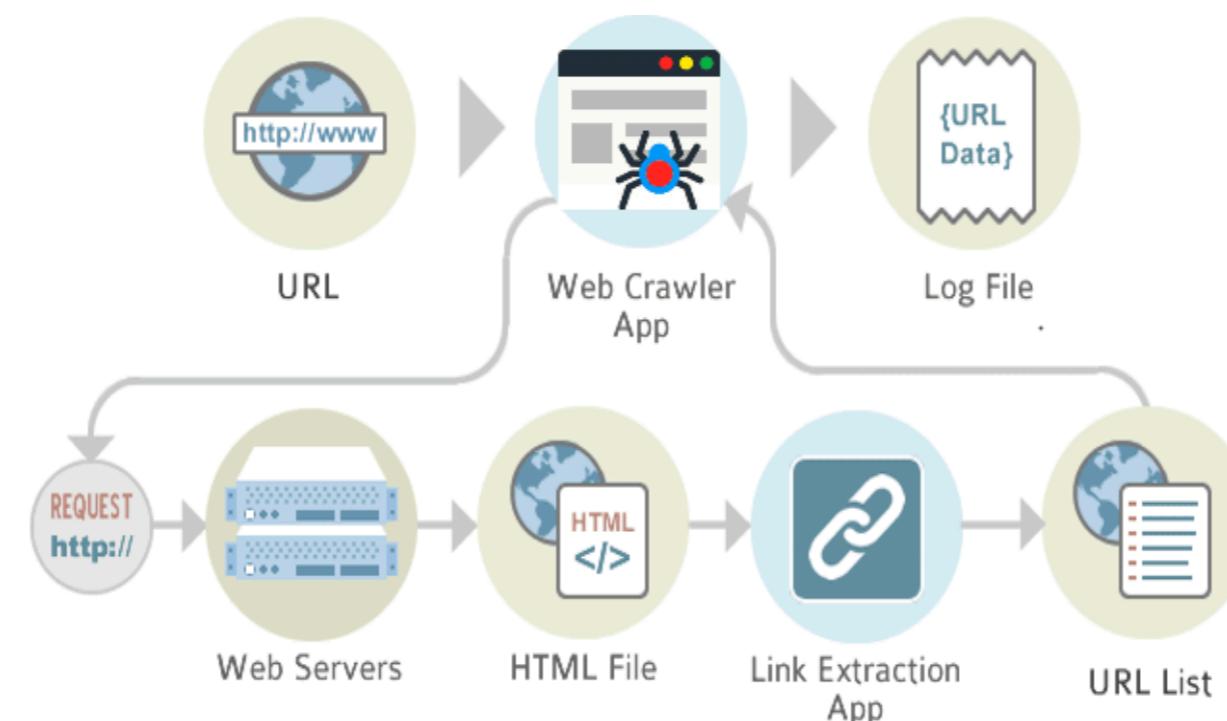
웹 크롤링 & 스크래핑

웹 크롤링(Web Crawling)의 사전적 의미

- 정제되지 않은 웹페이지에서 필요한 데이터를 추출/수집하는 행위
- 즉시 활용 가능한 데이터(파일, 데이터베이스 등)를 제외하고, 웹페이지에 게시된 데이터를 수집하는 기술
- API(Application Programming Interface)를 통해 데이터를 제공하는 경우가 있지만 제약이 많아 웹크롤링을 통한 수집이 요구됨

웹 크롤링 vs 웹 스크래핑, 원론적 의미의 차이점

- **웹 크롤링 (crawling)** : 웹페이지의 하이퍼링크를 돌아다니며 페이지 정보를 추출하고 저장하는 방법
- **웹 스크래핑 (scrapping)** : 웹페이지에서 필요한 정보만 추출하여 저장하는 방법
- 두 가지 모두 웹페이지 정보를 저장하는 방법으로 데이터를 수집하고 활용하기 위한 사전작업에 해당함



웹 크롤링 & 스크래핑

데이터 무단수집과 저작권 침해

- 웹 스크래핑은 원래 검색엔진 등의 인터넷 사이트에서 데이터를 최신 상태로 유지하기 위해 사용하는 기법임
- 웹 스크래핑을 활용하여 타사 컨텐츠를 무단 활용하는 것은 불법행위에 해당하며, 과도한 데이터 수집은 대상 웹서비스 운영과 서비스 관리에 안 좋은 영향을 끼침
- 무분별한 웹 스크래핑은 경쟁사 간의 상도덕 문제 또는 개인 양심상의 문제로 확장될 수 있음

채용정보 무단복제 '사람인HR', 잡코리아에 120억 지급

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

이준영 기자 | 승인 2018.02.09 12:31 | 댓글 0

JOBKOREA

saramin

사진=각사

제 및 게재 행위를 하지 않고 공정한 경쟁질서의 확립에 힘쓸 것"이라고 밝혔다.

채용공고 불법 복제 및 게재하는 웹크롤링 행위를 두고 10여 년간 갈등을 빚어온 사람인과 잡코리아가 마침내 합의를 이뤘다.

사람인은 웹크롤링 소송 합의금으로 잡코리아에 120억을 지불했다. 사람인은 이 같은 내용을 공시하고 10일 동안 사람인의 인터넷 웹사이트에 사과문을 공고함으로써 "향후 잡코리아 채용정보 복

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



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.

웹 크롤링 & 스크래핑

Robots.txt

- 웹 사이트에 배치된 텍스트 파일로, 웹 스크래핑 접근권한에 대해 명시해 놓은 문서
- 원칙적으로 웹 스크래핑은 Robots.txt 파일에서 허용하는 범위 내에서만 가능하며 그 외의 수집에 대한 책임은 모두 본인에게 있음
- 웹 스크래핑이 허용되더라도 대상 웹 사이트 운영에 피해를 주지 않는 선에서 필요한 만큼만 수집해야함
- Robots.txt 파일이 없는 경우 서비스 관리자에 직접 허락을 구한 후 수집해야함

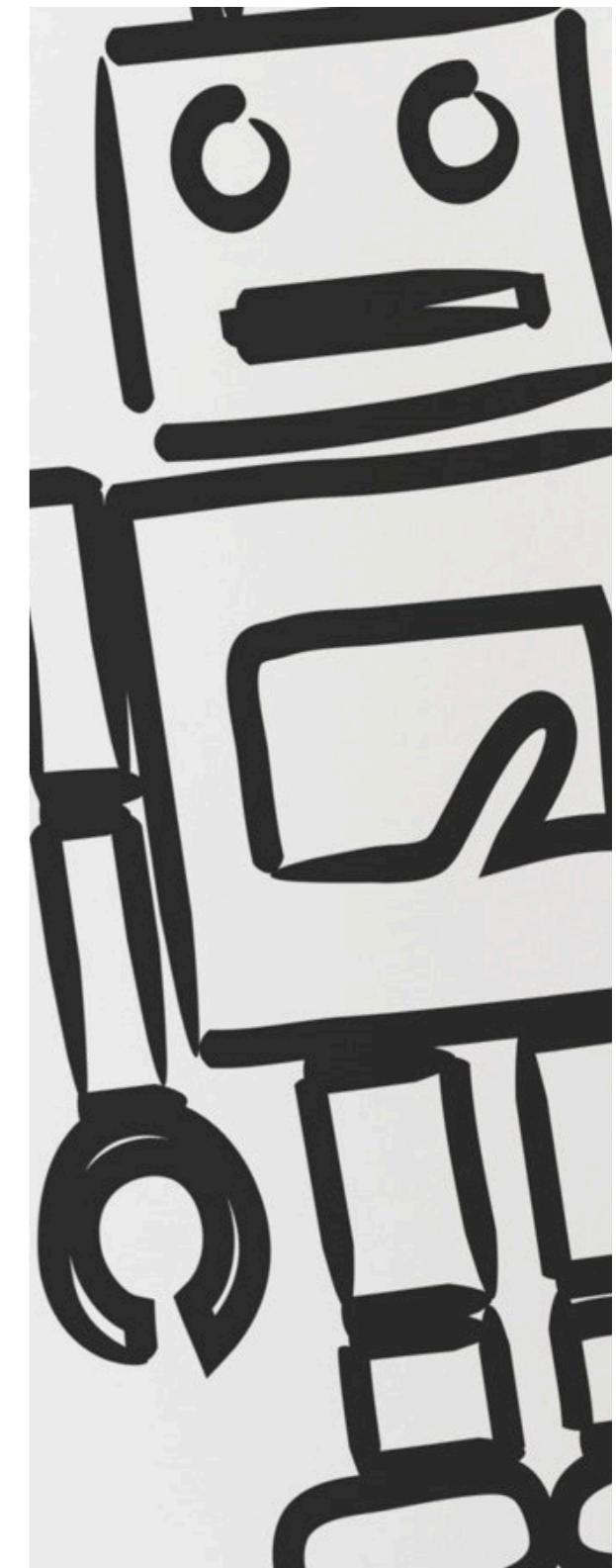


```
User-agent: Bingbot
Allow: /ajax/pagelet/generic.php/PagePostsSectionPagelet
Allow: /safetycheck/

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/
```



웹 크롤링 & 스크래핑

Python을 활용하는 이유

- 가장 널리 사용되는 프로그래밍 언어로 배우기 쉽고 따라하기 쉬운 장점이 있음
- 풍부한 외부 라이브러리로 프로그래밍이 쉽고 확장성이 좋음
- 데이터 분석과 머신러닝/딥러닝에 가장 많이 사용되는 프로그래밍 언어로, 데이터 수집 뿐만 아니라 수집된 데이터에 대한 전처리, 분석으로 연계하기에 유리함

웹 스크래핑 방법 선택

라이브러리	장점	단점
Requests (urllib)	Python에서 동작하는 가장의 작고 빠른 브라우저로, 거의 모든 플랫폼에서 구동 가능 Selenium에 비해 속도가 수백배 빠름	웹서버로부터 초기 HTML만 받아 동적인 처리결과가 반영되지 않음
Selenium	웹브라우저를 Python을 활용해 원격으로 컨트롤 하는 라이브러리 (Chrome, Firefox, IE, PhantomJS 등)	직접 브라우저를 사용하여 동적인 처리를 위한 리소스가 많이 필요함



E.O.D

Contact

-  <http://www.teanaps.com>
-  fingeredman@gmail.com

WEEK 03

텍스트 마이닝 프로세스: 전처리

HOW?

텍스트 전처리를 하루루 하면
하루분석결과만 나온다

텍스트 마이닝 용어

문서 (Document)

- 한 덩어리의 텍스트로서 문장에서 긴 문서까지 모두 포함하며, 문서의 집합을 말뭉치 (corpus)라고 함
- 문서의 레벨에 따라 말뭉치의 레벨이 바뀔 수 있음 (문장, 문단, 페이지, 댓글 등)

어휘사전 (Lexicon)

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

형태소 (Morpheme)

- 뜻을 가진 가장 작을 말의 단위로서 동사, 명사, 조사, 문장부호 등의 품사 (part of speech) 단위를 의미함
- **형태소분석 (morphological analysis)** : 문장을 형태소 단위로 구분하고 품사를 태깅하는 과정

