MIRAE ASSET

미래에셋대우 빅데이터 페스티벌

키워드 추출 / 별난것들

MIRAE ASSET

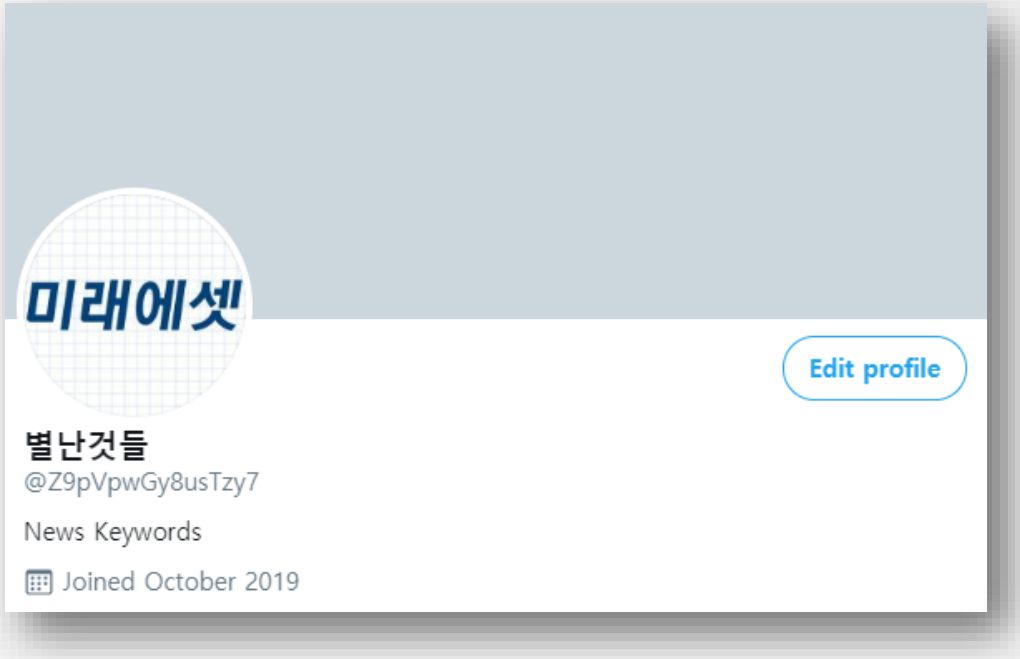
- 트위터 운영, 별난것들
- 필 요 성 및 목 적
- 정확한 추출, LDA 알고리즘
- 알고리즘 아키텍처 구조도
- 키 워 드 추 출 결 과
- 프 로 토 타 입

MIRAE ASSET

1

트위터 운영, 별난것들

트위터 운영, 별난것들 별난것들 소개



■ 트위터 운영, 별난것들

- 저희는 트위터를 운영하고 있는 **별난것들** 팀입니다.

■ 뉴스 키워드 추출 및 제공

- 저희는 해외 뉴스의 키워드를 '**정확히**' 추출합니다.
- 저희는 뉴스의 키워드를 **자동으로** 업로드 합니다.

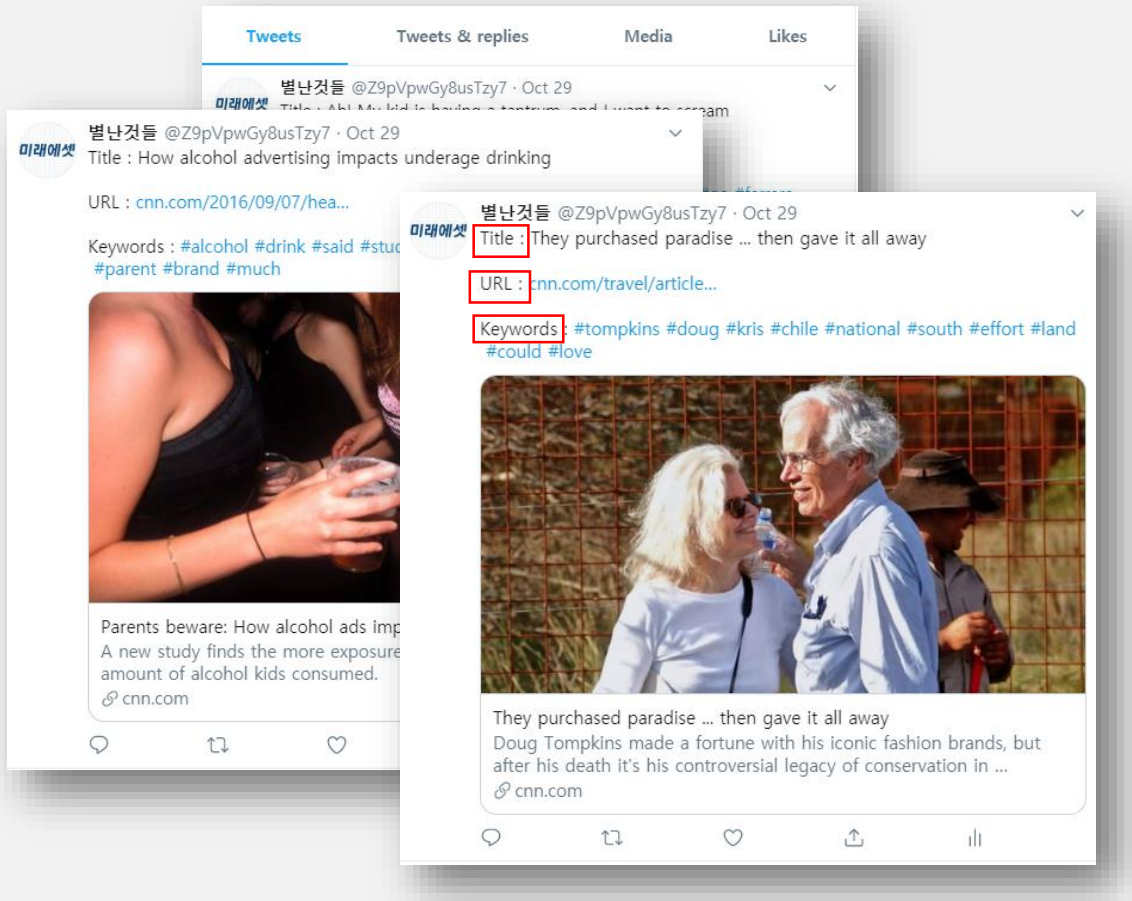
■ 별난것들@Z9pVpwGy8usTzy7

- 저희가 제공하는 컨텐츠를 설명 해 드리기 앞서, **FOLLOW** 부탁드리도록 하겠습니다.
https://twitter.com/Z9pVpwGy8usTzy7

