

ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)

WEEK 06

Representing Text Data

문서를 표현하는 방법

단어 주머니 (Bag of Words)

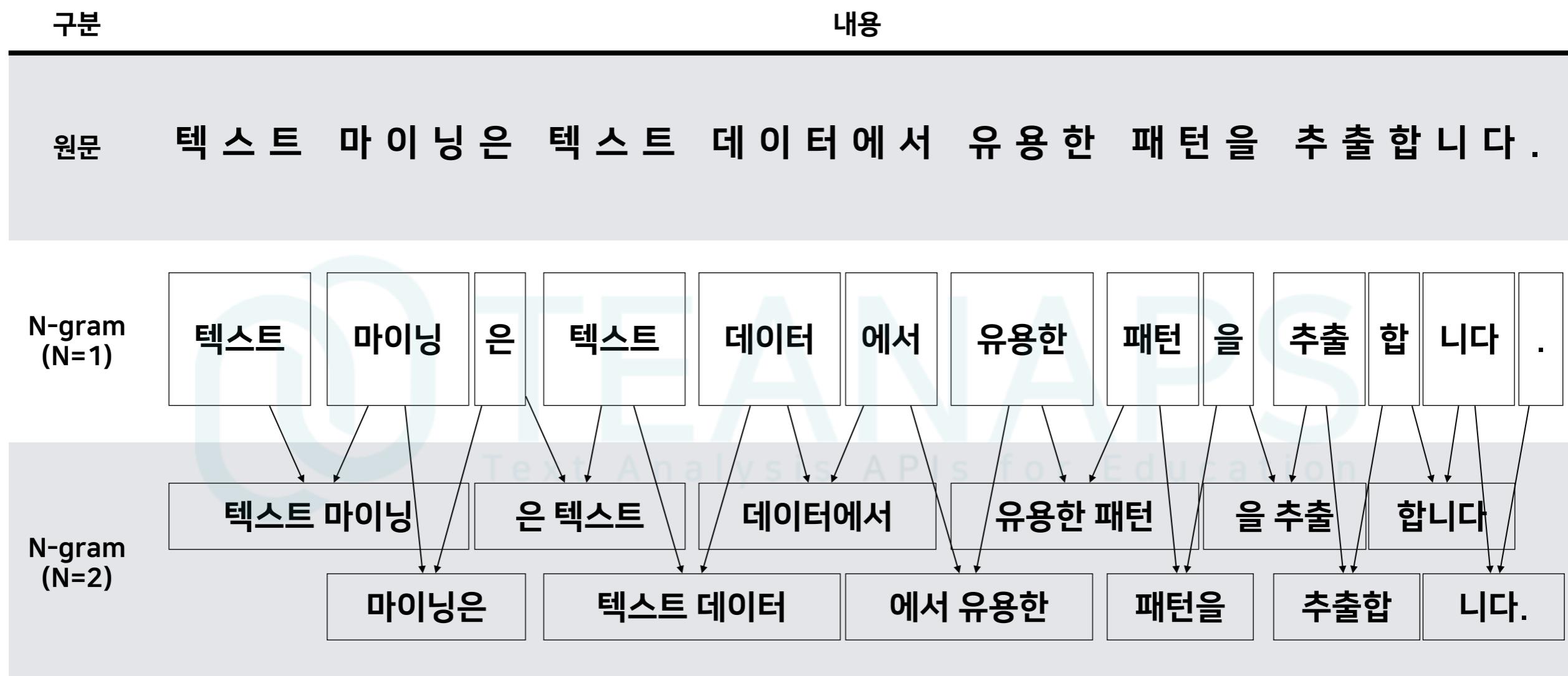
- 문서에 함께 사용된 단어의 집합
- 중복된 단어는 하나로 취급하며, 순서에 의미를 고려하지 않음
- “아버지가 방에 들어가신다.” → [“아버지”, “가”, “방”, “에”, “들어가다”]



문서를 표현하는 방법

N-그램 (N-gram)