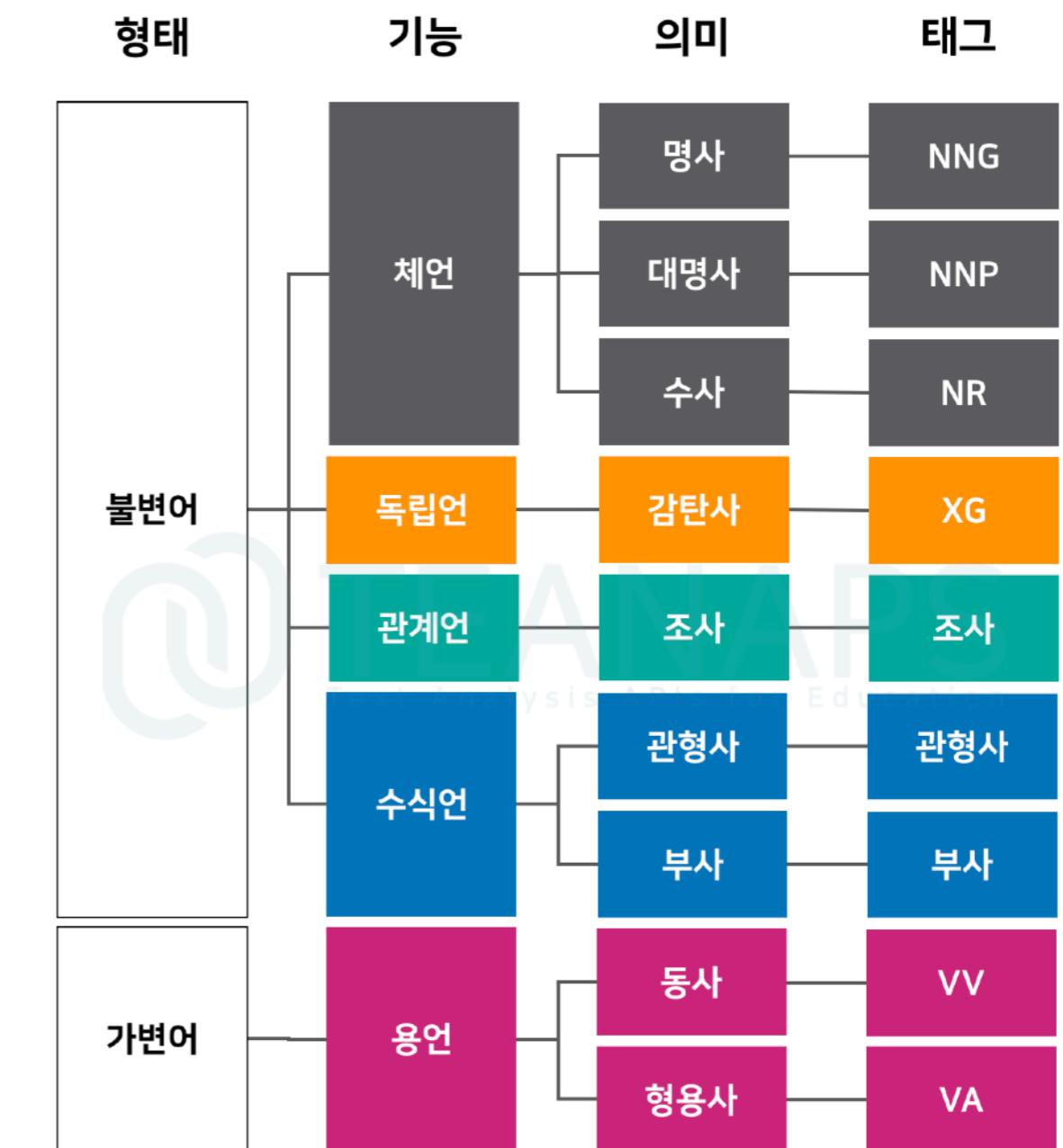
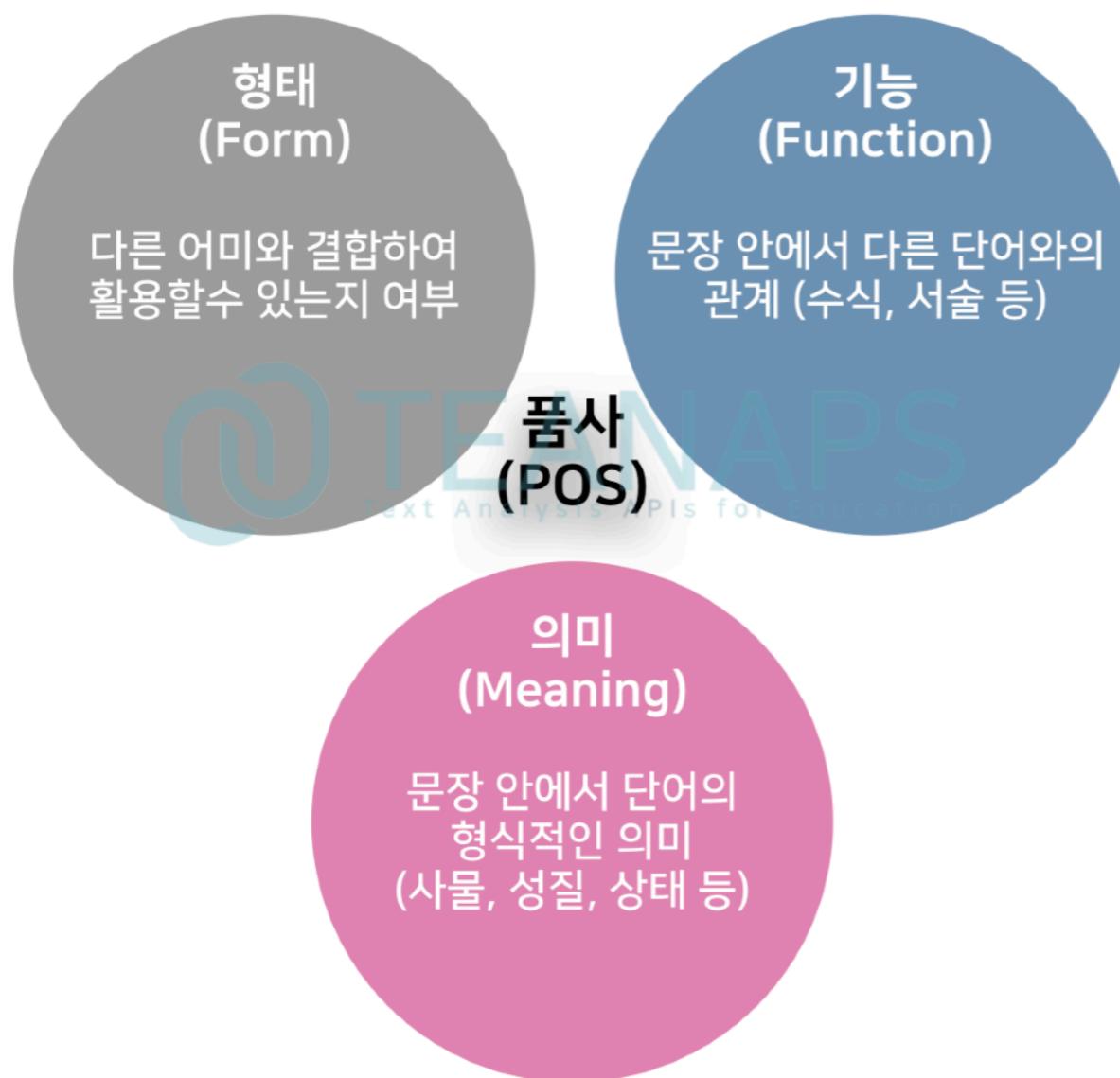
토큰 (Token)

- 유용한 의미적 단위로 함께 모여지는 일련의 문자열
- 구분 기호 사이의 문자열 시퀀스

한국어의 언어학적 특징

한국어 5언 9품사

- 한국어는 단어를 기능(function), 의미(meaning), 형태(form)의 세 가지 기준에 의해 분류함



한국어의 언어학적 특징

교착어, 굴절어, 그리고 고립어

- **교착어** (*agglutinative language*) : 어근에 접사가 결합되어 각 단어의 기능을 나타내는 언어
- **굴절어** (*inflectional language*) : 단어 자체의 형태변화로 그 단어의 문법성을 나타내는 언어
- **고립어** (*isolating language*) : 단어의 형태변화 없이 문법적 관계는 어순에 의해 정해지는 언어

어근	피동	높힘	과거	추측	전달	어미	파생된 단어
일어나						+다	일어나다
일어나	+지					+다	일어나지다
일어나	+지	+시				+다	일어나지시다
일어나	+지	+시	+었			+다	일어나지셨다
일어나				+았		+다	일어났다
일어나					+겠	+다	일어나겠다
일어나					+더라		일어나더라
일어나		+지	+었			+다	일어나졌다
일어나		+지	+었	+겠		+다	일어나졌겠다
일어나	+지	+었	+겠		+더라		일어나졌겠더라
일어나			+았	+겠		+다	일어났겠다
일어나	+지	+시	+았	+겠	+더라		일어나지셨겠더라

한국어의 언어학적 특징

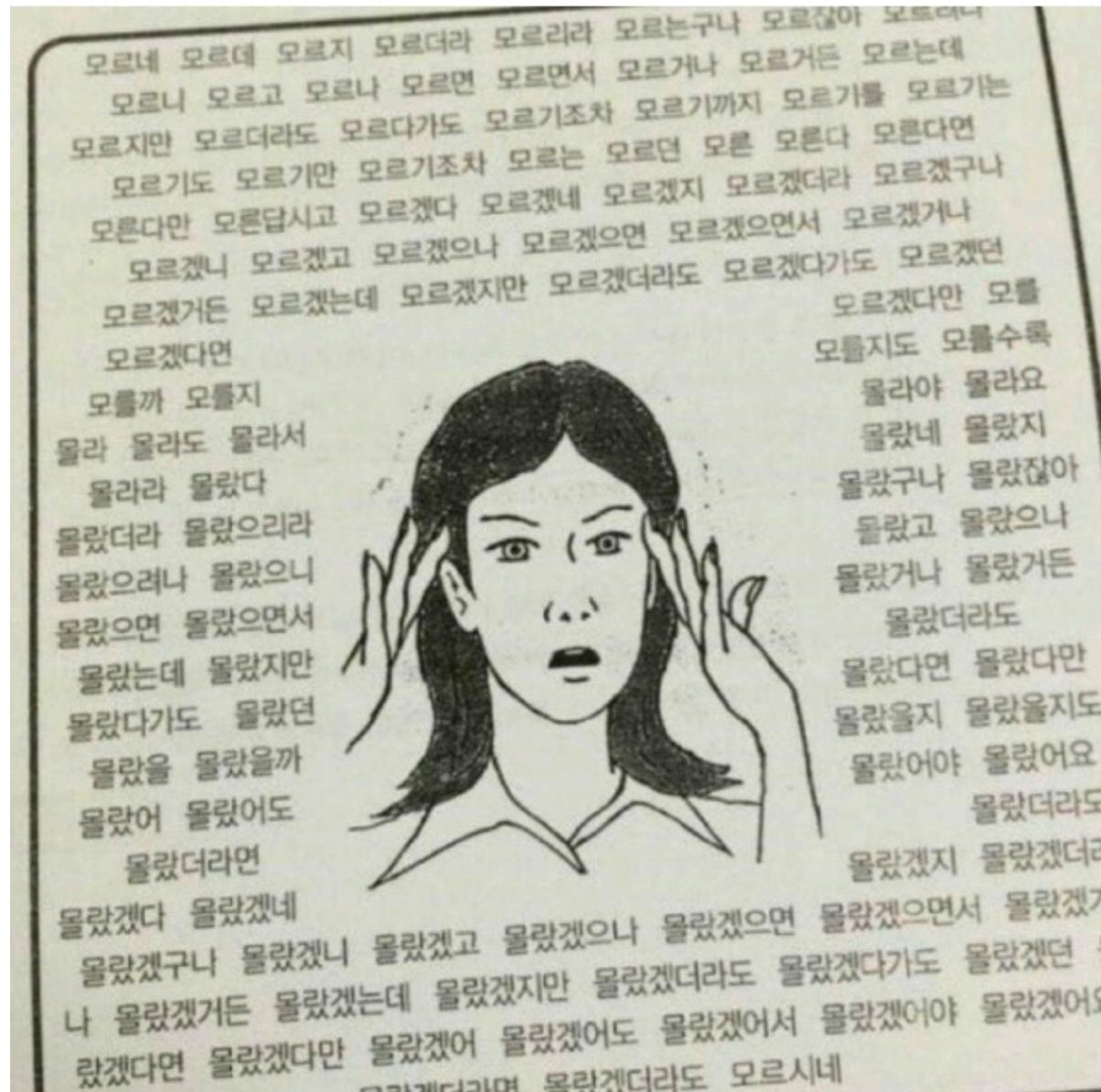
교착어 텍스트 분석이 어려운 이유

- 동일한 단어임에도 불구하고 어근에 따라 단어가 다양한 형태로 파생되어 표현됨
- 단어가 다양하게 생겨나므로 하나의 어근에서 비롯된 비슷한 의미의 단어가 매우 많이 생성됨
- 형태소 분석 시 수많은 경우의 수를 다르게 처리하거나 추가적인 토큰화가 필요함
- 어근에 따라 단어의 역할이 정의되기 때문에, 어순이 전체적인 문장의 의미에 미치는 영향이 상대적으로 매우 적음

구분	한국어	영어
원문	나는 오전수업을 들으러 학교에 간다.	I morning classes to take to school go.
문장 1	간다 나는 오전수업을 들으러 학교에.	Go I to morning classes take to school.
문장 2	학교에 간다 나는 오전수업을 들으러.	To school go I morning classes to take.
문장 3	오전수업을 들으러 학교에 간다 나는.	Morning classes to take to school go I.
문장 4	나는 간다 학교에 들으러 오전수업을.	I go to school to take morning classes.
문장 5	간다 학교에 오전수업을 들으러 나는.	Go to school morning classes to take I.
문장 6	학교에 들으러 나는 간다 오전수업을.	To school to take I go morning classes.
문장 7	나는 간다 들으러 학교에 오전수업을.	I go to take to school morning classes.
문장 8	들으러 학교에 나는 오전수업을 간다.	To take school I morning classes go to.

텍스트 마이닝이 어려운 이유

Review



I just got here.

상기 문장은 영어로 "나 막 도착했어" 가 된다. 자연스럽게 위 문장을 바꿀 수 있는

I have just arrived 하나 정도다.

한국어에서 저 just라는 표현은 대체 수십 가지로 가능하다.

나 막 왔어.

나 방금 왔어.

나 지금 왔어.

나 금방 왔어.

나 온 지 조금/좀 됐어. (조금에 강세)

나 온 지 별로/얼마 안 됐어.

나 이제 왔어.

나 바로 막 왔어.

게다가 위의 모든 표현의 '왔어'를 "도착했어"로 바꿔도 말이 된다.

- 시발ㅋ, 시발ㅋㅋ : 웃김
- 오 시발 : 놀라움
- 마 시발 : 마쉬움
- 시발... : 슬픔
- 시발! : 분노
- 시발; : 어이없음
- 시발ㅠㅠ : 격한슬픔
- 시발;; : 당황스러움
- 시바ㄹ : 급함
- 시ㅂ : 더욱 급함
- tlqkf : 정말로 급함

한국어 형태소 분석

토큰화 (Tokenization)

- 문서 또는 문자열이 주어졌을 때 이를 의미있는 최소단위의 토큰(형태소) 단위로 분리하는 작업
- 경우에 따라 구두점 등 불필요하거나 의미를 담고 있지 않은 토큰은 제외하기도 함
- 영어는 언어학적 특성상 단어에 조사가 붙지 않아 토큰화가 비교적 쉬우나 (어절단위 분리과정과 동일함), 한국어는 언어학적 특성으로 다른 언어에 비해 토큰화가 매우 어려움

구분	내용													
원문	손 흥 민 이 골 을 작 렬 하 며 토 트 넘 홋 스 퍼 의 승 리 를 이 끌 었 다 .													
토큰화	손흥민	이	골	을	작렬	하	며	토트넘 홋스퍼 (?)	의	승리	를	이끌	었	다 .