트위터 운영, 별난것들 콘텐츠 예시

트위터 별난것들에서 자동으로 제공하는 컨텐츠

트위터 운영, 별난것들 콘텐츠 소개



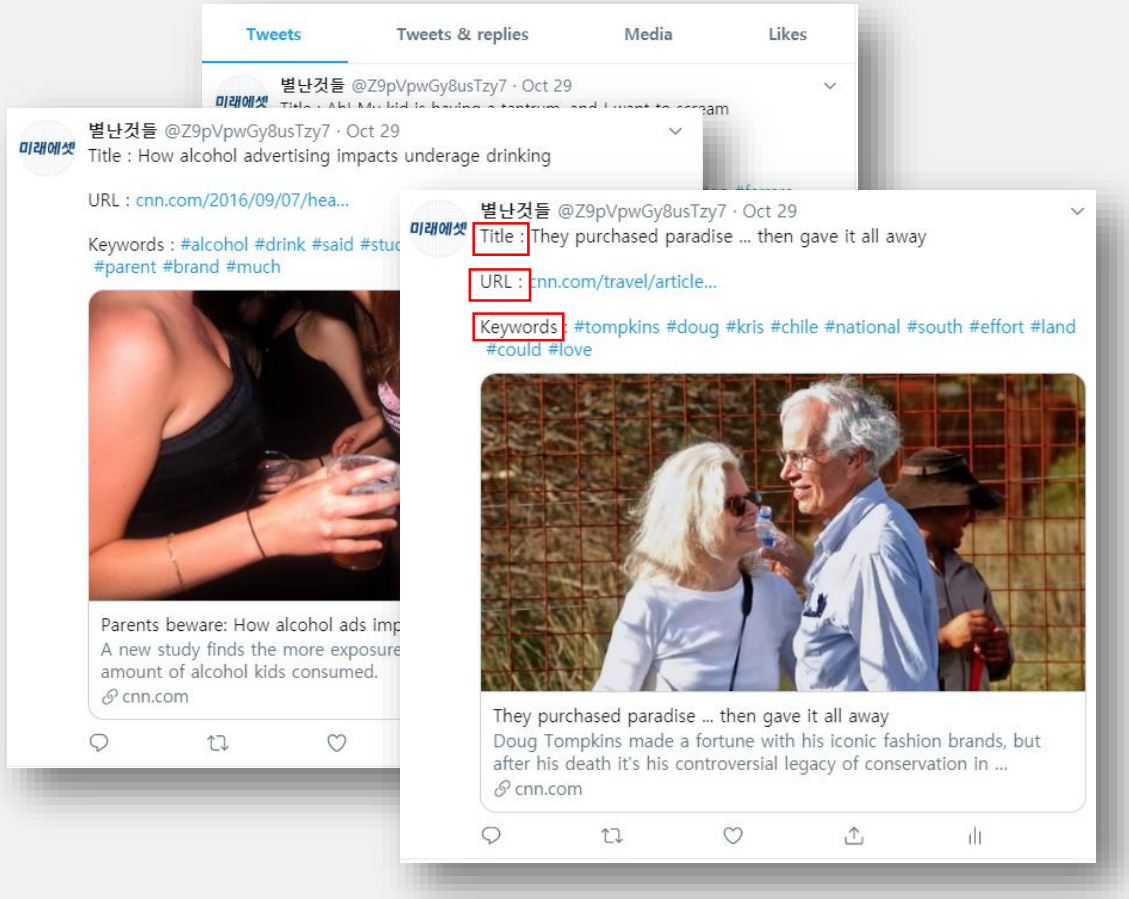
- Title

  - 저희는 글로벌 뉴스의 'Title'인 뉴스 제목을 제공해 줍니다.

- URL

  - 저희는 글로벌 뉴스의 'URL'을 제공합니다.
  - 키워드를 보고 뉴스가 궁금한 독자가 들어가 볼 수 있도록 도와줍니다.

트위터 운영, 별난것들 콘텐츠 소개



Keywords

- 저희는 글로벌 뉴스의 키워드 10개를 제공합니다.
- 저희는 뉴스의 정확한 키워드를 제공합니다.
- 저희는 해쉬태그를 통해 해당 키워드의 연관 게시물의 접근성을 용이하게 해줍니다.

자동 업로드

- 저희는 트위터 API를 이용해, 자동으로 업로드를 합니다.

MIRAE ASSET

트위터 운영, 별난것들 별난것들 홍보

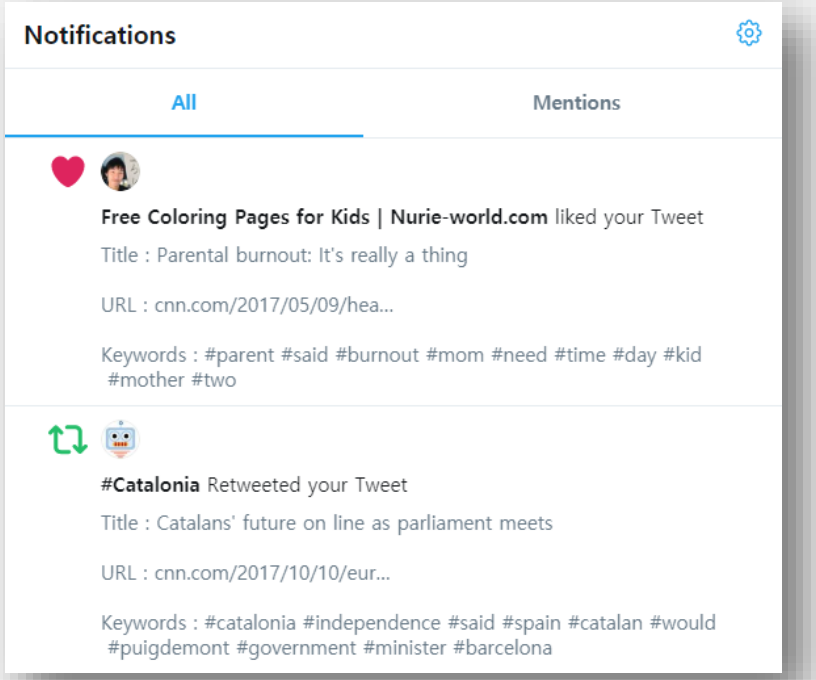
지금, 트위터 별난것들을 Follow 하고

무료로 뉴스의 키워드를 빠르게 제공 받으세요

#별난것들 #뉴스 #키워드 #정확성 #자동 #무료 #미래에셋대우 #빅데이터

MIRAE ASSET

트위터 운영, 별난것들 뜨거운 반응



글로벌 독자들과 국내외 독자들의 뜨거운 반응!

을 기다리고 있습니다...

MIRAE ASSET

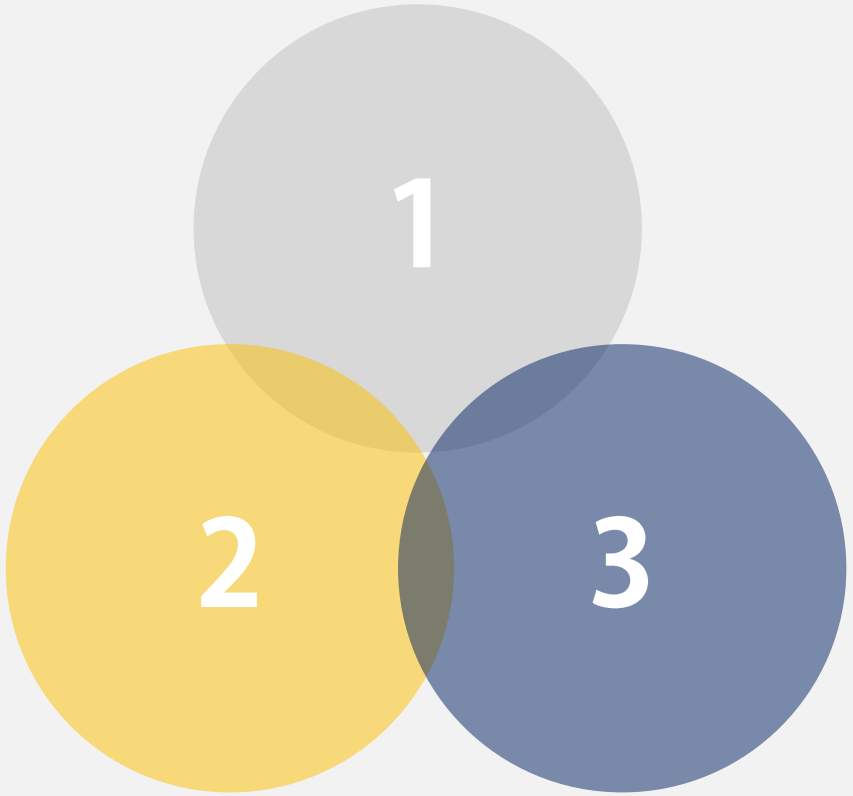
2

필요성 및 목적

MIRAE ASSET

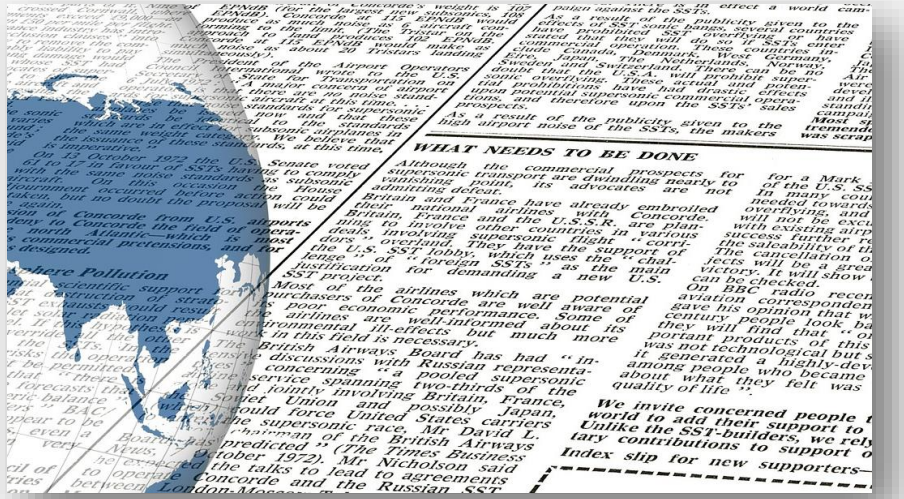
필요성 및 목적 별난것들이 필요한 이유

- 뉴스와 금융
- 뉴스와 빅데이터
- 신속성, 자동성, 정확성



필요성 및 목적

뉴스와 금융



- 금융권의 흐름은 경제, 사회, 문화, 정치의 상황을 반영합니다.
- 금융권의 흐름을 파악하기 위해 사회, 문화, 정치, IT, 스포츠 등 다양한 뉴스를 읽으며 경제 상황과 기업의 현황을 파악하는 것이 좋습니다.