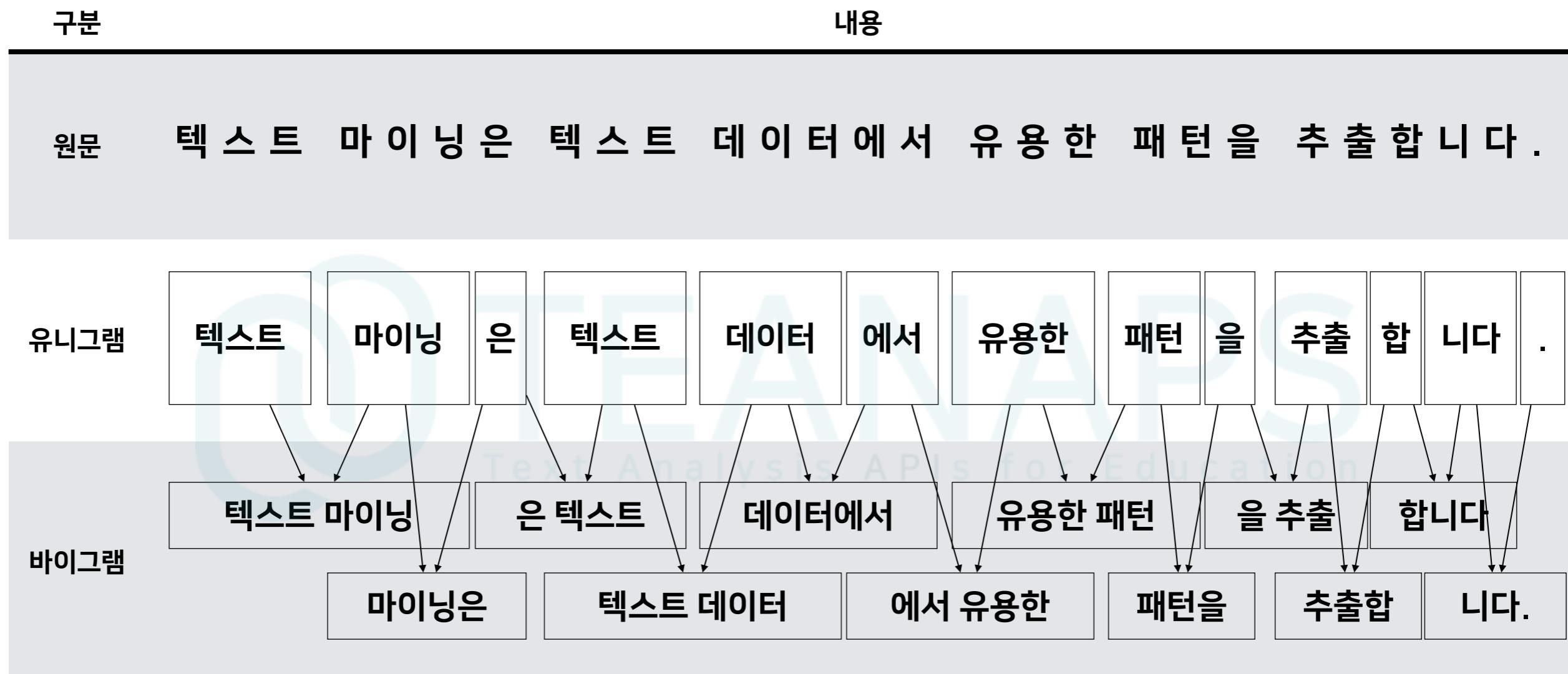
- N 개의 단어 또는 형태소를 하나의 토큰(token)으로 취급하여 문서를 표현하는 방법
- N이 증가할 수록 텍스트 처리에 필요한 연산량이 기하급수적으로 증가하기 때문에 보통 N은 3을 넘기 어려움
- 경우에 따라 불용어를 제외한 (명사+명사), (형용사+명사) 등 조합을 가진 N-그램만 사용하기도 함



문서를 표현하는 방법

유니-그램 & 바이-그램 (Uni-gram & Bi-gram)

- 유니그램 (uni-gram) : 독립된 하나의 단어 또는 형태소를 토큰으로 취급하여 문서를 표현하는 방법
- 바이그램 (bi-gram) : 두 개의 단어 또는 형태소를 토큰으로 취급하여 문서를 표현하는 방법
- N-그램에서 N이 클수록 하나의 토큰에 더 많은 정보를 내포할 수 있음



문서를 표현하는 방법

불용어 (Stopword)

- 텍스트 데이터를 분석함에 있어, 분석 결과에 출현하더라도 큰 의미가 없거나 방해되는 단어 또는 형태소
(그거, 여기, 이제, 은, 는, 이, 등, a, an, the, of, the, !, @, #, \$, %, & 등)
- 정보를 포함하기 보다는 주로 기능적인 역할(조사, 어미 등)을 하는 단어 또는 형태소에 해당하는 경우가 많음
- 일반적으로 한국어는 외자, 영어는 알파벳 두 자 단어 또는 형태소를 대부분 불용어 취급함
- 빈출어 (Common-word) : 너무 많이 출현하여 분석 결과에서 의미 또는 중요도가 떨어지는 단어
(기사, 기자, 제목, 사진, 네이버, 검색, 보다, 연기, 평점, 공감, 비공감 등)