한국어 형태소 분석

품사태깅 (Part of Speech Tagging)

- 토큰화된 문장의 각 토큰에 대해 앞뒤 문맥상 적합한 품사를 태깅하는 작업
- 토큰화가 잘못되거나, 동일한 토큰임에도 문맥에 따라 다양한 의미를 가지는 토큰이 존재하는 경우 품사태깅 결과가 잘못될 가능성이 큼
- 도메인에 특화 문서는 별도의 형태소 사전을 통해 토큰의 유형과 형태소 태그를 관리하여 품사태깅에 반영해야함



한국어 형태소 분석

형태소 품사 태그표

대분류	세종 품사 태그		KKMA 단일 태그 V 1.0					
	태그	설명	묶음 1	묶음 2	태그	설명	활태그	저장사전
체언	NNG	일반 명사	N	NN	NNG	보통 명사	NN	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	수사		
	NP	대명사			NP	대명사		
용언	VV	동사	V	V	VV	동사	EF	verb.dic
	VA	형용사			VA	형용사		
	VX	보조 용언			VX	VXV 보조 동사		
					VXA	보조 형용사		
	VCP	긍정 지정사		VC	VCP	긍정 지정사, 서술격 조사 '이다'		raw.dic
	VCN	부정 지정사			VCN	부정 지정사, 형용사 '아니다'		
관형사	MM	관형사	M	MD	MDT	일반 관형사	E	
					MDN	수 관형사		
부사	MAG	일반 부사		MA	MAG	일반 부사		
	MAJ	접속 부사			MAC	접속 부사		
감탄사	IC	감탄사	I	IC	IC	감탄사	IC	
조사	JKS	주격 조사	J	JK	JKS	주격 조사	EF	
	JKC	보격 조사			JKC	보격 조사		
	JKG	관형격 조사			JKG	관형격 조사		
	JKO	목적격 조사			JKO	목적격 조사		
	JKB	부사격 조사			JKM	부사격 조사		
	JKV	호격 조사			JKI	호격 조사		
	JKQ	인용격 조사		JC	JKQ	인용격 조사		
	JX	보조사			JX	보조사		
	JC	접속 조사			JC	접속 조사		
	EP	선어말 어미	EP		EPH	존칭 선어말 어미	EP	raw.dic
선어말 어미					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	활태그	저장사전	
어말 어미	EF	종결 어미	E	EF	EFN	평서형 종결 어미	EF		
	EFQ	의문형 종결 어미			EFO	명령형 종결 어미			
	EFA	청유형 종결 어미			EFI	감탄형 종결 어미			
	EFR	존칭형 종결 어미			ETN	명사형 전성 어미	ET		
	ECE	대등 연결 어미			ECD	의존적 연결 어미			
	ECS	보조적 연결 어미			ETM	관형형 전성 어미			
접두사	XPN	체언 접두사	XP	XPN	XPN	체언 접두사	XPN	simple.dic	
					XPV	용언 접두사			
접미사	XSN	명사 파생 접미사	XS	XSN	XSN	명사 파생 접미사	XSN		
	XSV	동사 파생 접미사			XSV	동사 파생 접미사			
	XSA	형용사 파생 접미사			XSA	형용사 파생 접미사			
	XSM	부사파생접미사			XSM	부사파생접미사			
	XSO	기타접미사			XSO	기타접미사			
어근	XR	어근	S	SF	XR	어근	SF	Symbol class	
부호	SF	마침표물음표, 느낌표			SP	쉼표, 가운뎃점, 콜론, 빗금			
	SS	따옴표, 괄호표, 줄표			SE	줄임표			
	SO	불임표(물결, 숨김, 빠짐)			SO	불임표(물결, 숨김, 빠짐)			
	SW	기타기호 (논리수학기호, 화폐기호)			SW	기타기호 (논리수학기호, 화폐기호)			
	NF	명사추정범주	U	UN	UN	명사추정범주	NNA	N/A	
	NV	용언추정범주			UV	용언추정범주			
	NA	분석불능범주			UE	분석불능범주			
한글 이외	SL	외국어	O	OL	OL	외국어	NNA		
	SH	한자			OH	한자			
	SN	숫자			ON	숫자			

한국어 형태소 분석

형태소 분석 (Morphological Analysis)

- 문장을 형태소 단위로 구분하고 품사를 구별하여 태깅하고 용언의 변형으로 탈락한 형태소를 복원하는 과정
- 형태소 분석기마다 형태소를 구분하는 방식이 다르기 때문에 특성에 맞는 형태소 분석기를 선택해야함
- 형태소 분석의 활용 범위
 - 1) 언어학적 측면 : 특정 언어현상의 생성과정을 설명하는 데 용이하게 쓰일 수 있음
 - 2) 전산학적 측면 : 정보검색이나 자연어 처리 자동 처리시스템의 구문 분석의 전 단계 등의 용도로 쓰일 수 있음



오픈소스 형태소 분석기 특징

한나눔 형태소 분석기 (Hannanum)

- 한국과학기술원(KAIST)의 SWRC(Semantic Web Research Center)에서 개발한 형태소 분석기
- 자동 띄어쓰기 모듈을 제공해 형태소 분석 결과를 활용하여 한글 문장에 대한 자동 띄어쓰기 수행 가능
- 사전 기반의 맞춤법 교정 모듈로 형태소 분석 결과를 활용하여 한글 단어에 대한 맞춤법 교정 수행 가능



오픈소스 형태소 분석기 특징

Okt (Twitter)

- 트위터에서 개발한 한국어 형태소 분석기
- SNS에서 발생하는 언어에서 자주 발생하는 인물명, 신조어 등을 잘 인식하는 장점이 있음
- 타 형태소 분석기 대비 속도가 빠른 편이지만 형태소 태그 구분이 명확하지 않고 품질이 상대적으로 낮음



오픈소스 형태소 분석기 특징

꼬꼬마 형태소 분석기 (KKMA)

- 서울대 IDS(Intelligent Data Systems) 연구실에서 자연어 처리를 위한 모듈구축과제로 개발한 형태소 분석기
- Java 언어를 기반으로 하며, Python-Java 연동을 통해 Python에서 사용 가능하도록 배포됨
- 가능한 모든 형태소 후보와 조합을 모두 찾아 그 중 가장 적합한 형태소를 판단함 → 매우느림



오픈소스 형태소 분석기 특징

Mecab (은전, 은전한닢)

- 검색에서 쓸만한 오픈소스 한국어 형태소 분석기를 목적으로 개발된 한국어 형태소 분석기
- 오픈소스 검색엔진 Elasticsearch에 형태소 분석 모듈로 적용되어 활용되고 있음
- 사용자 사전 등록기능을 제공하여 다양한 도메인에서 생성되는 단어들을 인식할 수 있도록 도와줌



오픈소스 형태소 분석기 특징

카카오 Khaiii

- 카카오에서 DHA2(Daumkakao Hangul Analyzer 2)를 계승하여 개발하고 2018년 공개한 형태소분석기
- 속도를 매우 중요시하며, 신경망 알고리즘 중에서 CNN(Convolutional Neural Network)을 사용하여 개발됨
- 사용자 사전 등록기능을 제공하여 다양한 도메인에서 생성되는 단어들을 인식할 수 있도록 도와줌



형태소 분석기 성능비교

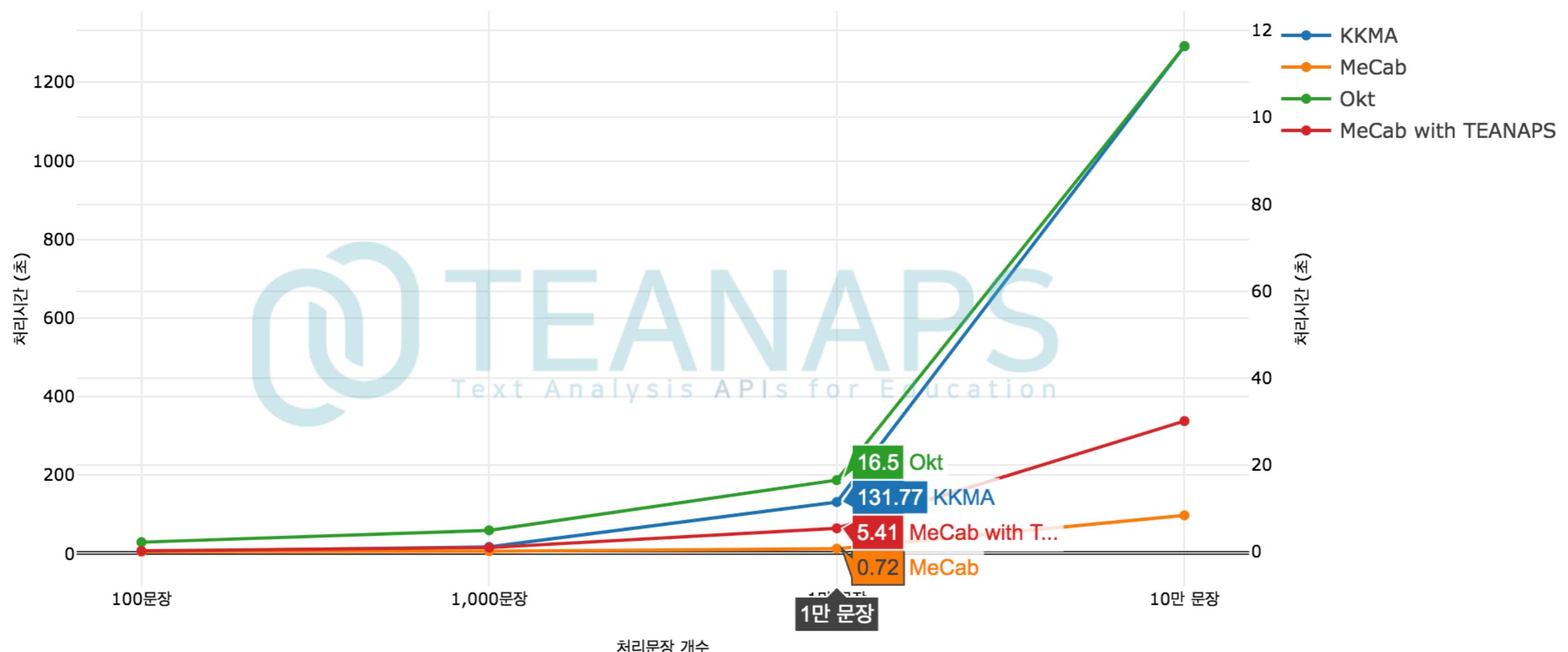
형태소 분석기 별 결과 비교

구분	내용																							
원문	손	흥	민	이	골	을	작	렬	하	며	토	트	넘	홋	스	퍼	의	승	리	를	이	끌	었	다
한나눔	손흥민		이		골	ㄹ	작렬하	이	며		토트넘			홋스퍼		의		승리	를	이끌	었다		.	
Okt	손흥민	이			골	을	작렬		하며		토트넘			홋스퍼		의		승리	를	이끌었다			.	
꼬꼬마	손	흥	민	이	골	을	작렬	하	며		토트	넘		홋스퍼		의		승리	를	이끌	었	다	.	
Mecab	손흥민	이			골	을	작렬	하	며		토트넘			홋스퍼의				승리	를	이끌	었	다	.	
Khaiii	손흥민	이			골	을	작렬	하	며		토트넘			홋스퍼		의		승리	를	이끌	었	다	.	

형태소 분석기 성능비교

형태소 분석기 별 수행시간 비교 (Time Analysis)

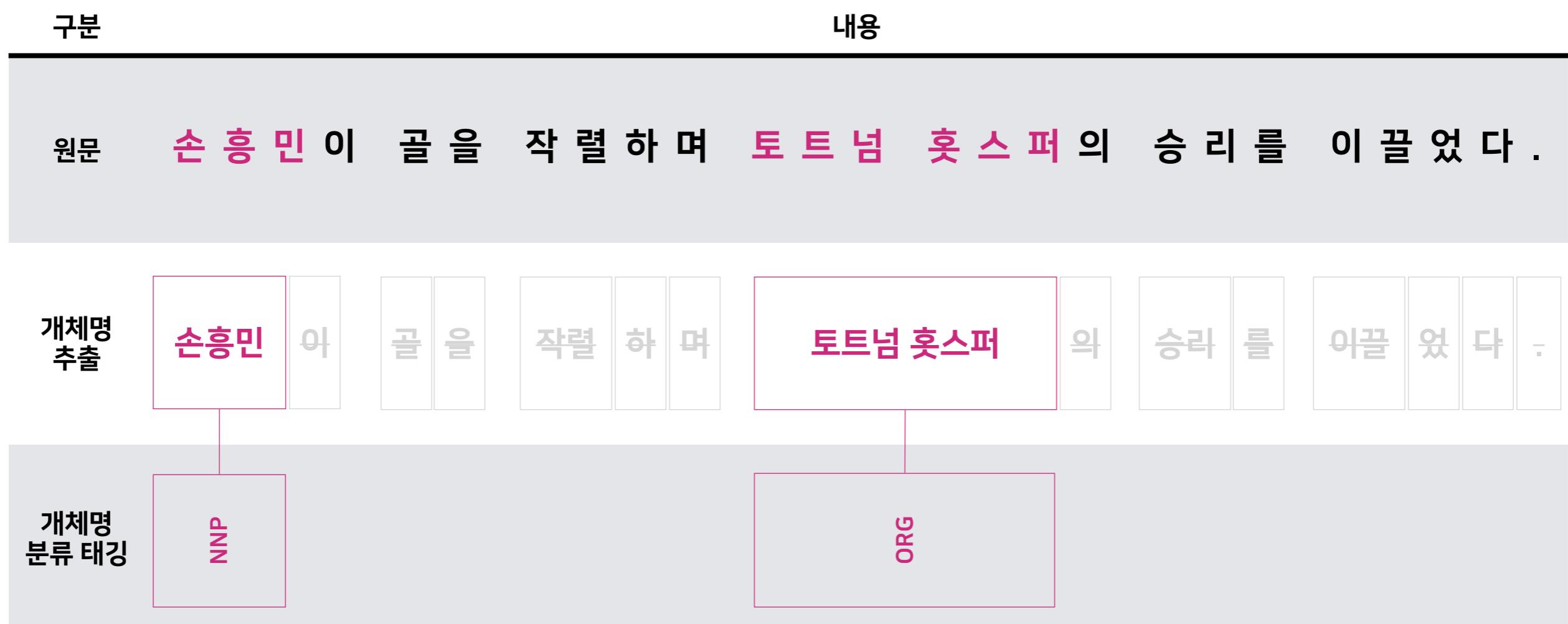
형태소 분리 및 품사태깅 평가결과

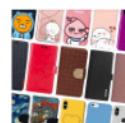


개체명 인식

개체명 인식 (Named Entity Recognition, PLO Tagging > PLODT Tagging > NER)

- 문서에서 하나의 개체로써 인식되어야하는 단어를 추출하고 분류를 태깅하는 과정 (지명, 인물명, 회사명 등)
- 개체명 태그 유형은 응용영역에 따라서 서로 다르게 지정할 수 있음 (개체명 사전 또는 개체명 태깅 학습셋 활용)
- 정보추출(Information Extraction), 정보검색(Information Retrieval) 분야에서 대부분의 지식(Who, Where, When, Which)은 개체명으로 표현될 수 있음





갤럭시A31 갤럭시A51퀀텀 갤럭시노트10 와이드3 LG G8 G7 G6 벨벳 케이스

4,900원 39,900원

N 구매하기

상세정보

리뷰 1,187

Q&A 85

반품/교환정보

리뷰 9건

랭킹순 | 최신순 | 평점 높은순 | 평점 낮은순

전체

포토/동영상

스토어PICK

한달사용리뷰

주제전체

색상

재질

가죽

만족도

수납

가격

디자인



★★★★★ 4

jjan**** · 20.08.12. · 옵션 : 옵션선택: 08.옴팡이 코지 클리어 젤리 / 색상: 러브스레드 | 신고

이미지와 실제 케이스가 똑같고 재질도 부드럽고 만족스럽습니다~



0



★★★★★ 4

jjan**** · 20.08.12. · 옵션 : 옵션선택: 18.귀염뽀짝 시즌1 사피아노 다이어리 / 색상: 푸들푸들 | 신고

이미지랑 똑같이 귀엽고 딱 맞아요. 케이스

내부 재질도 부드럽고 아주 좋습니다. 저렴한 가격은 아닌것 같지만 선물용으로 만족합니다.