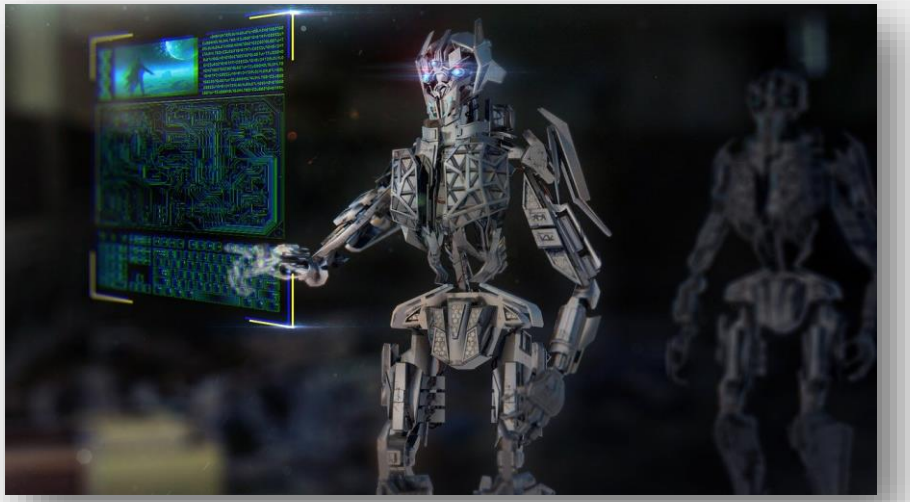
필요성 및 목적

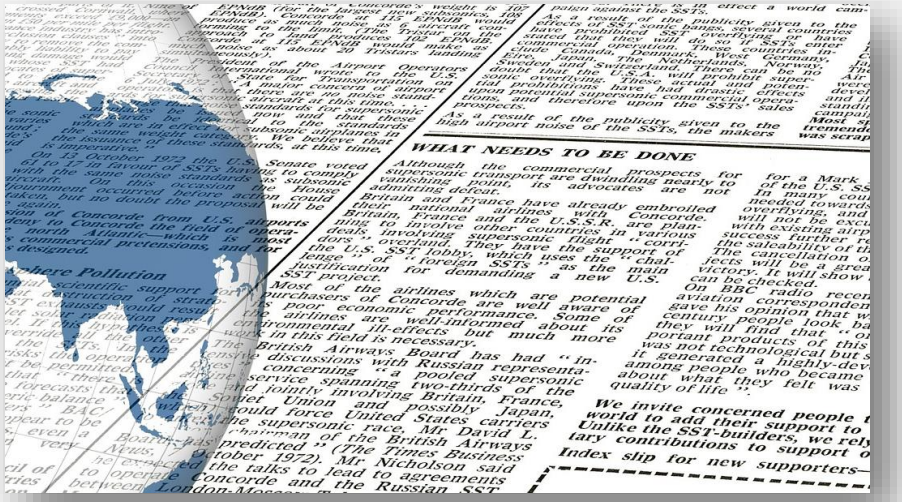
뉴스와 빅데이터

- 2016년 4월 기준 축적된 뉴스량은 3천만건 입니다. 데이터는 하루 평균 1만 5천건이 증가합니다.
- 하루에도 수 만 개씩 쏟아지는 뉴스 기사를 '키워드 ' 를 통해 요약해 줘서, 독자가 키워드를 통해 뉴스를 선택할 수 있도록 도와야 합니다.

필요성 및 목적

신속성, 자동성, 정확성





- 트위터 운영을 통해 ‘뉴스 제목(Title), URL, 키워드’를 독자들에게 신속하고 정확하게 제공합니다.
- 트위터 운영을 통해 자동으로 콘텐츠를 제공해서, 독자가 다양한 뉴스를 요약 제공 받을 수 있도록 도와줍니다.

MIRAE ASSET

필요성 및 목적 알고리즘 개발의 기본 방향

정확도

자동성

- 저희는 키워드 추출의 정확도를 올리기 위해 노력했습니다.
- 저희는 트위터로 뉴스와 키워드를 자동 제공할 수 있도록 노력했습니다.

MIRAE ASSET

3

정확한 추출, LDA 알고리즘

MIRAE ASSET

정확한 추출, LDA 알고리즘 LDA 알고리즘이란?

LDA 알고리즘

- LDA는 **토픽 모델링 알고리즘**입니다.
- LDA는 **주어진 문서에 어떤 주제**가 있는지 찾아주는 모델입니다.
- LDA는 **토픽별 단어의 분포**와 **문서별 토픽의 분포**를 추정해 냅니다.
- LDA는 토픽에서 단어가 나타날 **확률**을 알려줍니다.
- 저희는 뉴스의 키워드를 가장 잘 찾는 토픽 모델링 알고리즘을 찾기 위해 **LDA, LSI, HDP의 알고리즘**을 실험했습니다. 그 중 **LDA 알고리즘**이 가장 정확했습니다.

MIRAE ASSET

정확한 추출, LDA 알고리즘 LDA 알고리즘이란?

LDA 알고리즘 수식

$$p(z_{d,i} = j|z_{-i}, w) = \frac{n_{d,k} + \alpha_j}{\sum_{i=1}^{K}(n_{d,i} + \alpha_i)} \times \frac{v_{k,w_{d,n}} + \beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j} + \beta_j)} = AB$$

- nd,k : k번째 토픽에 할당된 d번째 문서의 단어 빈도
- vk,wd,n : 전체 말뭉치에서 k번째 토픽에 할당된 단어 wd,n의 빈도
- wd,n : d번째 문서에 n번째로 등장한 단어
- α : 문서의 토픽 분포 생성을 위한 디리클레 분포 파라메터
- β : 토픽의 단어 분포 생성을 위한 디리클레 분포 파라메터
- K : 사용자가 지정하는 토픽 수
- V : 말뭉치에 등장하는 전체 단어 수
- A : d번째 문서가 k번째 토픽과 맺고 있는 연관성 정도
- B : d번째 문서의 n번째 단어(wd,n)가 k번째 토픽과 맺고 있는 연관성 정도

MIRAE ASSET

정확한 추출, LDA 알고리즘 정확도를 높이기 위한 노력

데이터

```
In [4]: data['ALL_TEXT'] = data['TITLE'] + ' ' + data['TEXT'] # 제목과 Text 모두 이용
        data
```

- 키워드를 정확히 추출하기 위해, **타이틀**이 중요하다고 판단하였습니다.
- 키워드를 정확히 추출하기 위해 데이터로 **Title과 Text를 모두 이용**했습니다.

MIRAE ASSET

정확한 추출, LDA 알고리즘 정확도를 높이기 위한 노력

이상 데이터 처리

```
In [5]: def data_token(x):
            alltext = data.loc[x:x, 'ALL_TEXT'].tolist() * 5 # 리스트로 만들기
            # alltext를 5배, 10배.. 변경시켜도 output에 변화가 있다.
            # alltext가 키워드 추출하기에 너무 적은 단어 수로 구성되어 있을 경우, 5배를 해 주는게 효과적이다.
            alltext = [[wordnet_lemmatizer.lemmatize(z) for z in tokenizer.tokenize(str(t).lower()) if z not in STOPWORDS] for t in alltext] # 불용
            return alltext

        alltext = data_token(0)
        print(alltext)
```

- 뉴스 기사가 **굉장히 짧은 경우** 키워드 추출이 불가능 했습니다.
- 해당 데이터를 **5배 증식**시켜 키워드를 추출할 수 있었고, 정확도도 더 높아졌습니다.