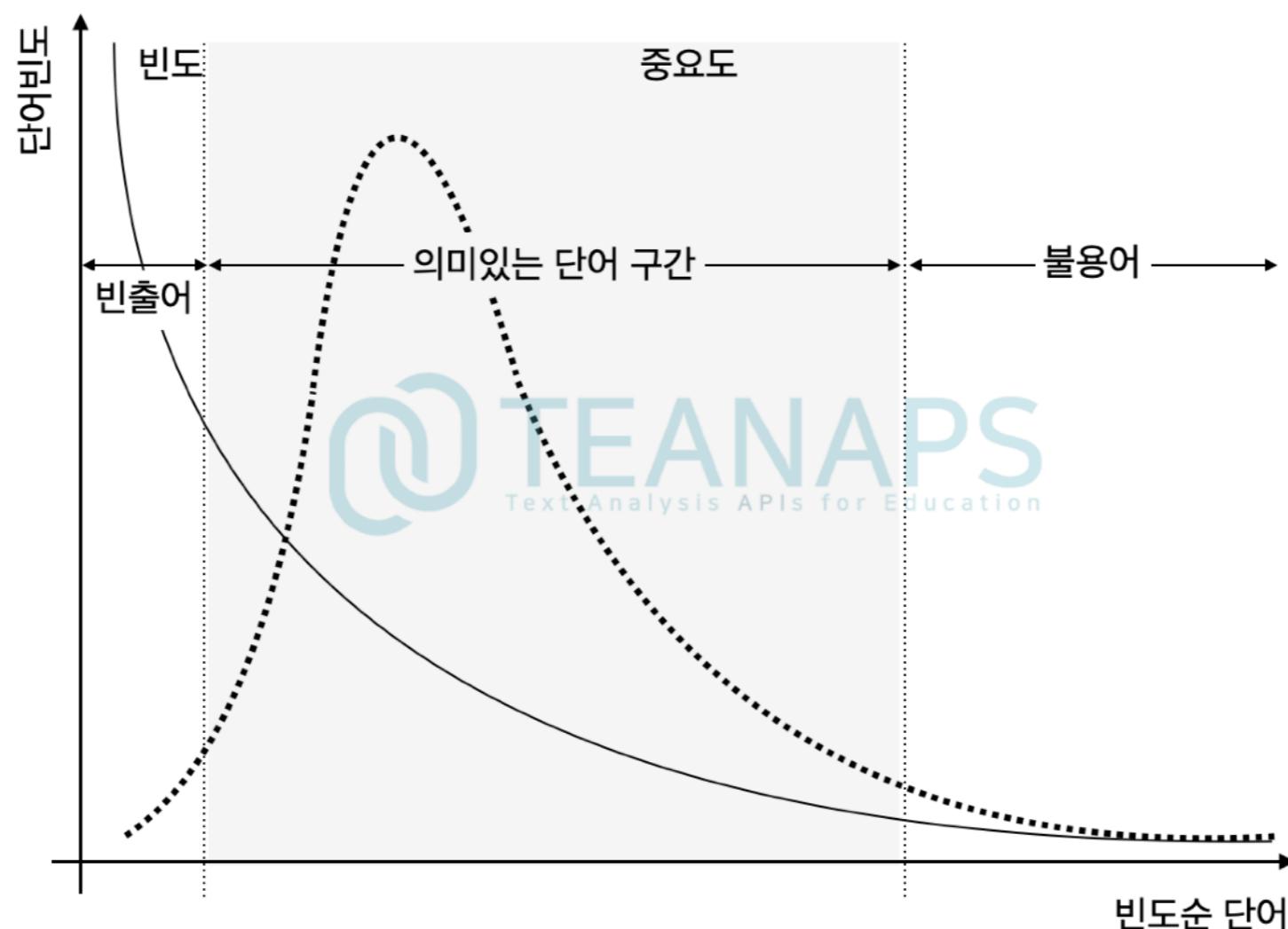


단어 가중치: 단어빈도

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

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

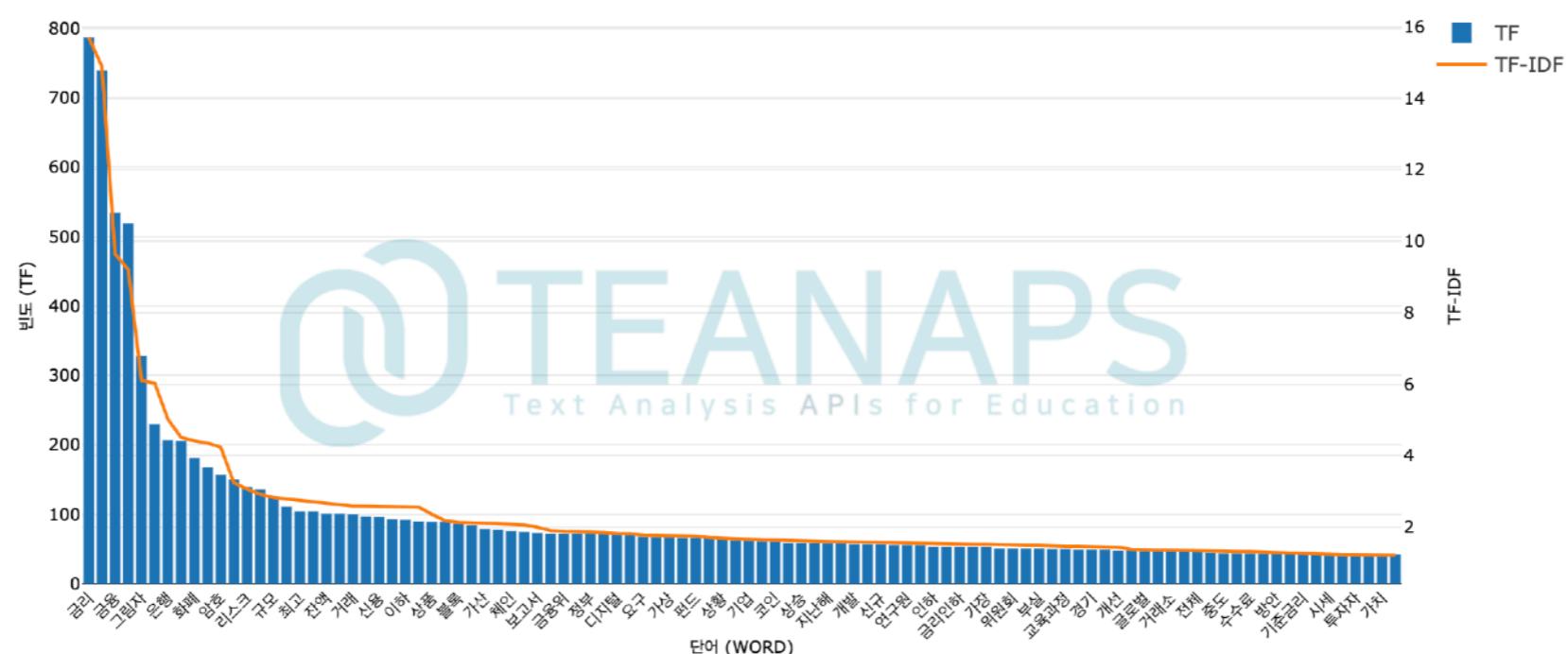


단어 가중치: 단어빈도

TF-IDF (TF-Inverse Document Frequency)

- 역문서빈도 (Inverse Document Frequency, IDF) : 단어가 출현한 문서가 적을수록 단어의 가중치를 높게 표현하는 방법 (희박성)