0



★★★★★ 4

revi**** · 20.08.03. · 옵션 : 옵션선택: 12.버핏 지퍼 다이어리 / 색상: 퍼플 | 신고

소재도 나쁘지않고 튼튼해보여서 좋아요



0



판매자 20.08.04. | 신고

소중한 리뷰 감사합니다♡



갤럭시A31 갤럭시A51퀀텀 갤럭시노트10 와이드3 LG G8 G7 G6 벨벳 케이스

4,900원 39,900원

N 구매하기

상세정보

리뷰 1,187

Q&A 85

반품/교환정보



★★★★★ 1

cbr0**** · 20.06.30. · 옵션 : 옵션선택: 10. 블루문 다이어리 / 색상: 와인 | 신고

0

새상품인데 상품에 끈끈이가 다 붙어있어 가지고 진짜. 기분나빴어요 물로지워지지도 않는 끈끈이가 왜 새 제품에 묻어잇는지 아직도 의문. 이상한상 품 제 돈주고 산 기분이네요 다른분들 잘 보고 쓰세요.

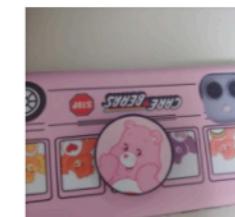


★★★★★ 1

expa**** · 20.03.17. · 옵션 : 디자인: G.케어베어 버스 (카드수납) / 컬러: 핑크 / 스마트톡: 핑크 | 신고

1

한달사용기 하드케이스는 벗겨지지도 안고 다시말하지만 별로네요..... 기대가큰건지 정말 생각보다별로네요



★★★★★ 1

expa**** · 20.03.17. · 옵션 : 디자인: A.카이카이키키 / 컬러: 화이트 / 스마트톡: 화이트 | 신고

0

한달사용기 샀는지 한달도 넘었는데 한번끼고 너무별로여서 그냥냅두고있어요ㅠㅠ 그립톡도 케이스에비해서 많이작은 느낌이네요



★★★★★ 1

dbwl**** · 20.03.15. · 옵션 : 디자인: L.공룡 / 컬러: 화이트 / 스마트톡: 화이트 | 신고

0

한달사용기 떨어뜨렸더니 가장자리가 깨졌어요



이 구매자의 처음 리뷰보기 >

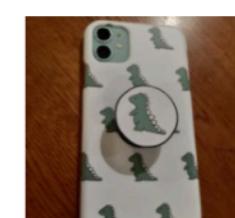


★★★★★ 1

dbwl**** · 20.02.11. · 옵션 : 디자인: L.공룡 / 컬러: 화이트 / 스마트톡: 화이트 | 신고

0

색 완전 달라요!! 참고하세요!!!



사기 당한느낌;;;

개체명 인식

개체명 사전 (Named Entity Dictionary)

구분	의학용어	인물명	회사명
1	불량 식품	사나	오피스디포
2	진행 암	쯔위	아마존
3	전진 피판	정연	구글
4	유해 효과	나연	한국수력원자력
5	무력증	황민현	코레일
6	유산소 운동	강다니엘	애플
7	산소 호흡	옹성우	깨끗한나라
8	공기 삼킴증	전병진	금호아시아나
9	분무제	진상형	몽블랑
10	에어로졸	서지석	샤넬
11	분무 주입법	배현진	티파니
12	대기 요법	현빈	신한생명
13	정동 장애	진세연	신한카드
14	정감성	남지현	삼성생명
15	정동성	주상욱	삼성전자
16	들신경	김태희	롯데칠성음료
17	협력 병원	허맹호	롯데백화점
18	친화력	유아인	인프라웨어
19	친화 크로마토그래피	이승기	카카오
20	무섬유소원 혈증	한예슬	네이버

구분	지역명	영문 지역명	광역시/구
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

E.O.D

Contact

-  <http://www.teanaps.com>
-  fingeredman@gmail.com

WEEK 03

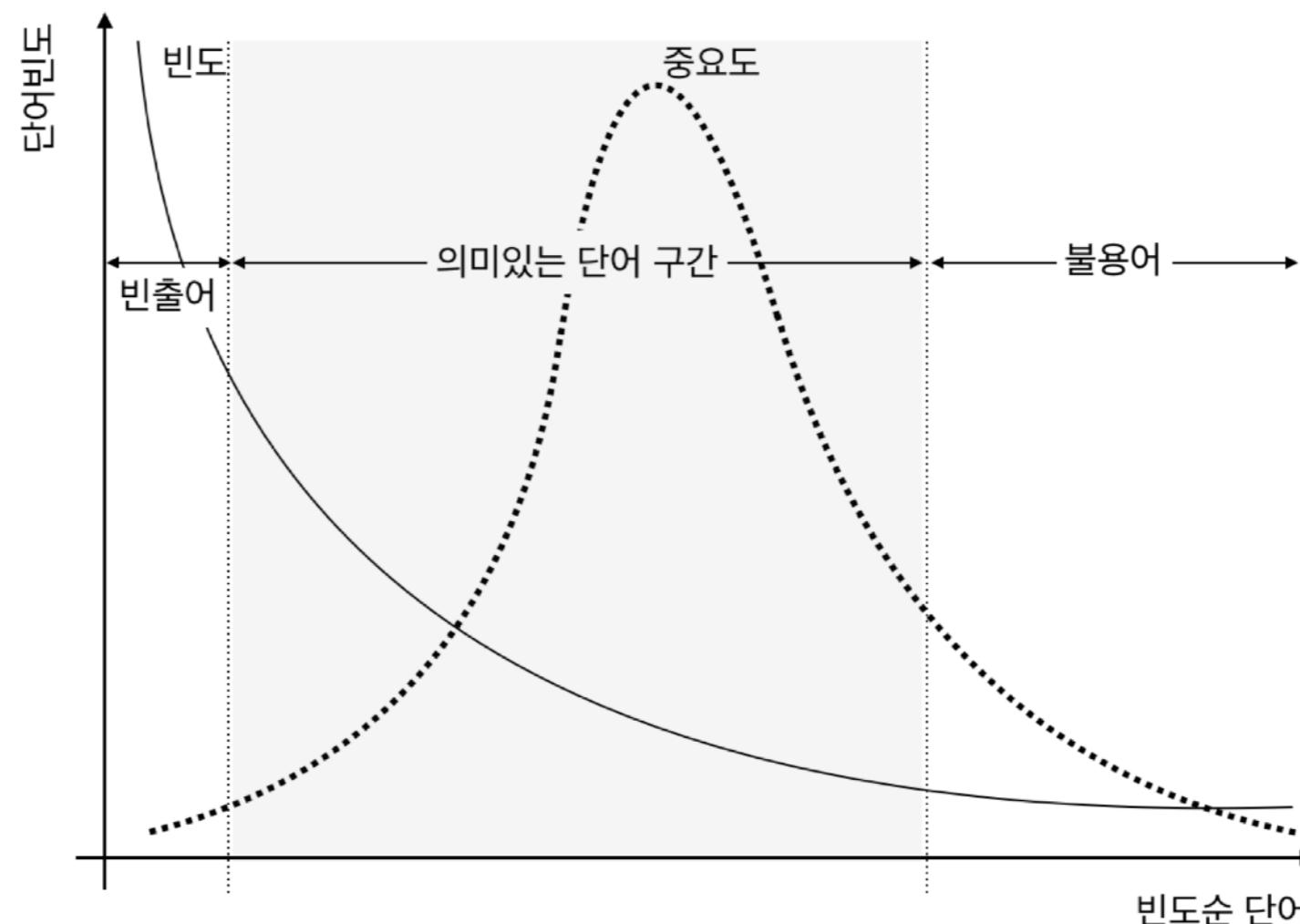
텍스트 마이닝 프로세스: 분석

단어 가중치: 단어빈도

단어빈도 (Term Frequency, TF)

- 특정 단어가 문서에 출현한 횟수로 단어의 특징을 표현하는 가장 간단한 방법
- 간단하지만 가장 빠르게 문서를 표현하고 파악할 수 있으며 기초통계와 같이 분석 전 반드시 거쳐야 하는 과정
- 단어가 너무 희귀한 경우 큰 의미를 부여하기 어려우며, 너무 흔한 경우 의미가 과도하게 부여될 가능성이 있음

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

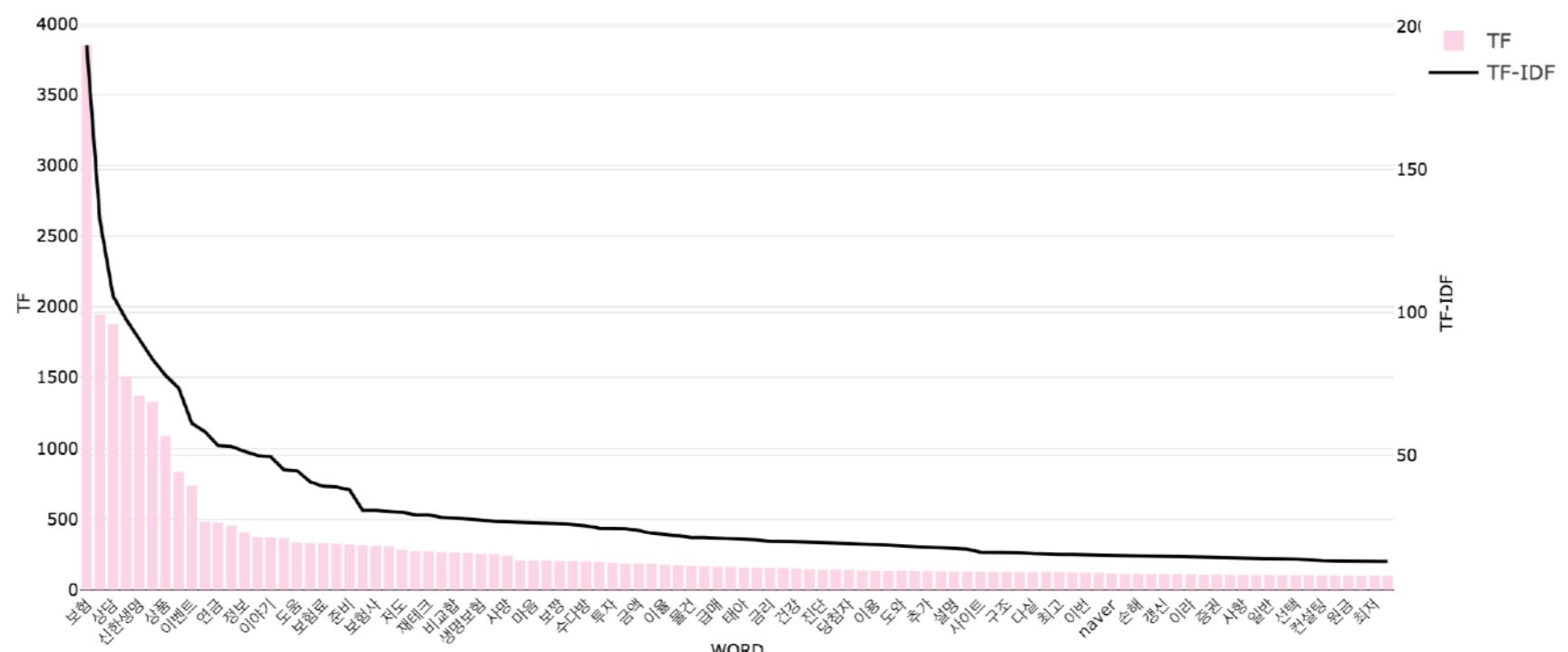


단어 가중치: 단어빈도

TF-IDF (TF-Inverse Document Frequency)

- 역문서빈도 (Inverse Document Frequency, IDF) : 단어가 출현한 문서가 적을수록 단어의 가중치를 낮게 표현하는 방법 (희박성)
 $IDF = 1 + \log(\text{document}/\text{count(word/document)})/\log(\text{document})$
- TF와 IDF 개념을 통합하여, 단어가 문서에 출현한 횟수와 희박성을 동시에 활용해 가중치를 표현하는 방법
 $TF - IDF = \text{Frequency} * IDF$
- 지프의 법칙 (Zipf's law)
 - 1) 자연어에 나타나는 단어들을 출현 횟수가 높은 순으로 정렬하면, 단어의 출현 횟수는 순위에 반비례함
 - 2) 가장 사용 빈도가 높은 단어는 두 번째 단어보다 빈도가 약 두 배 높으며, 세 번째 단어보다는 빈도가 세 배 높음

TF & TF-IDF Graph

* 김수인, 김재원, and 배휘동, 왜 프로그래밍에는 창의성이 필요하다고 할까요, 2017.5.25., <https://medium.com/elice/>.