MIRAE ASSET

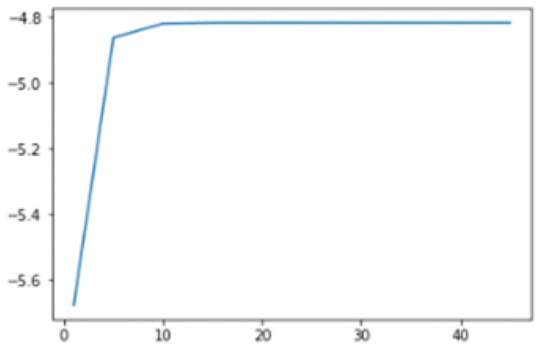
정확한 추출, LDA 알고리즘 정확도를 높이기 위한 노력

LDA 파라미터 튜닝

```
In [10]: x = find_pass
         y = find_perplexity
         plt.plot(x, y)

         # 결론 : perplexity는 낮을 수록 좋으므로, pass를 1로 설정하자.

Out[10]: [<matplotlib.lines.Line2D at 0x1bb74c70f98>]
```



- LDA 파라미터 중 pass의 최적의 파라미터를 찾았습니다.
- LDA 파라미터인 pass의 값에 따른 Perplexity를 확인 해 본 결과 최적의 값인 1을 찾았습니다.

MIRAE ASSET

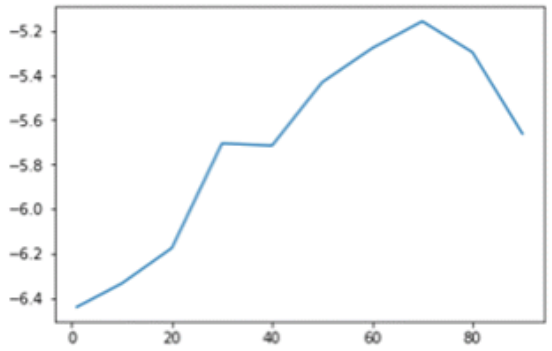
정확한 추출, LDA 알고리즘 정확도를 높이기 위한 노력

LDA 파라미터 튜닝

```
In [12]: x = find_Iteration
         y = find_perplexity
         plt.plot(x, y)

         # 결론 : perplexity는 낮을 수록 좋으므로, Iteration을 1로 설정하자.

Out[12]: [<matplotlib.lines.Line2D at 0x1bb74cd8ef0>]
```



- LDA 파라미터 중 Iteration의 최적의 파라미터를 찾았습니다.
- LDA 파라미터인 Iteration의 값에 따른 Perplexity를 확인 해 본 결과 최적의 값인 1을 찾았습니다.

MIRAE ASSET

정확한 추출, LDA 알고리즘 최종 모델 성능은?

모델 성능

- LDA 모델 평가 결과 -6.437의 Perplexity를 얻을 수 있었습니다.
- LDA 모델 평가 결과 0.477의 Coherence를 얻을 수 있었습니다.

MIRAE ASSET

정확한 추출, LDA 알고리즘 평가 척도 두가지

평가 척도

10. LDA 모델 평가

```
In [23]: lda.log_perplexity(corpus) # Perplexity는 낮을 수록 좋다.
         2019-10-29 00:47:09,461 : INFO : -6.437 per-word bound, 86.7 perplexity estimate based on a held-out corpus of 5 documents with 845 words
Out[23]: -6.437271865020842

In [24]: coherence_model_lda = CoherenceModel(model=lda, texts=alltext, dictionary=dictionary) # Coherence Model는 높을 수록 좋다.

In [25]: coherence_lda = coherence_model_lda.get_coherence()
         2019-10-29 00:47:09,725 : INFO : using ParallelWordOccurrenceAccumulator(processes=3, batch_size=64) to estimate probabilities from sliding windows
         2019-10-29 00:47:10,546 : INFO : 1 batches submitted to accumulate stats from 64 documents (300 virtual)
         2019-10-29 00:47:13,473 : INFO : 3 accumulators retrieved from output queue
         2019-10-29 00:47:13,489 : INFO : accumulated word occurrence stats for 300 virtual documents

In [26]: print(coherence_lda)
         0.4777908195146071
```

- Perplexity는 혼란도를 의미합니다. 토픽모델링의 확률 모델을 평가하는 척도 중 하나로, 낮을 수록 좋습니다.
- Coherence는 주제 일관성을 의미합니다. 단어 유사도와 토픽의 유의미를 평가합니다. 높을 수록 좋습니다.