 $IDF = \text{count}(\text{document}) / (1 + \text{count}(\text{word} | \text{document}))$
- 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

단어빈도 및 TF-IDF (TF & TF-IDF)



* 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"

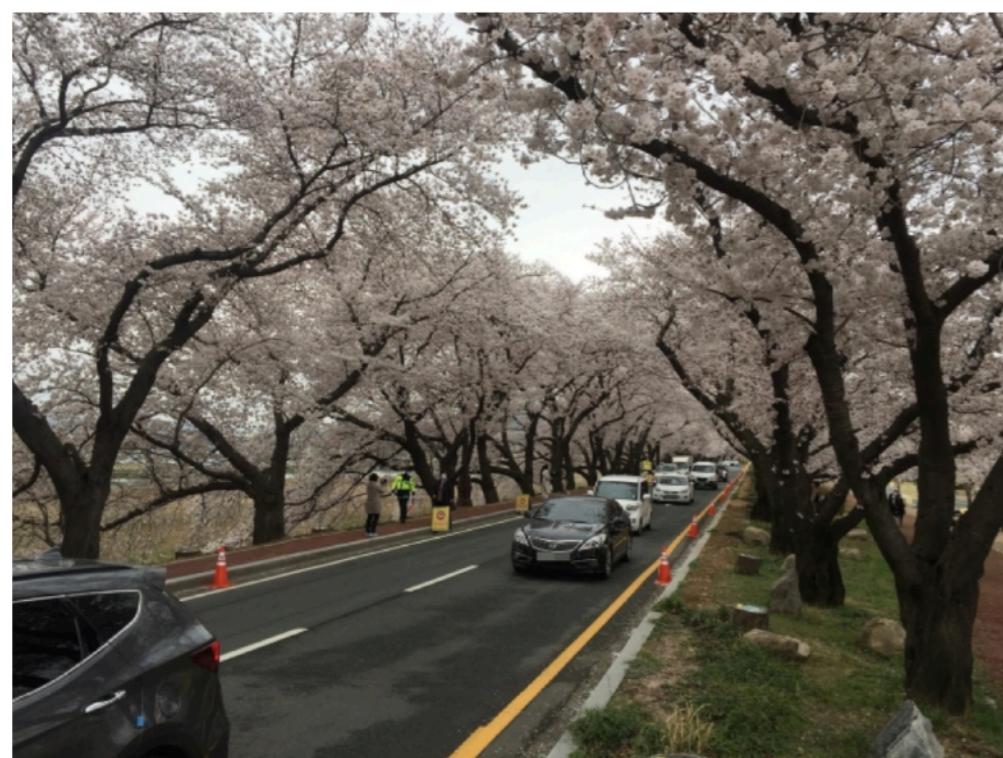
기사입력 2020.03.28. 오후 8:30 기사원문 스크랩 본문듣기 설정