** references

*** references

문서를 단어 가중치로 표현하는 방법

문서 내 단어의 빈도 계산하기

OhmyNews

"벚꽃 상춘객 올까봐 불도 꺼... 올해는 제발 참아달라"

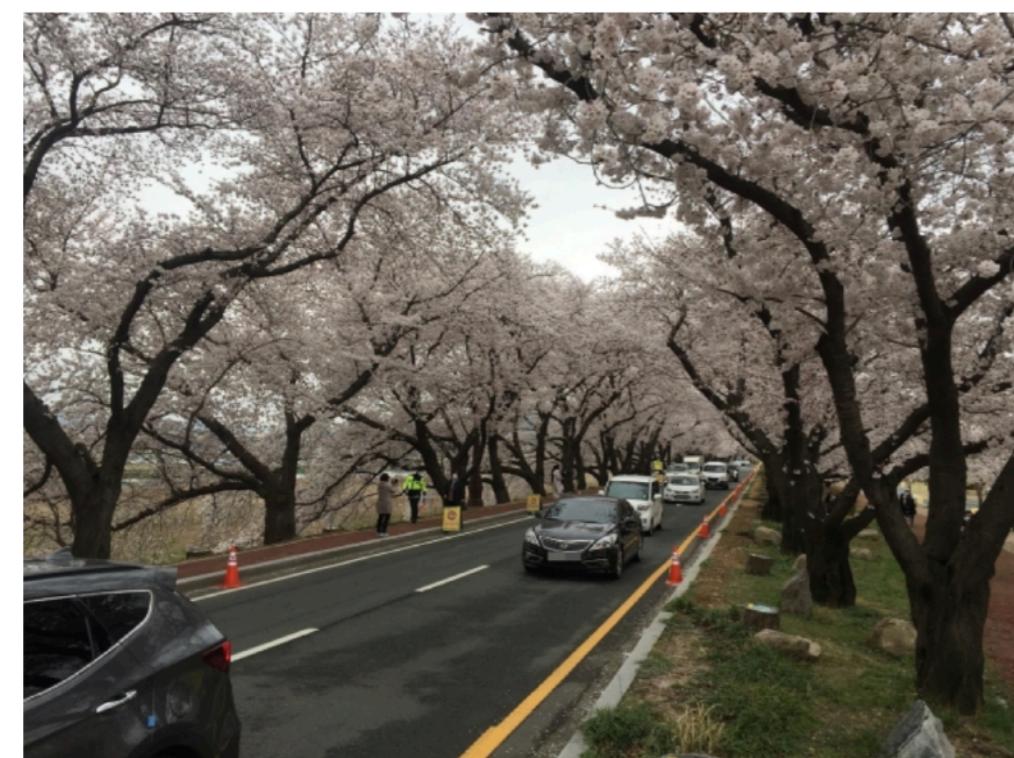
기사입력 2020.03.28. 오후 8:30 기사원문 스크랩 본문듣기 설정

좋아요 27 댓글 14

요약봇 가 둘러보기

[사진] 경주 벚꽃터널, 주정차 금지... 상춘객 몰릴 것에 대비해 야간경관조명도 꺼

[오마이뉴스 한정환 기자]



▲ 28일 주말 오전, 조금은 한산한 경주 흥무로 벚꽃길 모습

© 한정환

천년고도 경주에 벚꽃이 활짝 피었다. 해마다 이맘때쯤이면 경주는 벚꽃 관광객 맞이로 분주했다. 그러나 코로나19 확산으로, 올해는 사정이 달라졌다.

28일 오전 주말을 맞아 벚꽃길로 유명한 흥무로 벚꽃터널을 찾아보았다. 예년에 비해 벚꽃 상춘객들은 많이 줄었다. 코로나19 여파로 많은 관광객이 경주 방문을 자제하고 있는 듯 보인다. 그러나 벚꽃 상춘객이 몰릴 것에 대비하여 벚꽃터널에는 도로 양방향 100m 간격으로 경찰관들이 배치되어 불법 주정차 단속을 하고 있다. 사진을 찍기 위해 차에서 잠시 내리는 것도 안 된다.

관광도시 경주의 특성상 다른 도시처럼 벚꽃 명소를 통제할 수가 없다. 도시 전체가 관광지이고, 대부분 가로수에 벚꽃이 심어져 있어 통제를 하게 되면, 도시 전체가 봉쇄가 되어야 하기 때문에 현수막을 걸어 두고 홍보만 하고 있다. 사진으로나 마 경주 벚꽃 소식을 전하기로 했던 '경주시 벚꽃 알리미'도 잠정 중단된 상태이다.

경주시 관계자는 "지난 22일 일제히 불을 밝힌 야간경관조명도 27일 저녁부터 벚꽃 상춘객이 몰릴 것에 대비하여 일제히 불을 껐다. 벚꽃은 해마다 피니, 올해는 집단 감염이 우려되는 벚꽃 나들이를 내년으로 미루고, 코로나19 확산 방지를 위해 모두가 힘을 합쳐달라"라고 말했다.

경주시는 지난 2월 22일 첫 코로나19 확진자가 발생한 이래 현재까지 총 40명이 양성 판정을 받았으며, 이 중에서 사망 1명, 완치 10명을 제외한 29명이 현재 자가격리 및 생활치료센터에 입소하여 치료를 받고 있다.

문서를 단어 가중치로 표현하는 방법

| 형태소 분석을 통한 단어주머니 생성

구분	문장
문장 01	천년고도 경주에 벚꽃이 활짝 피었다.
문장 02	해마다 이맘때쯤이면 경주는 벚꽃 관광객 맞이로 분주했다.
문장 03	그러나 코로나19 확산으로, 올해는 사정이 달라졌다.
문장 04	8일 오전 주말을 맞아 벚꽃길로 유명한 흥무로 벚꽃터널을 찾아보았다.
문장 05	예년에 비해 벚꽃 상춘객들은 많이 줄었다.
문장 06	코로나19 여파로 많은 관광객이 경주 방문을 자제하고 있는 듯 보인다.
문장 07	그러나 벚꽃 상춘객이 몰릴 것에 대비하여 벚꽃터널에는 도로 양방향 100m 간격으로 경찰관들이 배치되어 불법 주정차 단속을 하고 있다.
문장 08	사진을 찍기 위해 차에서 잠시 내리는 것도 안 된다.
문장 09	관광도시 경주의 특성상 다른 도시처럼 벚꽃 명소를 통제할 수가 없다.
문장 10	도시 전체가 관광지이고, 대부분 가로수에 벚꽃이 심어져 있어 통제를 하게 되면, 도시 전체가 봉쇄가 되어야 하기 때문에 현수막을 걸어 두고 홍보만 하고 있다.
문장 11	사진으로나마 경주 벚꽃 소식을 전하기로 했던 '경주시 벚꽃 알리미'도 잠정 중단된 상태이다.
문장 12	경주시 관계자는 "지난 22일 일제히 불을 밝힌 야간경관조명도 27일 저녁부터 벚꽃 상춘객이 몰릴 것에 대비하여 일제히 불을 껐다.
문장 13	벚꽃은 해마다 피니, 올해는 집단 감염이 우려되는 벚꽃 나들이를 내년으로 미루고, 코로나19 확산 방지를 위해 모두가 힘을 합쳐달라"라고 말했다.
문장 14	경주시는 지난 2월 22일 첫 코로나19 확진자가 발생한 이래 현재까지 총 40명이 양성 판정을 받았으며, 이 중에서 사망 1명, 완치 10명을 제외한 29명이 현재 자가격리 및 생활치료센터에 입소하여 치료를 받고 있다.

문서를 단어 가중치로 표현하는 방법

| 형태소 분석을 통한 단어주머니 생성



문서를 단어 가중치로 표현하는 방법

형태소 분석을 통한 단어주머니 생성

구분	유니그램 (품사=NNG NNP, 불용어 제거)	단어주머니 (74 단어)
문장 01	고도, 경주, 벚꽃	벚꽃 경주 도로 대부분 제외 저녁
문장 02	이맘때, 경주, 벚꽃, 관광객, 분주	도시 단속 잠정 자제
문장 03	코로나, 확산, 올해, 사정	코로나 내년 자제
문장 04	오전, 주말, 벚꽃, 길, 유명, 흥무, 벚꽃, 터널	상춘객 관광지 자가 입소 관광 관계자 고도 유명
문장 05	예년, 벚꽃, 상춘객	전체 사진 치료 터널 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 06	코로나, 여파, 관광객, 경주, 방문, 자제	이맘때 여파 우려 완치 오전 예년 우려 완치
문장 07	벚꽃, 상춘객, 대비, 벚꽃, 터널, 도로, 양방향, 간격, 경찰관, 배치, 불법, 주정차, 단속	대비 확산 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 08	사진	관광객 발생 모두 방문 봉쇄 방지 배치 분주 불법 사망 사정 명소 흥무 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 09	관광, 도시, 경주, 특성, 도시, 벚꽃, 명소, 통제	경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 10	도시, 전체, 관광지, 대부분, 가로수, 벚꽃, 통제, 전체, 봉쇄, 현수막, 흥보	전체 봉쇄 방지 배치 분주 불법 사망 사정 명소 흥무 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 11	사진, 경주, 벚꽃, 소식, 경주시, 벚꽃, 알리, 잠정, 중단, 상태	경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 12	경주시, 관계자, 야간, 경관, 조명, 저녁, 벚꽃, 상춘객, 대비	경주시 관계자 야간 경관 조명 저녁 벚꽃 상춘객 대비 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 13	벚꽃, 올해, 집단, 감염, 우려, 벚꽃, 나들이, 내년, 코로나, 확산, 방시, 위해, 모두	벚꽃 올해 집단 감염 우려 벚꽃 나들이 내년 코로나 확산 방시 위해 모두 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치
문장 14	경주시, 지난, 코로나, 확진, 발생, 현재, 양성, 판정, 사망, 완치, 제외, 격리, 생활, 치료, 센터, 입소, 치료	경주시 지난 코로나 확진 발생 현재 양성 판정 사망 완치 제외 격리 생활 치료 센터 입소 치료 경찰관 경관 격리 감염 간격 나들이 센터 상태 생활 홍보 현재 현수막 판정 특성 집단 중단 주정차 양성 여파 우려 완치 오전 예년 우려 완치

문서를 단어 가중치로 표현하는 방법

형태소 분석을 통한 단어주머니 생성

구분	벚꽃	경주	도시	코로나	상춘객	경주시	전체	사진	치료	터널	통제	올해	대비	확산	관광객	발생	모두	방문	봉쇄	방지	배치	분주	불법	사망	사정	명소	흥무
문장 01	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 02	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	
문장 03	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	
문장 04	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
문장 05	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 06	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	
문장 07	2	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	
문장 08	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 09	1	1	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
문장 10	1	0	2	0	0	0	2	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
문장 11	2	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 12	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
문장 13	2	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	
문장 14	0	0	0	1	0	1	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	

단어 가중치: 네트워크 중심성

동시출현 분석 (Co-word Analysis)

- 문서에 서로다른 두 단어의 동시출현 횟수와 네트워크 중심성을 통해 단어의 특징을 표현하는 방법
 - 두 단어 사이의 동시출현을 연관성의 척도로 취급하고, 그 관계를 네트워크 중심성으로 표현하여 가중치를 계산함
 - **연관어** (공기어, Co-word) : 하나의 문서에서 함께 출현하여 서로 밀접한 의미관계를 가지는 단어