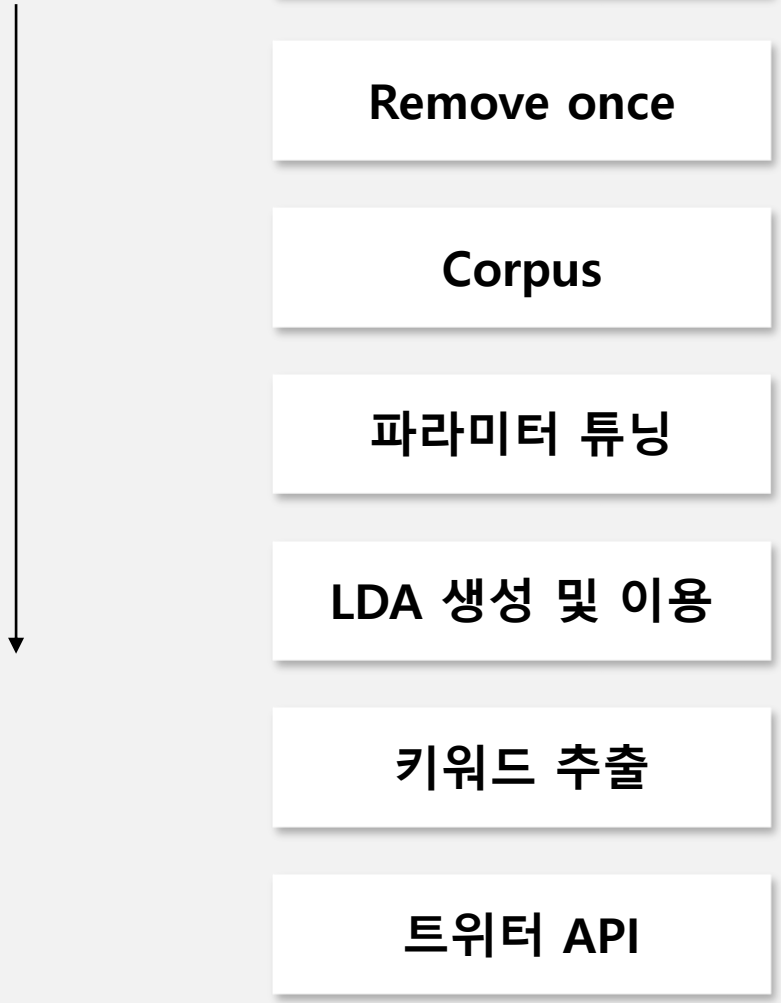
MIRAE ASSET

4

알고리즘 아키텍처 구조도

MIRAE ASSET

알고리즘 아키텍처 구조도



MIRAE ASSET

5

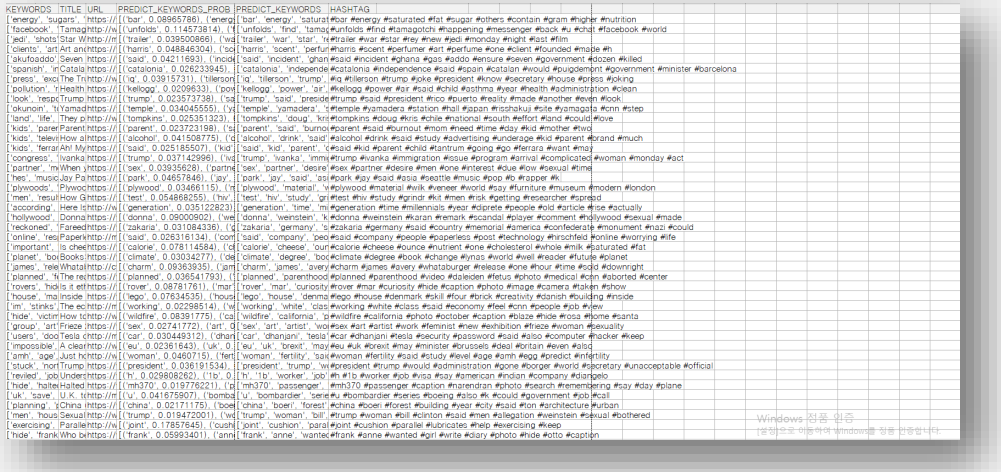
키워드 추출 결과

MIRAE ASSET

키워드 추출 결과 CSV 파일

KEYWORDS	TITLE	URL	PREDICT_KEYWORDS_PROB	PREDICT_KEYWORDS	HASHTAG
['energy', 'sugars', '		https://	[('bar', 0.08965786), ('energ	['bar', 'energy', 'saturat	#bar #energy #saturated #fat #sugar #others #contain #gram #higher #nutrition
['facebook', '	Tamag	http://w	[('unfolds', 0.114573814), ('	['unfolds', 'find', 'tamag	#unfolds #find #tamagotchi #happening #messenger #back #u #chat #facebook #world
['jedi', 'shots	Star W	http://m	[('trailer', 0.039500866), ('wa	['trailer', 'war', 'star', 're	#trailer #war #star #rey #new #jedi #monday #night #last #film
['clients', 'art	Art and	https://	[('harris', 0.048846304), ('sc	['harris', 'scent', 'perfun	#harris #scent #perfumer #art #perfume #one #client #founded #made #h
['akufoaddo'	Seven	https://	[('said', 0.04211693), ('incide	['said', 'incident', 'ghan	#said #incident #ghana #gas #addo #ensure #seven #government #dozen #killed
['spanish', 'in	Catala	https://	[('catalonia', 0.026233945), '	['catalonia', 'independe	#catalonia #independence #said #spain #catalan #would #puigdemont #government #minister #barcelona
['press', 'exc	The Tr	http://w	[('iq', 0.03915731), ('tillerson	['iq', 'tillerson', 'trump',	#iq #tillerson #trump #joke #president #know #secretary #house #press #joking
['pollution', 'r	Health	https://	[('kellogg', 0.0209633), ('pov	['kellogg', 'power', 'air',	#kellogg #power #air #said #child #asthma #year #health #administration #clean
['look', 'respo	Trump	https://	[('trump', 0.023573738), ('sa	['trump', 'said', 'preside	#trump #said #president #rico #puerto #reality #made #another #even #look
['okunoin', 't	Yamad	https://	[('temple', 0.034045555), ('ya	['temple', 'yamadera', 's	#temple #yamadera #station #hall #japan #risshakuji #site #yamagata #cnn #step
['land', 'life',	They p	http://w	[('tompkins', 0.025351323), '	['tompkins', 'doug', 'kris	#tompkins #doug #kris #chile #national #south #effort #land #could #love
['kids', 'parer	Parent	https://	[('parent', 0.023723198), ('sa	['parent', 'said', 'burnou	#parent #said #burnout #mom #need #time #day #kid #mother #two
['kids', 'televi	How a	https://	[('alcohol', 0.041508775), ('c	['alcohol', 'drink', 'said'	#alcohol #drink #said #study #advertising #underage #kid #parent #brand #much
['kids', 'ferrar	Ah! My	https://	[('said', 0.025185507), ('kid'	['said', 'kid', 'parent', 'c	#said #kid #parent #child #tantrum #going #go #ferrara #want #may
['congress', '	Ivanka	https://	[('trump', 0.037142996), ('iva	['trump', 'ivanka', 'immi	#trump #ivanka #immigration #issue #program #arrival #complicated #woman #monday #act
['partner', 'm	When y	https://	[('sex', 0.03935628), ('partne	['sex', 'partner', 'desire'	#sex #partner #desire #men #one #interest #due #low #sexual #time
['hes', 'music	Jay Pa	https://	[('park', 0.04657846), ('jay',	['park', 'jay', 'said', 'asi	#park #jay #said #asia #seattle #music #pop #b #rapper #k
['plywoods', '	Plywoo	https://	[('plywood', 0.03466115), ('n	['plywood', 'material', 'w	#plywood #material #wilk #veneer #world #say #furniture #museum #modern #london
['men', 'result	How G	https://	[('test', 0.054868255), ('hiv',	['test', 'hiv', 'study', 'gri	#test #hiv #study #grindr #kit #men #risk #getting #researcher #spread
['according',	Here Is	http://w	[('generation', 0.035122823),	['generation', 'time', 'mi	#generation #time #millennials #year #diprete #people #old #article #rise #actually
['hollywood',	Donna	https://	[('donna', 0.09000902), ('we	['donna', 'weinstein', 'k	#donna #weinstein #karan #remark #scandal #player #comment #hollywood #sexual #made
['reckoned', '	Fareed	https://	[('zakaria', 0.031084336), ('g	['zakaria', 'germany', 's	#zakaria #germany #said #country #memorial #america #confederate #monument #nazi #could
['online', 'res	Paperl	http://m	[('said', 0.026316134), ('com	['said', 'company', 'peo	#said #company #people #paperless #post #technology #hirschfeld #online #worrying #life
['important',	Is chee	https://	[('calorie', 0.078114584), ('ch	['calorie', 'cheese', 'oun	#calorie #cheese #ounce #nutrient #one #cholesterol #whole #milk #saturated #fat
['planet', 'bo	Books	https://	[('climate', 0.03034277), ('de	['climate', 'degree', 'boo	#climate #degree #book #change #lynas #world #well #reader #future #planet
['james', 'rele	What a	http://c	[('charm', 0.09363935), ('jam	['charm', 'james', 'avery	#charm #james #avery #whataburger #release #one #hour #time #sold #downright
['planned', 'fe	The re	https://	[('planned', 0.036541793), ('	['planned', 'parenthood	#planned #parenthood #video #daleiden #fetus #photo #medical #cnn #aborted #center
['rovers', 'hid	Is it et	https://	[('rover', 0.08781761), ('mar	['rover', 'mar', 'curiosity	#rover #mar #curiosity #hide #caption #photo #image #camera #taken #show
['house', 'ma	Inside	https://	[('lego', 0.07634535), ('hous	['lego', 'house', 'denma	#lego #house #denmark #skill #four #brick #creativity #danish #building #inside
['im', 'stinks',	The ec	http://m	[('working', 0.02298514), ('w	['working', 'white', 'clas	#working #white #class #said #economy #feel #cnn #people #job #view
['hide', 'victim	How to	http://w	[('wildfire', 0.08391775), ('ca	['wildfire', 'california', 'p	#wildfire #california #photo #october #caption #blaze #hide #rosa #home #santa
['group', 'art'	Frieze	https://	[('sex', 0.02741772), ('art', 0	['sex', 'art', 'artist', 'wo	#sex #art #artist #work #feminist #new #exhibition #frieze #woman #sexuality
['users', 'doo	Tesla c	http://m	[('car', 0.030449312), ('dhan	['car', 'dhanjani', 'tesla'	#car #dhanjani #tesla #security #password #said #also #computer #hacker #keep
['impossible',	A clear	http://w	[('eu', 0.02361643), ('uk', 0.	['eu', 'uk', 'brexit', 'may	#eu #uk #brexit #may #minister #brussels #deal #britain #even #also
['amh', 'age',	Just ho	http://w	[('woman', 0.0460715), ('fert	['woman', 'fertility', 'sai	#woman #fertility #said #study #level #age #amh #egg #predict #infertility
['stuck', 'nor	Trump	https://	[('president', 0.036191534), '	['president', 'trump', 'wo	#president #trump #would #administration #gone #borger #world #secretary #unacceptable #official
['reviled', 'job	Unders	https://	[('h', 0.029808262), ('1b', 0.	['h', '1b', 'worker', 'job'	#h #1b #worker #job #visa #say #american #indian #company #diangelo
['hide', 'halte	Halted	https://	[('mh370', 0.019776221), ('p	['mh370', 'passenger', '	#mh370 #passenger #caption #narendran #photo #search #remembering #say #day #plane
['uk', 'save',	U.K. to	http://m	[('u', 0.041675907), ('bomba	['u', 'bombardier', 'serie	#u #bombardier #series #boeing #also #k #could #government #job #call
['planning', 'p	China	https://	[('china', 0.02171175), ('boe	['china', 'boeri', 'forest'	#china #boeri #forest #building #year #city #said #ton #architecture #urban
['men', 'hous	Sexual	http://w	[('trump', 0.019472001), ('wo	['trump', 'woman', 'bill',	#trump #woman #bill #clinton #said #men #allegation #weinstein #sexual #bothered
['exercising',	Paralle	http://w	[('joint', 0.17857645), ('cush	['joint', 'cushion', 'paral	#joint #cushion #parallel #lubricates #help #exercising #keep
['hide', 'frank	Who be	https://	[('frank', 0.05993401), ('anna	['frank', 'anne', 'wanted	#frank #anne #wanted #girl #write #diary #photo #hide #otto #caption