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 경주시 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	통제 올해 대비 확산 관광객 발생 모두 방문 봉쇄 방지 배치 분주 불법 사망 사정 명소 흥무
문장 07	벚꽃, 상춘객, 대비, 벚꽃, 터널, 도로, 양방향, 간격, 경찰관, 배치, 불법, 주정차, 단속	간격 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 08	사진	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	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	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	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	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	0	1	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	

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/29>

** 워드클라운드 kr 열정 긍정적 - 워드클라운드 2017.11.5. <http://wordcloud.kr/1295>

*** 퀘드클라우드.kr, 월정 긍정식 - 퀘드클라우드, 2017.11.5., <http://wordcloud.kr/1295>.

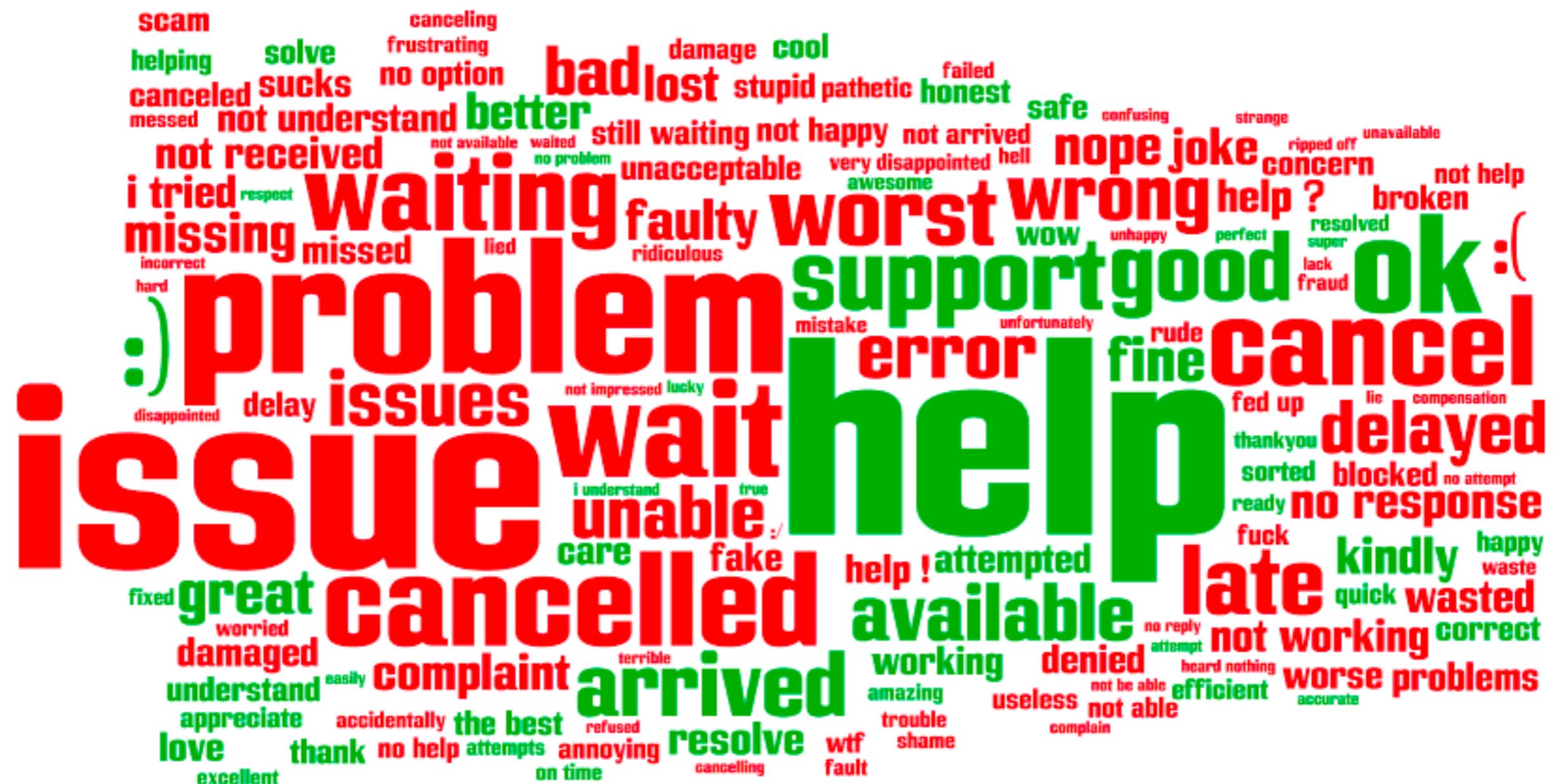
**** Kumo - Java Word Cloud, <http://kennycason.com/posts/2014-07-03-kumo-wordcloud.html/>.