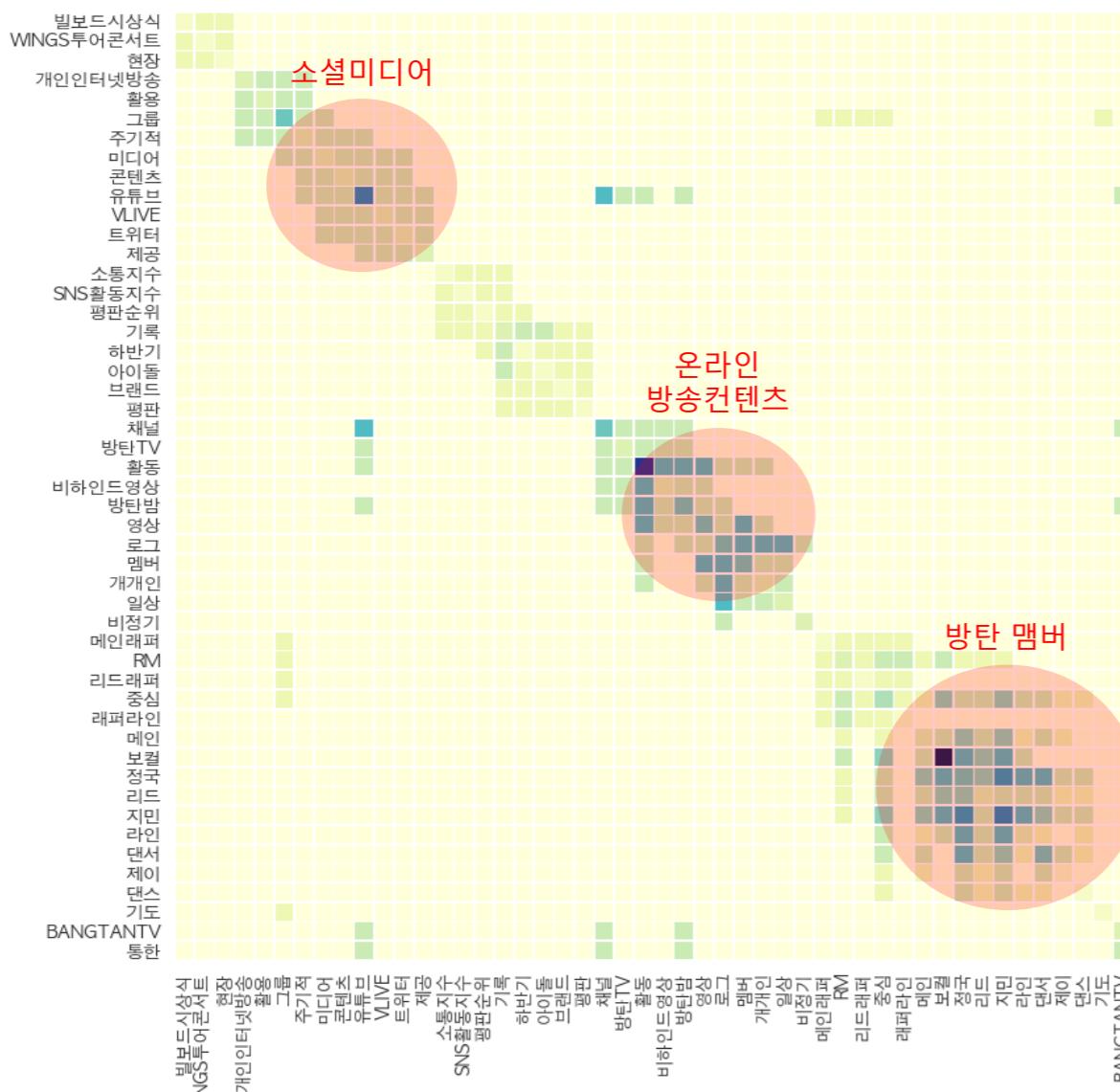
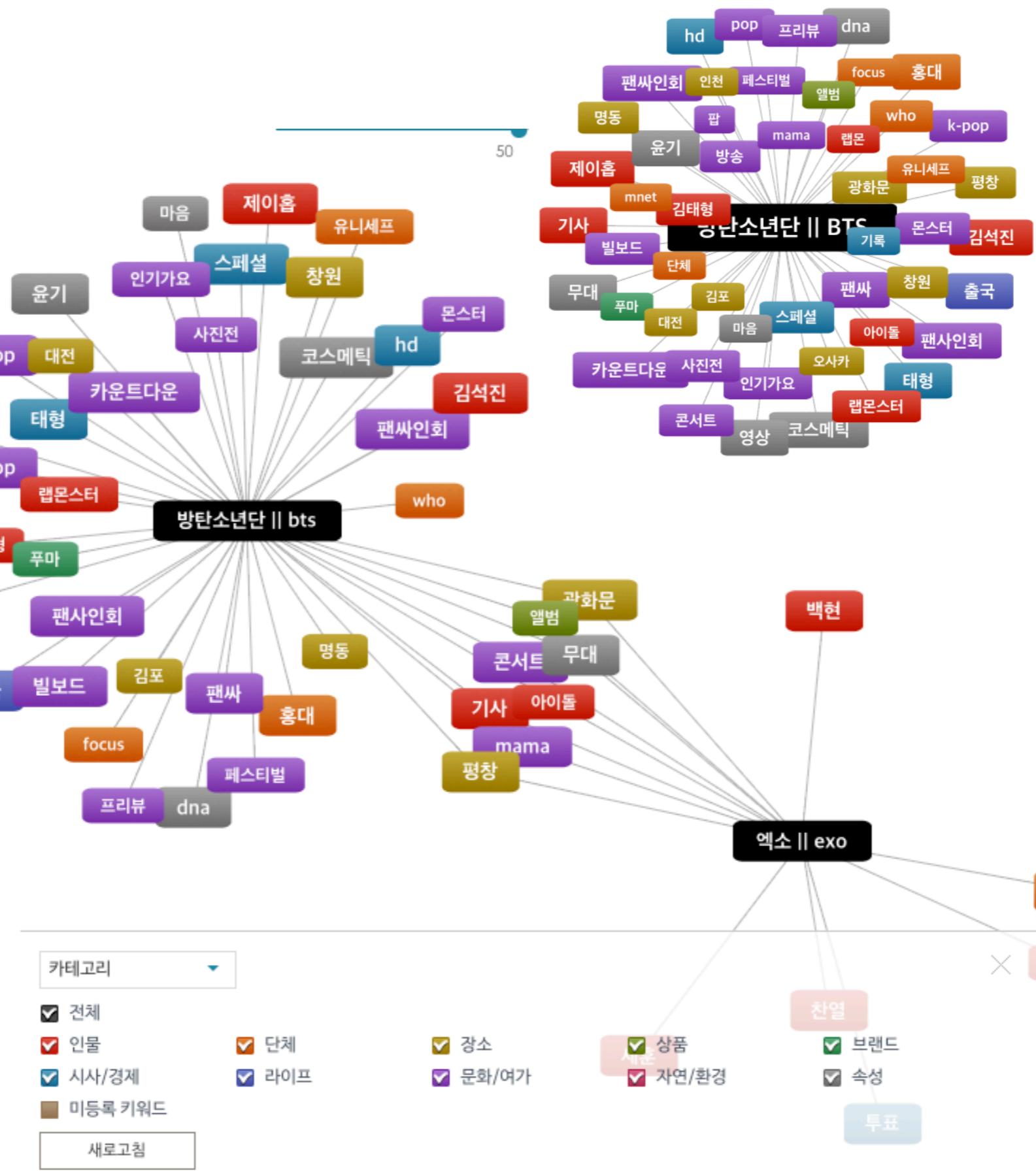


표1 ‘아쿠르트 아줌마’ 연관어 변화

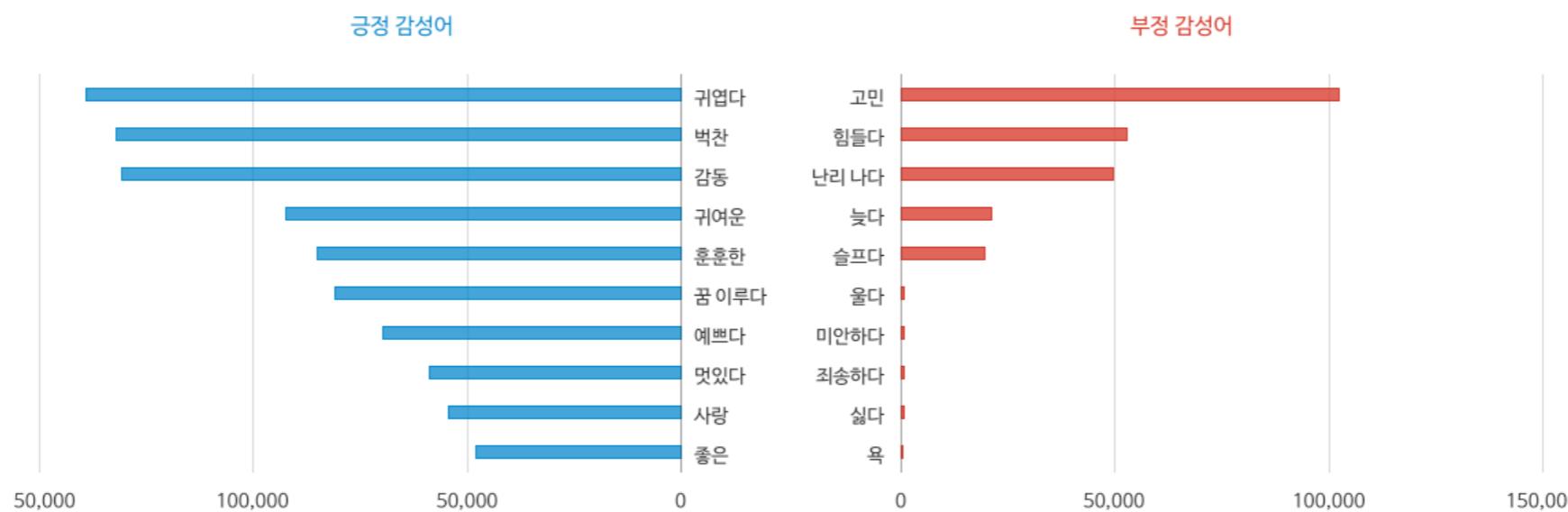
아쿠르트 아줌마는 여전히 '아쿠르트' 와의 연관도가 가장 높지만 2016년 들어 '커피' 및 '크림치즈' 제품 연관어와 '10일'이라는 키워드가 등장. 아쿠르트 아줌마는 '배달하는' 역할에서 막상 제품을 위해 '만나고' '찾고' '발견하는' 대상으로 변화 중.

2013년		2014년		2015년		2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	야쿠르트	21.3%	1	야쿠르트	26.3%	1	야쿠르트	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	야쿠르트 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%						



NO.	연관어 - 엑소 exo	건수
1	■ 투표	539,724
2	■ 콘서트	469,536
3	■ 백현	373,347
4	■ 찬열	363,191
5	■ 세훈	342,956
6	■ 엑소엘	333,724
7	■ mama	302,523
8	■ 시우민	185,113
9	■ 그룹	153,995
10	■ 대상	153,270
11	■ 티켓	147,756
12	■ 멜론	144,508
13	■ 파워	136,761
14	■ video	134,713
15	■ 평창	134,577
16	■ 광화문	133,053
17	■ 단독콘서트	120,636
18	■ 앨범	118,598
19	■ 레이	111,334
20	■ 멤버	107,329
21	■ 가수	106,322
22	■ 디오	104,295
23	■ 기사	98,662
24	■ 아이돌	98,416
25	■ 고척스카이돔	95,535
26	■ music	88,950
27	■ 올림픽	88,537
28	■ 종대	88,349

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								
<input checked="" type="checkbox"/> 전체	1	방탄소년단	826,236	방탄소년단	987,650	방탄소년단	798,031	방탄소년단	549,312	방탄소년단	1,481,373
<input checked="" type="checkbox"/> 인물	2	태형	449,263	태형	552,652	기사	260,201	김석진	147,539	평창	289,317
<input checked="" type="checkbox"/> 단체	3	코스메틱	404,925	홍대	428,101	태형	217,125	태형	141,507	광화문	250,964
<input checked="" type="checkbox"/> 장소	4	기사	260,803	기사	423,304	출국	205,573	hd	108,600	콘서트	214,887
<input checked="" type="checkbox"/> 상품	5	명동	253,605	dna	362,319	푸마	187,739	무대	106,730	기사	205,638
<input checked="" type="checkbox"/> 브랜드	6	팬싸인회	207,776	카운트다운	355,310	김석진	165,320	타이페이	95,120	유니세프	205,279
<input checked="" type="checkbox"/> 라이프	7	팬싸인회	200,843	팬싸인회	329,385	dvd	128,894	잼	92,580	캠페인	131,943
<input checked="" type="checkbox"/> 시사/경제	8	마음	198,856	영상	297,121	hd	127,530	제이홉	77,808	무대	119,974
<input checked="" type="checkbox"/> 문화/여가	9	who	186,415	팬싸인회	283,711	dna	124,366	콘서트	77,700	스페셜	119,423
<input checked="" type="checkbox"/> 자연/환경	10	dna	184,025	팬싸	222,826	제이홉	117,073	윤기	76,071	리허설	113,370
<input checked="" type="checkbox"/> 속성	11	팬싸	175,731	출국	197,729	mama	107,045	대만	66,871	올림픽	109,404

표1 '야쿠르트 아줌마' 연관어 변화

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

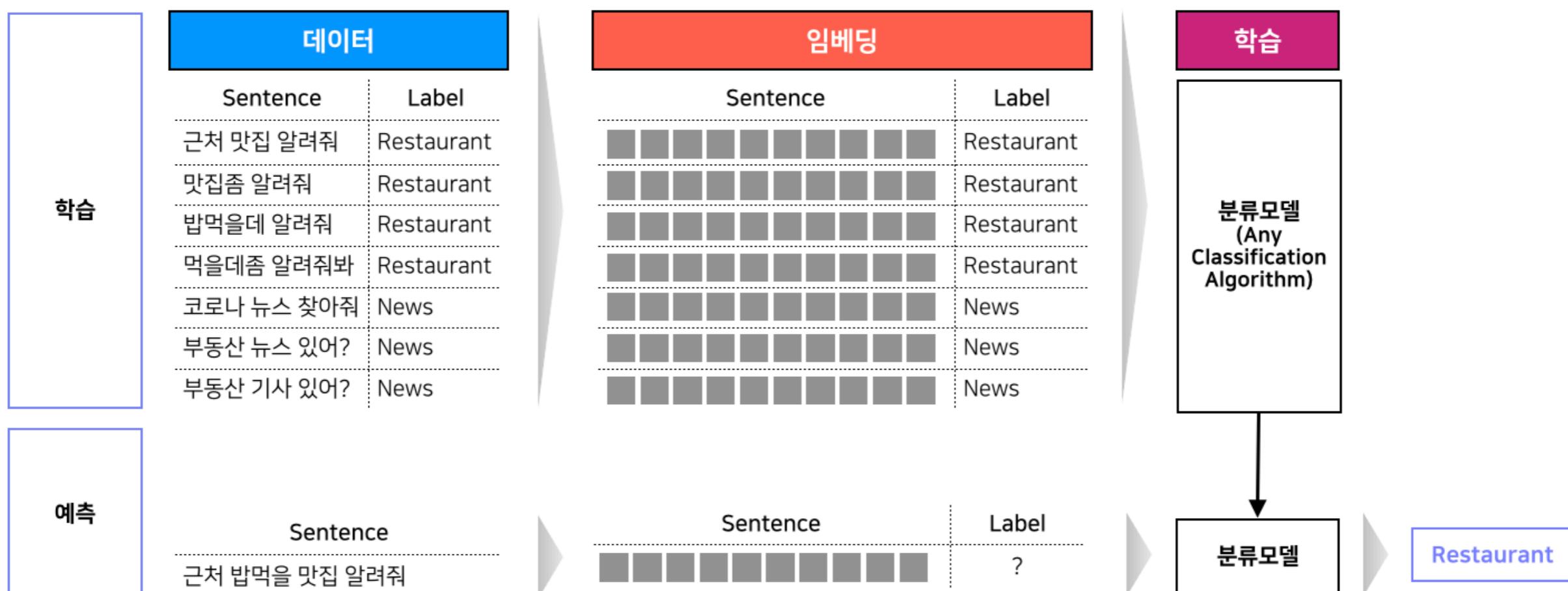
2013년			2014년			2015년			2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	야쿠르트	21.3%	1	야쿠르트	26.3%	1	야쿠르트	26.6%	1	야쿠르트	13.1%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%	2	콜드브루	8.2%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%	3	커피	7.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%	4	맛	6.6%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%	5	끼리	5.7%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%	6	치즈	5.3%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%	7	과자	5.0%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%	8	아메리카노	4.1%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%	9	먹다	3.3%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%	10	크림치즈	3.1%
11	야쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%	11	라떼	2.8%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%	12	만나다	2.7%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%	13	가격	2.4%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%	14	찾다	1.9%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%	15	아침	1.8%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%	16	10일	1.6%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%	17	엄마	1.5%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%	18	우유	1.4%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%	19	팔다	1.3%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%	20	발견하다	1.3%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%	21	사다	1.2%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%	22	인기	1.2%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%	23	편의점	1.2%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%	24	끼리딥앤크런치	1.1%
...				
29	팔다	1.4%	29	묻다	1.3%	29	배달하다	1.3%	29	구입하다	1.0%