키워드 추출 결과 CSV 파일 설명



KEYWORDS	TITLE	URL	PREDICT_KEYWORDS_PROB	PREDICT_KEYWORDS	HASHTAG
['energy', 'sugars',		'https://[	[('bar', 0.08965786), ('energ	['bar', 'energy', 'saturat	#bar #energy #saturated #fat #sugar #others #contain #gram #higher #nutrition
['facebook', 'Tamag		http://w	[('unfolds', 0.11457314), ('f	['unfolds', 'find', 'tama	#unfolds #find #tamagotchi #happening #messenger #back #u #chat #facebook #world
['jedi', 'shots	Star W	http://n	[('trailer', 0.039500866), ('wa	['trailer', 'war', 'star', 'n	#trailer #war #star #ley #new #jedi #monday #night #last #film
['clients', 'art'	Art and	https://[	[('harris', 0.048846304), ('so	['harris', 'scent', 'perfur	#harris #scent #perfumer #art #perfume #one #client #founded #made #h
['akufoaddo'	Seven	https://[	[('said', 0.04211693), ('incid	['said', 'incident', 'ghan	#said #incident #ghana #gas #addo #ensure #seven #government #dozen #killed
['spanish', 'in	Catala	https://[	[('catalonia', 0.026233945), (	['catalonia', 'independe	#catalonia #independence #said #spain #catalan #would #suggestment #government #minister #barcelona
['press', 'exci	The Tri	http://w	[('q', 0.03915731), ('tillerson	['q', 'tillerson', 'trump',	#q #tillerson #trump #joke #president #know #secretary #house #press #joking
['pollution', 'r	Health	https://[	[('kellogg', 0.0209633), ('pov	['kellogg', 'power', 'air',	#kellogg #power #ar #said #child #asthma #year #health #administration #clean
['look', 'respo	Trump	https://[	[('trump', 0.023573738), ('sa	['trump', 'said', 'preside	#trump #said #president #rico #puerto #reality #made #another #even #look
['okunoin', 'N	Yamad	https://[	[('temple', 0.034045555), ('yi	['temple', 'yamadera', 'r	#temple #yamadera #station #hall #japan #rishakuji #site #yamagata #nnn #step
['land', 'life', 'They p		http://w	[('tompkins', 0.025351323), (	['tompkins', 'doug', 'kris	#tompkins #doug #kris #chile #national #south #effort #land #could #love
['kids', 'parer	Parent	https://[	[('parent', 0.023723198), ('sj	['parent', 'said', 'burnou	#parent #said #burnout #mom #need #time #day #kid #mother #two
['kids', 'telev	How al	https://[	[('alcohol', 0.04150877 5), ('q	['alcohol', 'drink', 'said'	#alcohol #drink #said #study #advertising #underage #kid #parent #brand #much
['kids', 'ferrar	Ahl My	https://[	[('said', 0.025185507), ('kid'	['said', 'kid', 'parent', 'c	#said #kid #parent #child #tantrum #going #go #ferrara #want #ma
['congress', '	Ivanka	https://[	[('trump', 0.037142996), ('wd	['trump', 'ivanka', 'imm	#trump #ivanka #immigration #issue #program #arrival #complicated #woman #monday #act
['partner', 'mi	When	https://[	[('sex', 0.03935828), ('partne	['sex', 'partner', 'desire	#sex #partner #desire #men #one #interest #due #lov #sexual #time
['hes', 'music	Jay Pa	https://[	[('park', 0.04657846), ('jay',	['park', 'jay', 'said', 'as	#park #jay #said #asia #seattle #music #pop #b #rapper #k
['plywoods', '	Plywoo	https://[	[('plywood', 0.03466115), ('n	['plywood', 'material', 'v	#plywood #material #wilk #veneer #world #say #furniture #museum #modern #london
['men', 'resul	How G	https://[	[('test', 0.054868255), ('hiv',	['test', 'hiv', 'study', 'gr	#test #hiv #study #grindr #kit #men #risk #getting #researcher #spread
['according', '	Here is	http://w	[('generation', 0.035122823)	['generation', 'time', 'm	#generation #time #millennials #year #dispute #people #old #article #use #actually
['hollywood', '	Donna	https://[	[('donna', 0.09000902), ('we	['donna', 'weinstein', 'k	#donna #weinstein #karan #remark #scandal #player #comment #hollywood #sexual #made
['reckoned', '	Fareed	https://[	[('zakaria', 0.031084336), ('g	['zakaria', 'germany', 'u	#zakaria #germany #said #country #memorial #america #confederate #monument #nazi #could
['online', 'res	Paper	http://n	[('said', 0.026316134), ('com	['said', 'company', 'peo	#said #company #people #paperless #post #technology #hirschfeld #online #worrying #life
['important', '	Is chee	https://[	[('calorie', 0.078114584), ('d	['calorie', 'cheese', 'our	#calorie #cheese #ounce #nutrient #one #cholesterol #whole #milk #saturated #fat
['planet', 'bo	Books	https://[	[('climate', 0.03034277), ('de	['climate', 'degree', 'boo	#climate #degree #book #change #jynas #world #well #reader #future #planet
['james', 'rele	What a	http://c	[('charm', 0.09636935), ('jam	['charm', 'james', 'avery	#charm #james #avery #whataburger #release #one #hour #time #sold #downright
['planned', 'N	The re	https://[	[('planned', 0.036541793), ('[	['planned', 'parenthood	#planned #parenthood #video #daleiden #fetus #photo #medical #cnn #aborted #center
['rovers', 'hid	Is it et	https://[	[('rover', 0.08781761), ('mar	['rover', 'mar', 'curiosity	#rover #mar #curiosity #hide #caption #photo #image #camera #taken #show
['house', 'ma	Inside	https://[	[('lego', 0.07634535), ('hous	['lego', 'house', 'denma	#lego #house #denmark #skill #four #brick #creativity #danish #building #inside
['im', 'stinks'	The ec	http://m	[('working', 0.02298514), ('w	['working', 'white', 'class	#working #white #class #said #economy #feel #cnn #people #job #jew
['hide', 'victim	How to	http://w	[('wildfire', 0.08391775), ('ca	['wildfire', 'california', 'p	#wildfire #california #photo #october #caption #blaze #hide #rosa #home #santa
['group', 'art'	Frieze	https://[	[('sex', 0.02741772), ('art', 'g	['sex', 'art', 'artist', 'wo	#sex #art #artist #work #feminist #new #exhibition #frieze #woman #sexuality
['users', 'doo	Tesla c	http://w	[('car', 0.03044931 2), ('dhan	['car', 'dhanjani', 'tesla'	#car #dhanjani #tesla #security #password #said #also #computer #hacker #keep
['impossible', '	A clear	http://w	[('eu', 0.02361643), ('uk', 0.0	['eu', 'uk', 'brexit', 'may	#eu #uk #brexit #may #minister #brussels #deal #britain #even #also
['amh', 'age',	Just h	http://w	[('woman', 0.0460715), ('fert	['woman', 'fertility', 'sai	#woman #fertility #said #study #level #age #amh #egg #predict #infertility
['stuck', 'nort	Trump	https://[	[('president', 0.036191534), (	['president', 'trump', 'wi	#president #trump #would #administration #gone #border #world #secretary #unacceptable #official
['revived', 'job	Under	https://[	[('h', 0.029808262), ('1b', 0.0	['h', '1b', 'worker', 'job'	#h #1b #worker #job #visa #say #american #indian #company #disrupto
['hide', 'halte	Halted	https://[	[('mh370', 0.01977621), ('p	['mh370', 'passenger', '	#mh370 #passenger #caption #narendran #photo #search #remembering #say #day #plane
['uk', 'save', 'U.K. t		http://m	[('u', 0.041675907), ('bomba	['u', 'bombardier', 'serie	#u #bombardier #series #boeing #also #k #could #government #job #call
['planning', 'j	China	https://[	[('china', 0.02171175), ('boei	['china', 'boeri', 'forest',	#china #boeri #forest #building #year #city #said #tan #architecture #urban
['men', 'hous	Sexual	http://w	[('trump', 0.019472001), ('wd	['trump', 'woman', 'bill',	#trump #woman #bill #clinton #said #men #allegation #weinstein #sexual #bothered
['exercising', '	Parall	http://w	[('joint', 0.17857645), ('cush	['joint', 'cushion', 'paral	#joint #cushion #parallel #lubricates #help #exercising #keep
['hide', 'frank	Who b	https://[	[('frank', 0.05993401), ('anne	['frank', 'anne', 'wantec	#frank #anne #wanted #girl #write #diary #photo #hide #otto #caption

- PREDICT_KEYWORDS_PROB

  - 모델을 통해 키워드를 추출하고, 키워드가 토픽일 확률을 알려줍니다.