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

***** CX DATA SCIENCE, SIMPLE SENTIMENT 2 SUPPORT, https://www.cxdatascience.com/ssvz_support
***** 니이바블로그(기지그드시), WordCloud(워드클라우드) 만들어 주는 사이트 사용 방법 안내 - taxedo_2016.2.3. <https://blog.naver.com/liberty264/220616866103>

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%BO%EC%9D%B4%ED%84%BO-%EB%B6%84%EC%84%9D-be3f6c557ca2/>

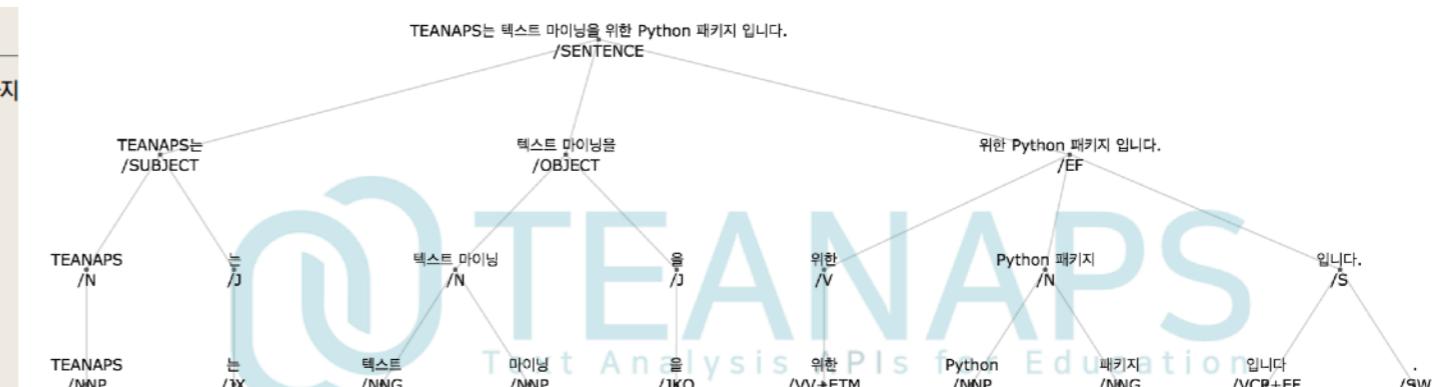
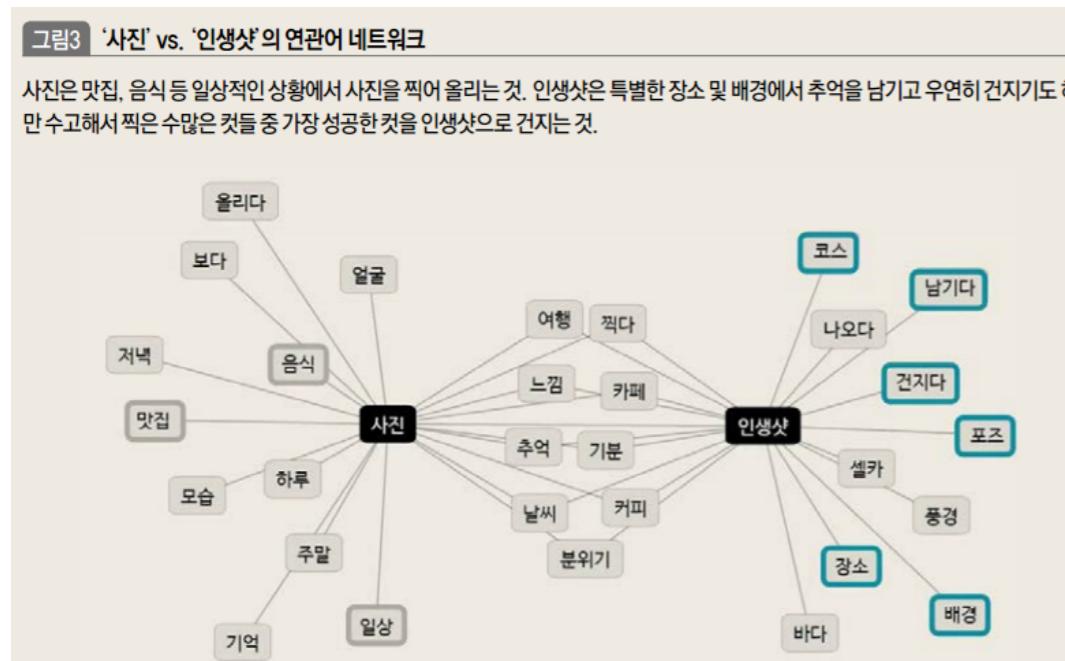
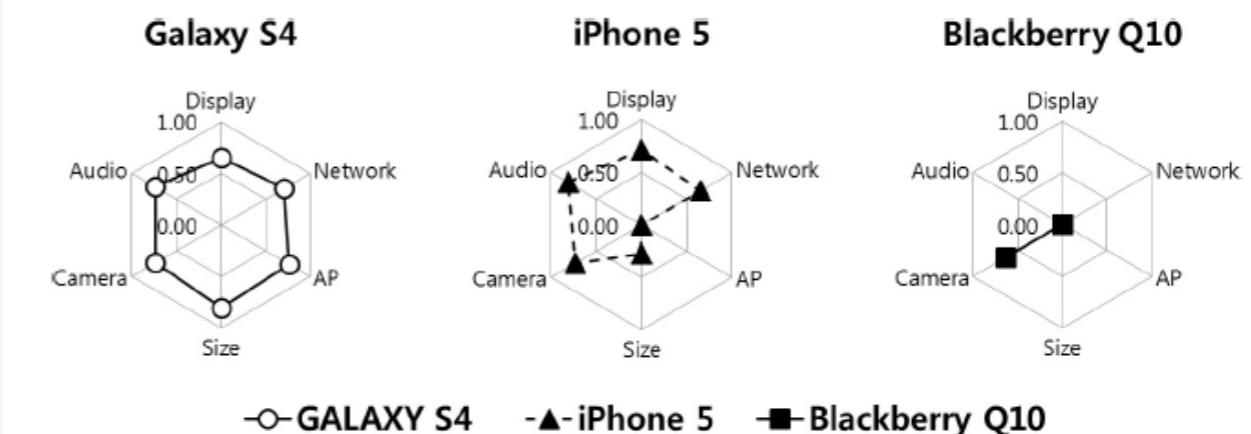
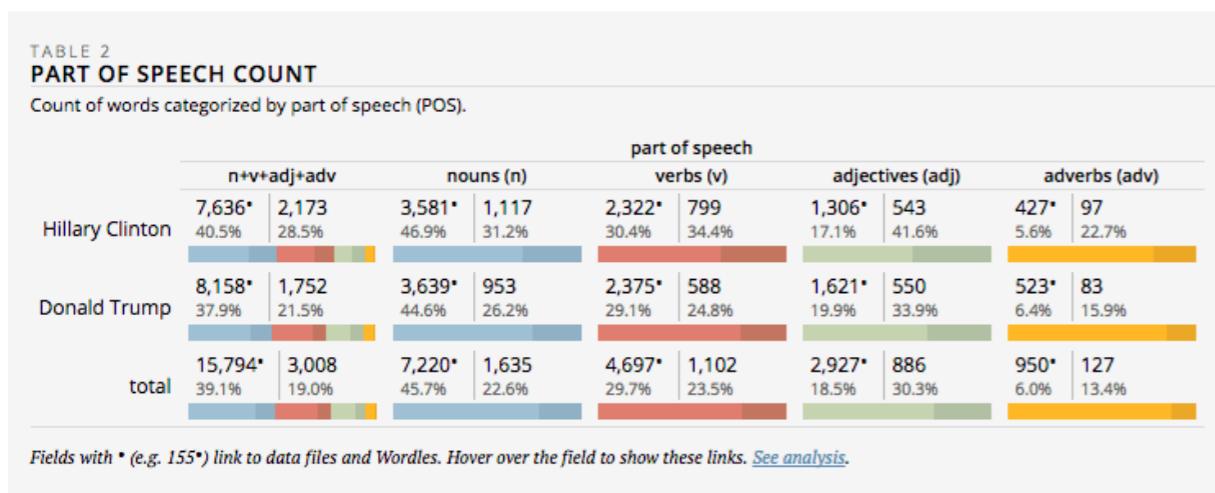
** 동아비즈니스리뷰 262호, 전형적인 워드클라우드, 2018.11., https://dbr.donga.com/graphic/view/gdbr_no/699