■ 상승 키워드 ■ 하락 키워드 ■ 신규 키워드

텍스트 분류하는 방법

텍스트 분류 (Text Classification)

- 텍스트, 문장 또는 문서를 입력으로 받아 사전에 정의된 클래스(class) 중에 어디에 속하는지 분류하는 과정
- 텍스트를 입력으로 받아 불연속적인 값(범주, descret data)으로 출력하는 문제를 해결하는 과정
- 나이브 베이즈, SVM 등 머신러닝 기법부터 RNN, CNN 등 딥러닝을 활용한 방식으로 문제해결이 가능함
- 활용 예 : 감성분석 (긍정, 중립, 부정), 스팸탐지 (스팸, 비스팸), 챗봇 의도분류, 뉴스기사 주제분류 등



문서의 감성수준을 평가하는 방법

감성분석 (Sentiment Analysis)

- 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.



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



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

[감성사전 예시]

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

문서의 감성수준을 평가하는 방법

| 상용/연구용 감성사전 종류

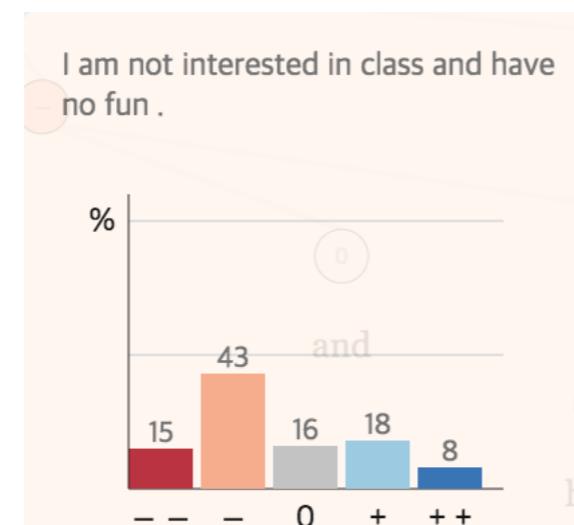
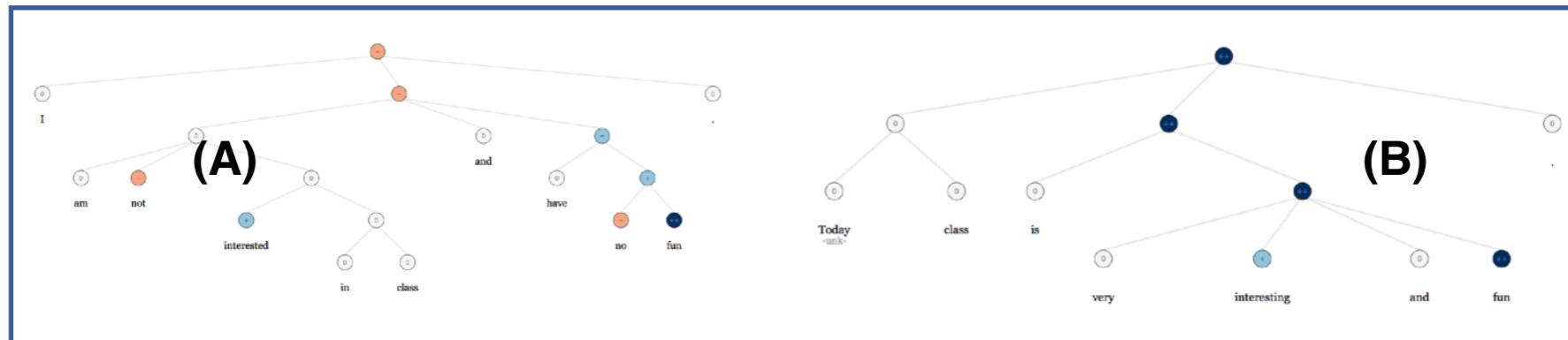
- Linguistic Inquiry and Word Count (LIWC) - <http://www.liwc.net/>
- MPQA Subjectivity Cues Lexicon - <http://www.cs.pitt.edu/>
- SentiWordNet - <http://sentiwordnet.isti.cnr.it/>
- KOSAC - <http://word.snu.ac.kr/kosac/icon.php>

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

문서의 감성수준을 평가하는 방법

[머신러닝(딥러닝) 기반 감성분석]

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



E.O.D

Contact

 <http://www.teanaps.com>

 fingeredman@gmail.com

WEEK 03

텍스트 마이닝 프로세스: 시각화

텍스트 데이터 시각화 (Visualization)

테이블 (Table)

- 분석결과를 테이블 형태로 행과 열을 구분하여 표현하는 방법

표1 '아쿠르트 아줌마' 연관어 변화											
2013년			2014년			2015년			2016년		
No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중	No.	연관어	언급 비중
1	아쿠르트	21.3%	1	아쿠르트	26.3%	1	아쿠르트	26.6%	1	아쿠르트	13.1%
2	먹다	4.9%	2	건강	4.5%	2	집	4.7%	2	콜드브루	8.2%
3	아침	4.4%	3	아침	4.0%	3	아침	4.4%	3	커피	7.4%
4	엄마	4.2%	4	집	3.6%	4	맛	3.9%	4	맛	6.6%
5	집	3.5%	5	제품	3.4%	5	먹다	3.4%	5	끼리	5.7%
6	오다	2.8%	6	엄마	3.3%	6	사다	2.8%	6	치즈	5.3%
7	사다	2.7%	7	맛	2.7%	7	주다	2.8%	7	과자	5.0%
8	주다	2.5%	8	같다	2.6%	8	다니다	2.7%	8	아메리카노	4.1%
9	구입하다	2.4%	9	우유	2.6%	9	엄마	2.6%	9	먹다	3.3%
10	아이	2.4%	10	주다	2.2%	10	우유	2.1%	10	크림치즈	3.1%
11	아쿠르트 주다	2.3%	11	먹다	2.2%	11	만나다	2.1%	11	라떼	2.8%
12	배달하다	2.3%	12	만나다	2.0%	12	제품	2.0%	12	만나다	2.7%
13	수입	2.3%	13	사다	1.9%	13	사진	2.0%	13	가격	2.4%
14	다니다	2.1%	14	알다	1.9%	14	나오다	2.0%	14	찾다	1.9%
15	얼려먹다	2.0%	15	배달하다	1.8%	15	팔다	1.9%	15	아침	1.8%
16	살다	2.0%	16	다니다	1.8%	16	지나가다	1.8%	16	10일	1.6%
17	제품	2.0%	17	하루야채	1.7%	17	하나	1.7%	17	엄마	1.5%
18	세븐	1.8%	18	나누다	1.7%	18	판매	1.7%	18	우유	1.4%
19	가다	1.8%	19	지나가다	1.6%	19	일하다	1.6%	19	팔다	1.3%
20	자녀	1.8%	20	세븐	1.5%	20	오다	1.6%	20	발견하다	1.3%
21	만나다	1.8%	21	수입	1.5%	21	찾다	1.6%	21	사다	1.2%
22	마시다	1.7%	22	찾다	2.3%	22	음료	1.5%	22	인기	1.2%
23	유산균	1.7%	23	노인	1.4%	23	마시다	1.4%	23	편의점	1.2%
24	일하다	1.7%	24	마시다	1.4%	24	길	1.4%	24	끼리딥앤크런치	1.1%
...				
29	팔다	1.4%	29	묻다	1.3%	29	배달하다	1.3%	29	구입하다	1.0%

■ 상승 키워드 ■ 하락 키워드 ■ 신규 키워드

<표 6> 불행요인 세부 토픽 모델링 결과

#	토 픽	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, 뱀띠, 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,	

텍스트 데이터 시각화 (Visualization)

워드클라우드 (Wordcloud)

- 단어의 가중치(TF , $TF-IDF$, 중심성 등)를 단어의 크기로 반영하여 그 분포를 아름답게 표현하는 방법
 - 가중치를 비롯해 단어의 색깔, 배치 등을 통해 더 많은 정보를 표현할 수 있음



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

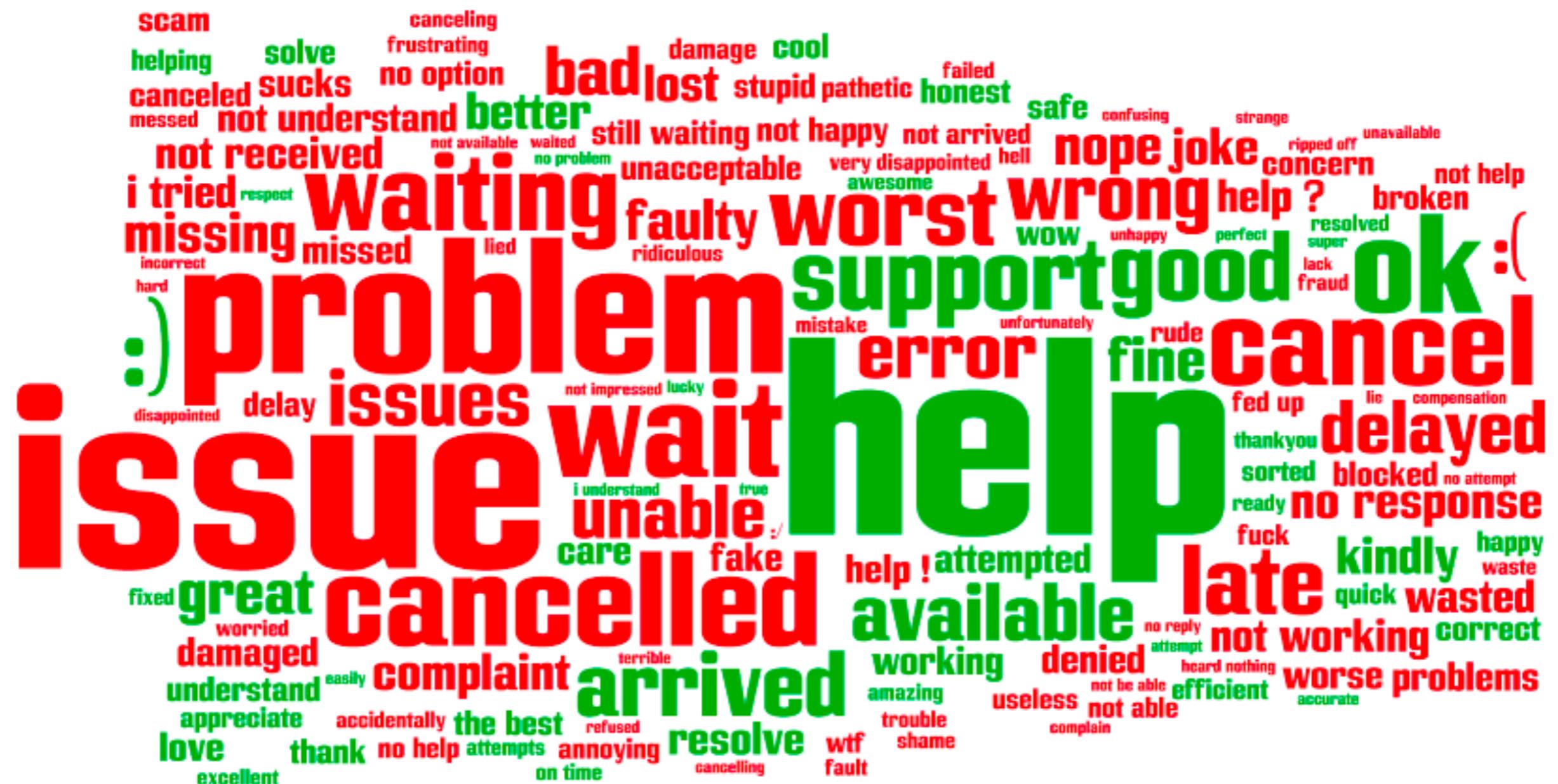
** 워드클리우드.kr, 열정 긍정적 - 워드클라우드, 2017.11.5., <http://wordcloud.kr/1295>

*** Kumo - Java Word Cloud, <http://kennycason.com/posts/2014-07-03-kumo-wordcloud.html>

**** CX DATA SCIENCE, SIMPLY SENTIMENT 2 SUPPORT, https://www.cxdatascience.com/ssv2_support

텍스트 데이터 시각화 (Visualization)

워드클라우드 (Wordcloud)



텍스트 데이터 시각화 (Visualization)

워드클라우드: 무엇이 잘못되었을까요?