- PREDICT_KEYWORDS

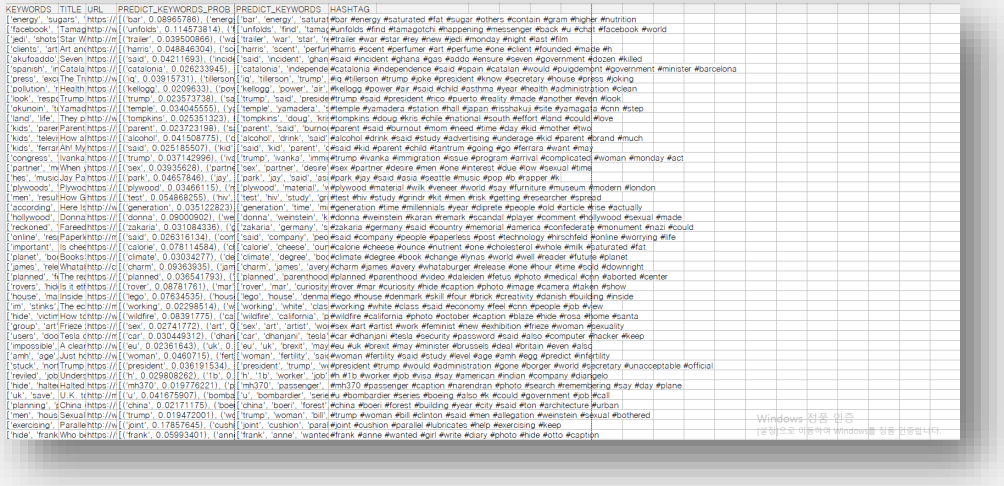
  - 모델을 통해 예측한 키워드입니다.

- HASHTAG

  - 모델을 통해 예측한 키워드에 HASHTAG를 붙여 트위터에 자동 업로드 할 수 있도록 했습니다.

MIRAE ASSET

키워드 추출 결과 키워드 추출 결과 예시



KEYWORDS	TITLE	URL	PREDICT_KEYWORDS_PROB	PREDICT_KEYWORDS	HASHTAG
['energy', 'sugars',		https://[[('bar', 0.08965786), ('energ	['bar', 'energy', 'saturat	#bar #energy #saturated #fat #sugar #others #contain #gram #higher #nutrition	
['facebook',	'Tamag	http://w	[('unfolds', 0.11457381), ('f	['unfolds', 'find', 'tamag	#unfolds #find #tamagotchi #happening #messenger #back #u #chat #facebook #world
['jedi', 'shots'	'Star W	http://m	[('trailer', 0.039500866), ('wa	['trailer', 'war', 'star', 'r	#trailer #war #star #rey #new #jedi #monday #night #last #film
['clients', 'art'	'Art an	https://[[('harris', 0.048846304), ('soc	['harris', 'scent', 'perfur	#harris #scent #perfumer #art #perfume #one #client #founded #made #h	
['akufoaddo'	'Seven	https://[[('said', 0.0421693), ('incid	['said', 'incident', 'ghan	#said #incident #ghana #gas #addo #ensure #seven #government #dozen #killed	
['spanish', 'm'	'Catala	https://[[('catalonia', 0.026233945),	['catalonia', 'independe	#catalonia #independence #said #spain #catalan #would #puigdemont #government #minister #barcelona	
['press', 'exci	'The Tr	http://w	[('rg', 0.03915731), ('tillerson	['rg', 'tillerson', 'trump',	#g #tillerson #trump #joke #president #know #secretary #house #press #joking
['pollution', 'r	'Health	https://[[('kellogg', 0.0209633), ('pow	['kellogg', 'power', 'air',	#kellogg #power #air #said #child #asthma #year #health #administration #clean	
['look', 'respc	'Trump	https://[[('trump', 0.023573738), ('sa	['trump', 'said', 'preside	#trump #said #president #rico #puerto #reality #made #another #even #look	
['okunoin', 'ti	'Yamad	https://[[('temple', 0.034045555), ('ya	['temple', 'yamadera', 't	#temple #yamadera #station #hall #japan #rishakuji #site #yamagata #cnn #step	
['land', 'life',	'They p	http://w	[('tompkins', 0.025351323), ([	['tompkins', 'doug', 'kris	#tompkins #doug #kris #chile #national #south #effort #land #could #love
['kids', 'parer	'Parent	https://[[('parent', 0.023723198), ('sa	['parent', 'said', 'burnou	#parent #said #burnout #mom #need #time #day #kid #mother #and	
['kids', 'televi	'How a	https://[[('alcohol', 0.041508775), ('d	['alcohol', 'drink', 'said'	#alcohol #drink #said #study #advertising #underage #kid #parent #brand #much	
['kids', 'ferrar	'Ah! M	https://[[('said', 0.025185507), ('kid'	['said', 'kid', 'parent', 'c	#said #kid #parent #child #tantrum #going #go #ferrara #want #may	
['congress',	'Ivanka	https://[[('trump', 0.037142996), ('iv	['trump', 'ivanka', 'immi	#trump #vanka #immigration #issue #program #arrival #complicated #woman #monday #act	
['partner', 'my	'When	https://[[('sex', 0.03935628), ('partne	['sex', 'partner', 'desire'	#sex #partner #desire #men #one #interest #due #low #sexual #time	
['hes', 'music	'Jay Pa	https://[[('park', 0.04657846), ('jay',	['park', 'jay', 'said', 'asi	#park #jay #said #asia #seattle #music #pop #b #rapper #k	
['plywoods', '	'Plywo	https://[[('plywood', 0.03466115), ('pl	['plywood', 'material', 'w	#plywood #material #wilk #veneer #world #say #furniture #museum #modern #london	
['men', 'resul	'How G	https://[[('test', 0.054868255), ('hiv',	['test', 'hiv', 'study', 'gri	#test #hiv #study #grindr #kit #men #risk #getting #researcher #spread	
['according',	'Here is	http://w	[('generation', 0.035122823),	['generation', 'time', 'mi	#generation #time #millennials #year #diprete #people #old #article #rise #actually
['hollywood',	'Donna	https://[[('donna', 0.09000902), ('we	['donna', 'weinstein', 'k	#donna #weinstein #karan #remark #scandal #player #comment #hollywood #sexual #made	
['reckoned',	'Fareed	https://[[('zakaria', 0.031084236), ('g	['zakaria', 'germany', 's	#zakaria #germany #said #country #memorial #america #confederate #monument #nazi #could	
['online', 'req	'Paper	http://m	[('said', 0.026316134), ('cont	['said', 'company', 'peo	#said #company #people #paperless #post #technology #hirschfeld #online #worrying #life
['important',	'Is che	https://[[('calorie', 0.078114584), ('ch	['calorie', 'cheese', 'our	#calorie #cheese #ounce #nutrient #one #cholesterol #whole #milk #saturated #fat	
['planet', 'bo	'Books	https://[[('climate', 0.03034277), ('de	['climate', 'degree', 'boc	#climate #degree #book #change #ymas #world #well #reader #future #planet	
['james', 'rele	'What a	http://c	[('charm', 0.09363935), ('jam	['charm', 'james', 'avery	#charm #james #avery #whataburger #release #one #hour #time #sold #downright	
['planned', 'N	'The re	https://[[('planned', 0.036541793), ([	['planned', 'parenthood	#planned #parenthood #video #daleiden #fetus #photo #medical #cen #aborted #center	
['rovers', 'hid	'Is it et	https://[[('rover', 0.08781761), ('mar	['rover', 'mar', 'curiosity	#rover #mar #curiosity #hide #caption #photo #image #camera #taken #show	
['house', 'ma	'Inside	https://[[('lego', 0.07634535), ('hous	['lego', 'house', 'denma	#lego #house #denmark #skill #four #brick #creativity #danish #building #inside	
['im', 'stinks'	'The ec	http://m	[('working', 0.02298514), ('wo	['working', 'white', 'class	#working #white #class #said #economy #feel #cnn #people #job #jew
['hide', 'victim	'How t	http://w	[('wildfire', 0.08391775), ('cal	['wildfire', 'california', 'p	#wildfire #california #photo #october #caption #blaze #hide #rosa #home #santa
['group', 'art'	'Frieze	https://[[('sex', 0.02741772), ('art', 0	['sex', 'art', 'artist', 'wo	#sex #art #artist #work #feminist #new #exhibition #frieze #woman #sexuality	
['users', 'doo	'Tesla c	http://m	[('car', 0.03044912), ('dhan	['car', 'dhanjani', 'tesla'	#car #dhanjani #tesla #security #password #said #also #computer #hacker #keep
['impossible',	'A clea	http://w	[('eu', 0.02361643), ('uk', 0.0	['eu', 'uk', 'brexit', 'may	#eu #uk #brexit #may #minister #brussels #deal #britain #even #also
['amh', 'age',	'Just h	http://w	[('woman', 0.0460715), ('fert	['woman', 'fertility', 'sai	#woman #fertility #said #study #level #age #amh #egg #predict #infertility
['stuck', 'nort	'Trump	https://[[('president', 0.036191534), (	['president', 'trump', 'wi	#president #trump #would #administration #gone #borger #world #secretary #unacceptable #official	
['reviled', 'job	'Under	https://[[('h', 0.029808262), ('1b', 0.0	['h', '1b', 'worker', 'job'	#h #1b #worker #job #visa #say #american #indian #company #dianpelo	
['hide', 'halte	'Halted	https://[[('mh370', 0.019776221), ('p	['mh370', 'passenger', '	#mh370 #passenger #caption #narendran #photo #search #remembering #say #day #plane	
['uk', 'save',	'U.K. t	http://m	[('u', 0.041675907), ('bomba	['u', 'bombardier', 'serie	#u #bombardier #series #boeing #also #k #could #government #job #call
['planning', '	'China	https://[[('china', 0.021771175), ('boe	['china', 'boeri', 'forest'	#china #boeri #forest #building #year #city #said #ton #architecture #urban	
['men', 'hous	'Sexual	http://w	[('trump', 0.019472001), ('wo	['trump', 'woman', 'bill'	#trump #woman #bill #clinton #said #men #allegation #weinstein #sexual #bothered
['exercising',	'Parallel	http://w	[('joint', 0.17857645), ('cushi	['joint', 'cushion', 'paral	#joint #cushion #parallel #lubricates #help #exercising #keep
['hide', 'frank	'Who b	https://[[('frank', 0.05993401), ('ann	['frank', 'anne', 'wantec	#frank #anne #wanted #girl #write #diary #photo #hide #otto #caption	