*** 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



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

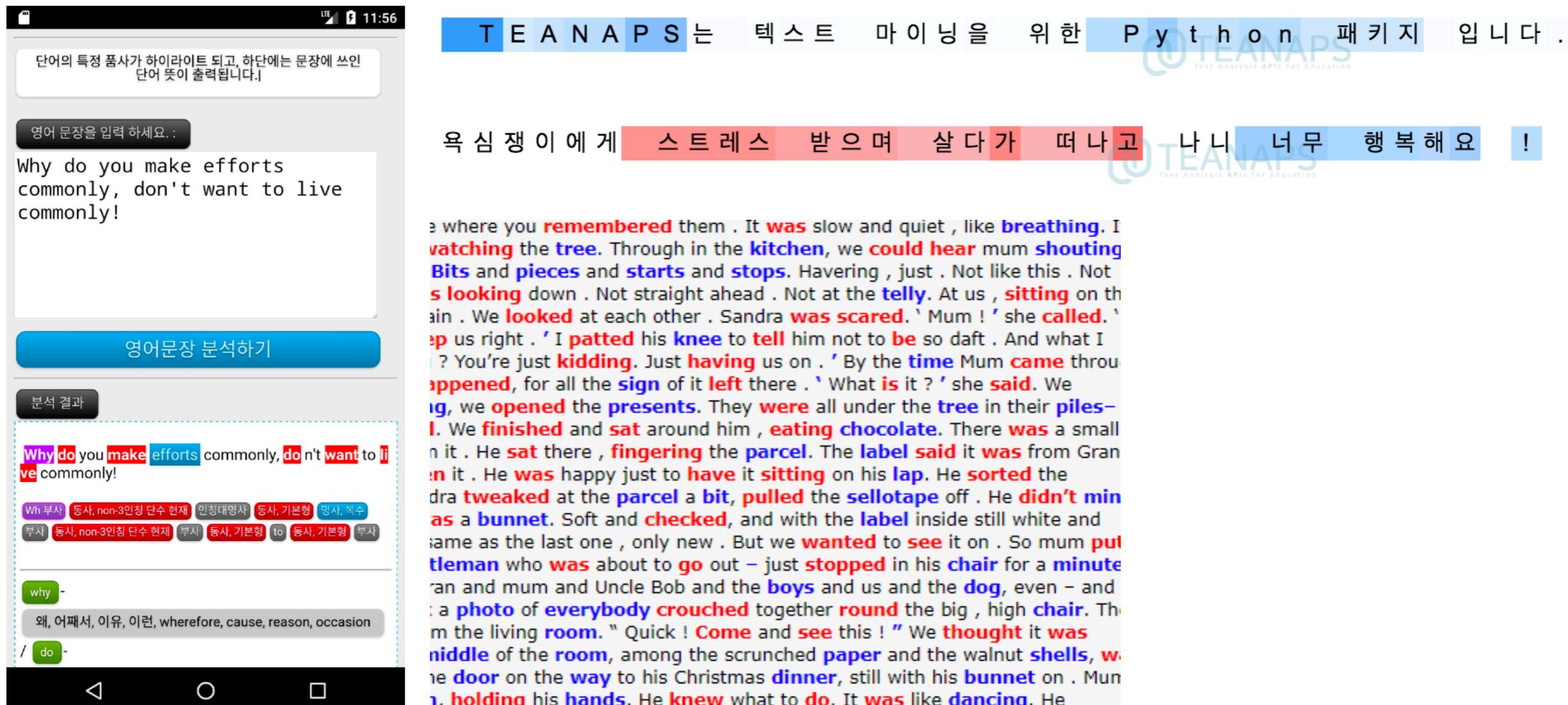
** 최홍규(슬로우뉴스), 2016 미국 대선을 보여주는 텍스트 마이닝 분석방법들, 2017.1.9., <http://slownews.kr/60919>.

*** 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).

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