* How we build MyRealTrip. 마이리얼트립 여행 후기 데이터 분석

<https://medium.com/myrealtrip-product/%EB%A7%88%EC%9D%B4%EB%A6%AC%EC%96%BC%ED%8A%B8%EB%A6%BD-%EC%97%AC%ED%96%89-%ED%9B%84%EA%B8%BO-%EB%8D%EC%9D%B4%ED%84%BO-%EB%B6%84%EC%84%9D-be3f6c557ca2/>

** 둘아비즈니스리뷰 262호 전형적인 워드클라우드. 2018.11.. https://dbr.donga.com/graphic/view/qdbr_no/699

*** references

텍스트 데이터 시각화 (Visualization)

그래프와 네트워크 (Graph & Network)

- 단어 사이의 관계와 그 강도를 선으로 연결하여 표현하는 방법
 - **그래프 (graph)** : 문서 또는 단어의 정량화된 특징을 도표로 표현하는 방법
 - **네트워크 (network)** : 단어를 노드, 단어들 사이의 관계를 엣지로 취급하여 네트워크를 표현하는 방법

TABLE 2
PART OF SPEECH COUNT

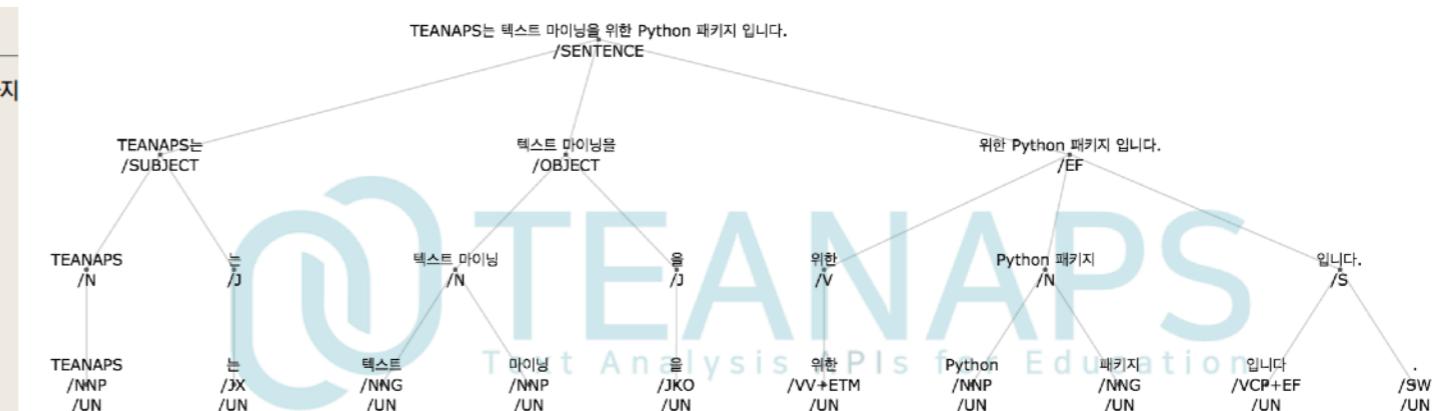
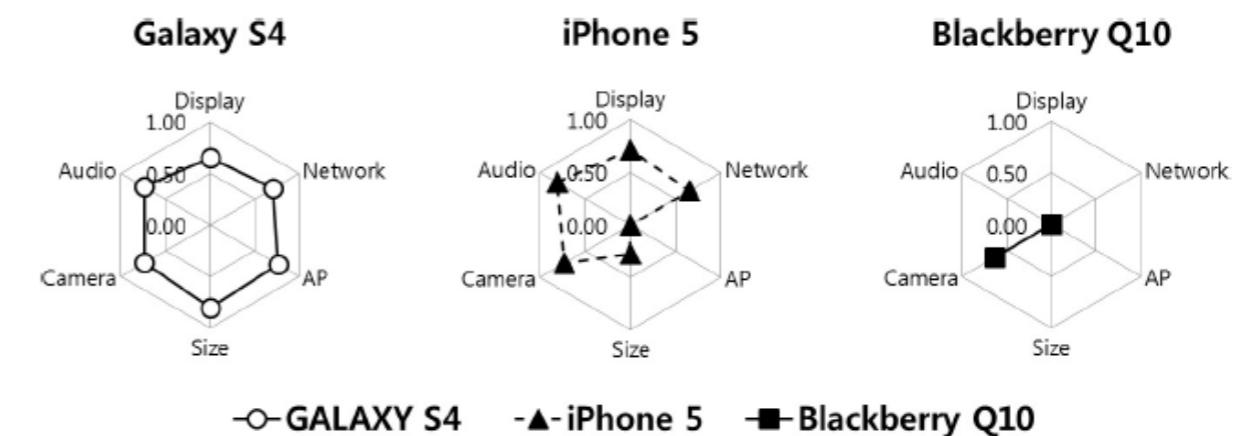
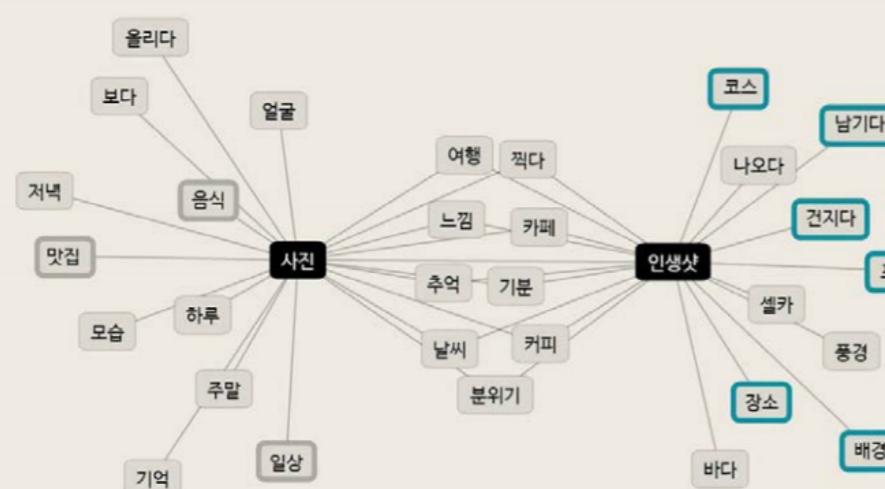
Count of words categorized by part of speech (POS)

	part of speech									
	n+v+adj+adv		nouns (n)		verbs (v)		adjectives (adj)		adverbs (adv)	
Hillary Clinton	7,636*	2,173	3,581*	1,117	2,322*	799	1,306*	543	427*	97
	40.5%	28.5%	46.9%	31.2%	30.4%	34.4%	17.1%	41.6%	5.6%	22.7%
Donald Trump	8,158*	1,752	3,639*	953	2,375*	588	1,621*	550	523*	83
	37.9%	21.5%	44.6%	26.2%	29.1%	24.8%	19.9%	33.9%	6.4%	15.9%
total	15,794*	3,008	7,220*	1,635	4,697*	1,102	2,927*	886	950*	127
	39.1%	19.0%	45.7%	22.6%	29.7%	23.5%	18.5%	30.3%	6.0%	13.4%

Fields with * (e.g. 155*) link to data files and Wordles. Hover over the field to show these links. See analysis.

그림3 ‘사진’ vs. ‘인생샷’의 연관어 네트워크

사진은 맛집, 음식 등 일상적인 상황에서 사진을 찍어 올리는 것. 인생샷은 특별한 장소 및 배경에서 추억을 남기고 우연히 건지기도 하지만 수고해서 찍은 수많은 커들 중 가장 성공한 커들을 인생샷으로 견дин는 것.



* 백경희(DBR), “매력을 소비하는 나는 덕후! 즐거움을 위해 기꺼이 지갑을 연다”, 2017.1., http://dbr.donga.com/article/view/1203/article_no/793

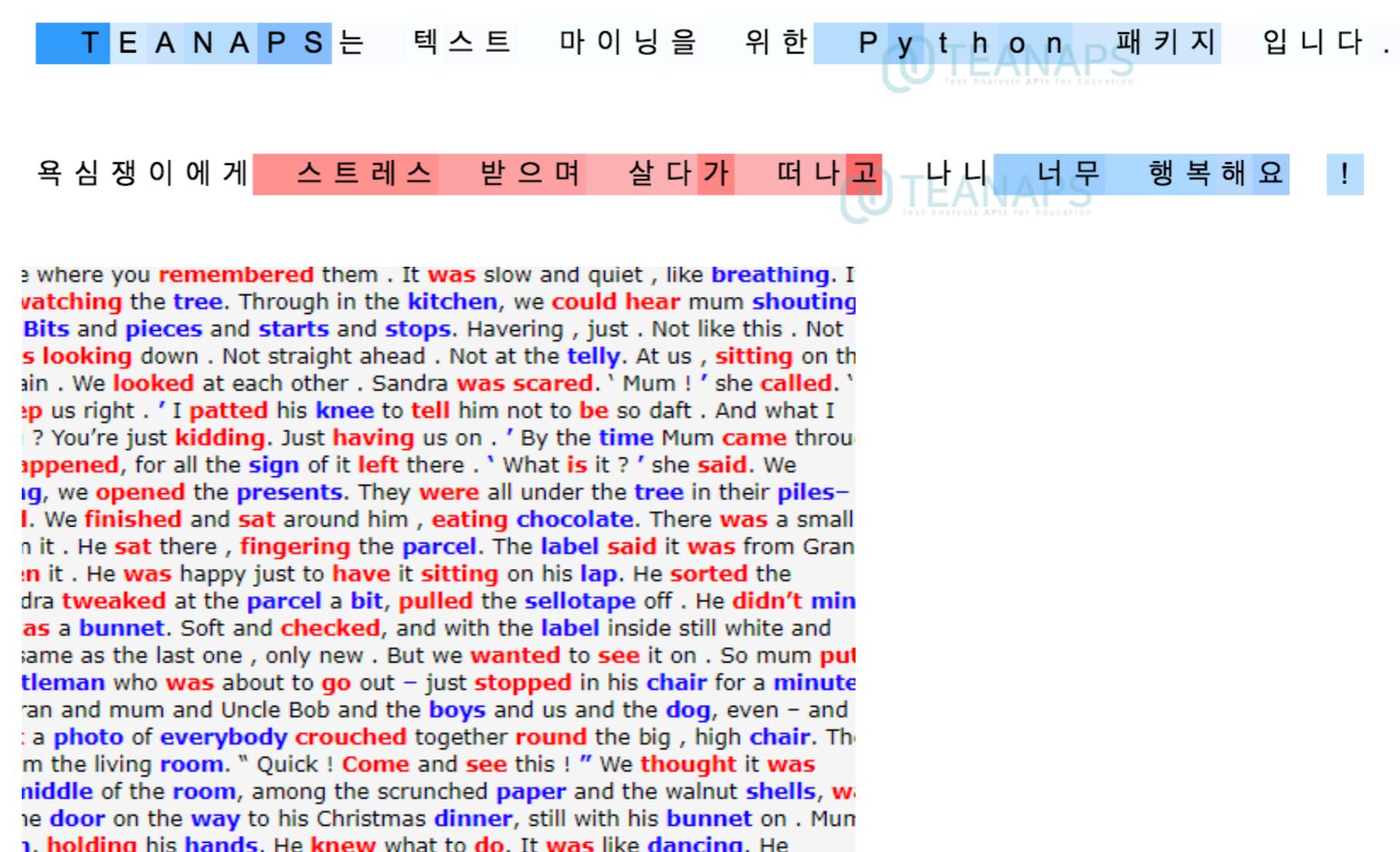
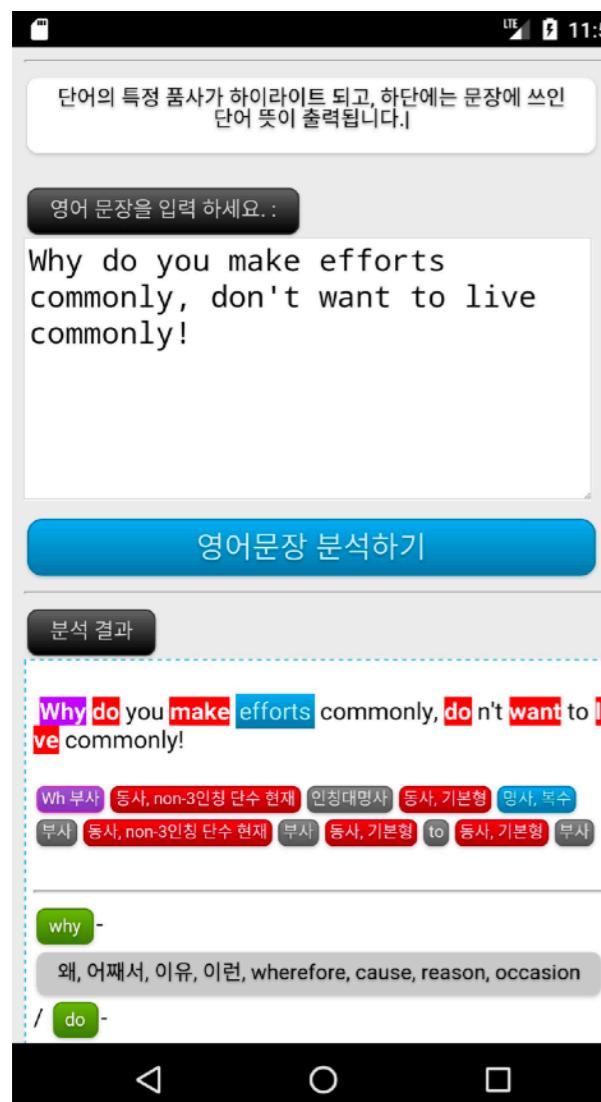
** 최홍규(슬로우뉴스), 2016 미국 대선을 보여주는 텍스트 마이닝 분석방법들 2017.1.9., <http://slownews.kr/609>

*** 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. 205-214).

텍스트 데이터 시각화 (Visualization)

강조표시 (Highlighting)

- 문서 또는 문장의 일부를 색상으로 강조하여 표현하는 방법
- 음절, 단어, 문장 단위로 강조범위를 지정할 수 있으며 색상에 따라 다양한 특징을 표현할 수 있음



E.O.D

Contact

 <http://www.teanaps.com>

 fingeredman@gmail.com