정답 키워드 : ['akufoaddo', 'tanker', 'incidents', 'dozens', 'streetsgovernment', 'work', 'station', 'seven', 'explosion', 'gas', 'ghana', 'nana', 'injured', 'ensure', 'killed']

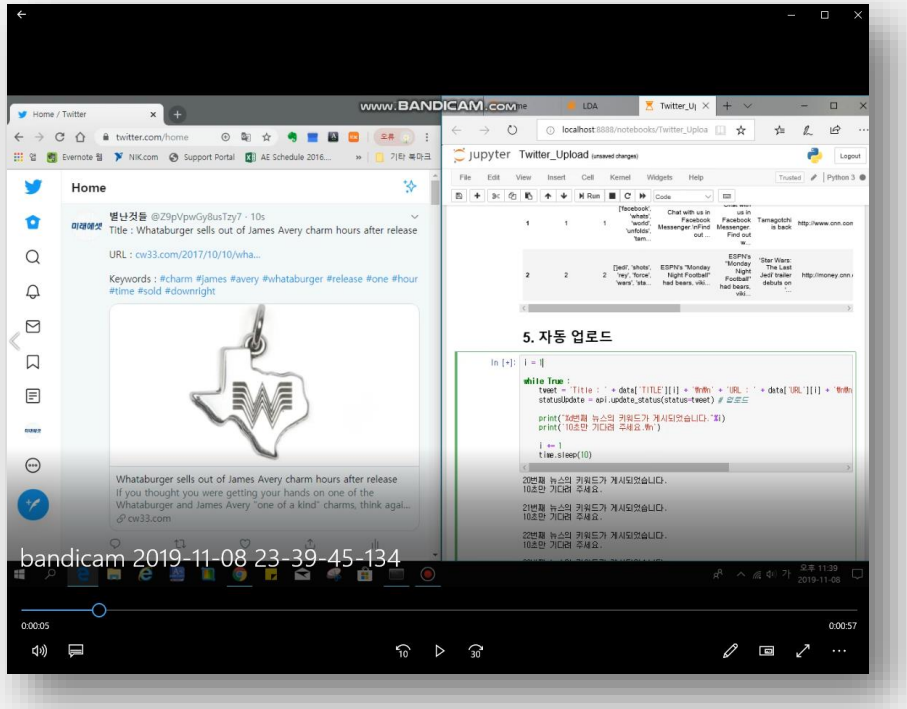
예측 키워드 : ['said', 'incident', 'ghana', 'gas', 'addo', 'ensure', 'seven', 'government', 'dozen', 'killed']

MIRAE ASSET

6

프로토 타입

프로토 타입 프로토 타입 설명



프로토 타입

- 프로토 타입은 실제 시연 영상입니다.

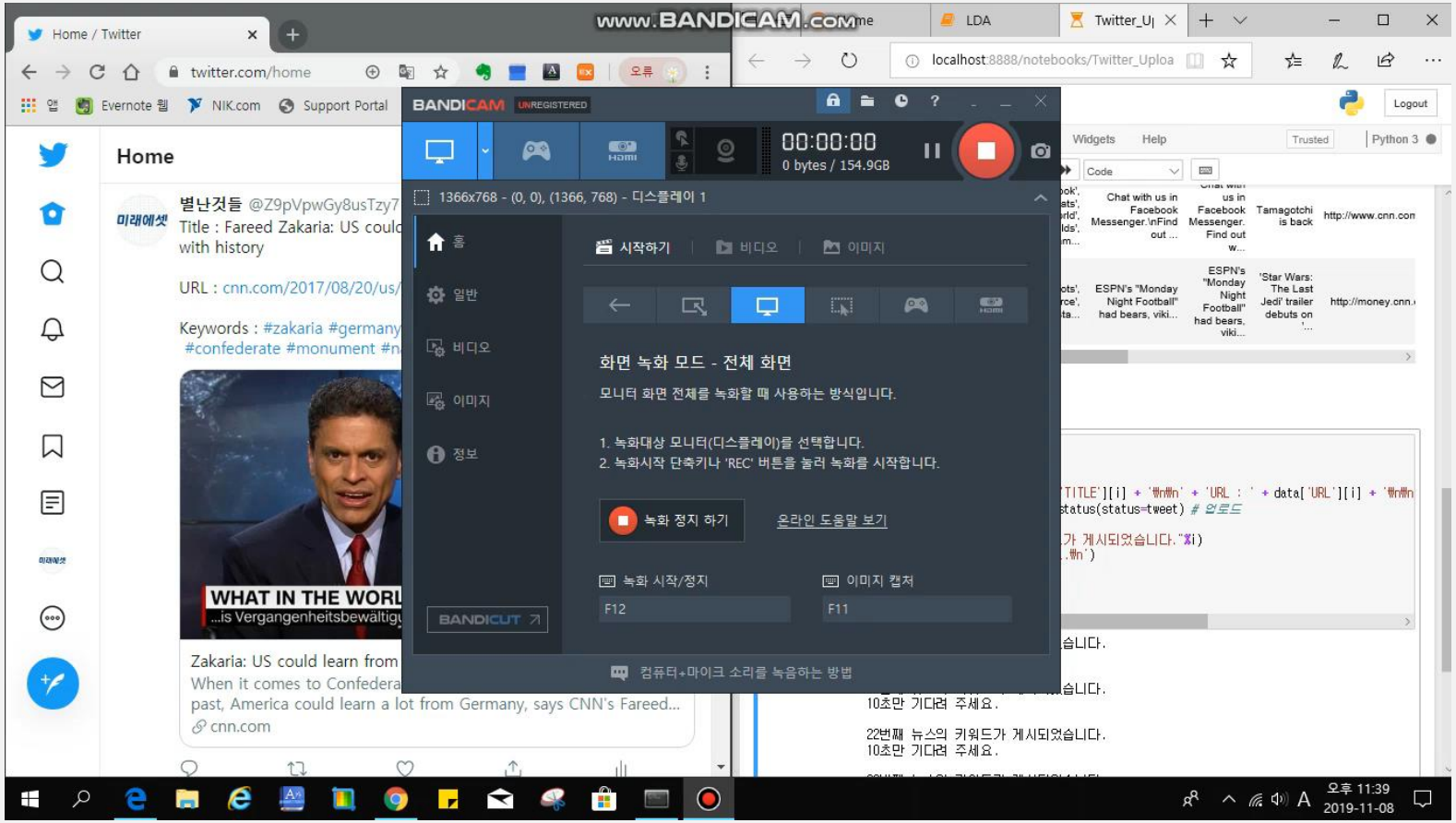
시연

- 시연을 위해 업로드 타임을 10초마다로 설정했습니다.
- 시연 이후 jupyter notebook을 종료했기 때문에, 이후 계속해서 콘텐츠가 업로드 되지는 않습니다.

MIRAE ASSET

프로토 타입

프로토 타입 영상



MIRAE ASSET

레퍼런스

http://www.bloter.net/archives/254773

https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/

https://bab2min.tistory.com/587

MIRAE ASSET

감사합니다