YES24 리뷰를 통한 감성 사전 구축

이상진

감성 사전 구축 과정

리뷰 스크랩(Selenium)

형태소 분석(Mecab)

리뷰, 부정 리뷰에서 출현 빈도가 높은 단어들을 각각 상위 15개씩 선정하여
이를 통해 '긍정·부정 기준단어집합(Standard Positive·Negative Words)' 생성

단어 임베딩(Word2Vec)

기준단어집합과 유사한 단어들의 집합 추출.
기준단어집합과 유사한 단어는 그 극성도 비슷할 것이라는 기본적인 개념에 기반함.

감성사전 구축에 이용할 데이터 수집

YES24.COM

대만여행 준비중이라면? 꼭 봐야할 책!

내용 ★★★★★ 편집/디자인 ★★★★★ | 째야 | 2019-10-13

내게 가이드북이란, 다르고 닳도록 읽고 또 읽어도 늘 설레임을 주는 책. 한 번쯤 꼭 가보고 싶은 여행지나 좋은 추억이 너무 많아 다시 가고 싶은 여행지가 생길 때면 서점에 들려 늘 가이드북을 사곤 한다. 어느 책을 읽을 때보다 경건한 마음으로 집중해서 정독을 하다보면 무심코 끌리는 그런 도시, 구석구석 보고 싶은 도시를 찾아낼 때가 있는데, 그 때의 희열...

1명이 이 리뷰를 추천합니다. ♡1 댓글 0 >

펼쳐보기

여행

내용 ★★★★★ 편집/디자인 ★★★★☆ | you507507 | 2019-10-11

대만 여행을 준비하면서 많은 도움이 되었던 책입니다.생소한 나라다보니 나라에 대한 기본 상식과 주의해야할 점, 교통, 날씨등 꼼꼼하게 알아가야 할거같았기에 이런 정보들을 블로그와 카페등에서 찾아보기 시작했습니다...역시나 단시간에 찾아보고싶었던 모든 내용을 찾아 정리하기에는 힘이 들었습니다.결국 책의 도움을 받기로 하였는데요 그 책이 바로 '타이...

이 리뷰가 도움이 되었나요? ♡0 댓글 0 >

펼쳐보기

포토리뷰

타이완 100배 즐기기!? 이미 타이완 100배 즐겼다^^

내용 ★★★★★ 편집/디자인 ★★★★★ | 오오모리노 | 2019-10-10

여행가기 전 반드시 해당 나라나 도시의 책을 꼭 구입하기 때문에 도서관에서 관련 책은 모두 비교해보고 구입합니다. 이번에는 엄마랑 단 둘이 가는 첫 여행이고, 타이완도 처음이어서 특히 신중하게 책을 고르게 되더라구요~ 각설하고!! 1) 장점 : 1. 비닐커버 _ 책을 가지고 다니기 때문에 이 부분이 꽤 신경쓰이는데 100배는 진짜 ?? 점입니다 ㅎㅎ 완전 센스 있...

이 리뷰가 도움이 되었나요? ♡0 댓글 0 >

펼쳐보기

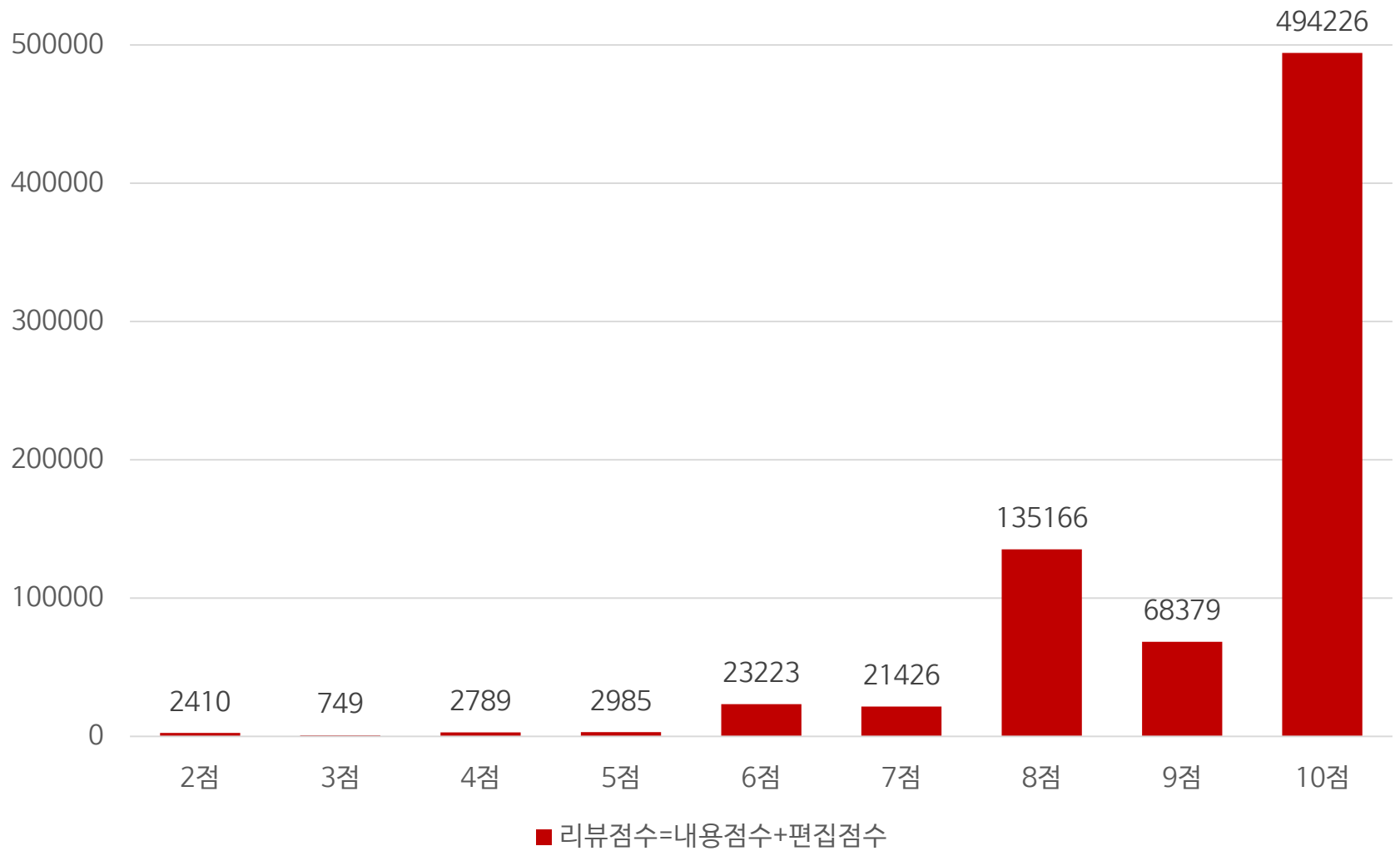
	STAR.	REVIEW_TITLE	REVIEW_CONTENT
	…	필터	필터
1	5	차라리 혼자 …	이책은 결혼한 사람 누구에게 물어봐도 10중 8,9는 결혼생활
2	5	차라리 혼자…	이 책의 제목을 처음 봤을 때는 요즘 결혼하지 않는 싱글족들
3	5	결혼전에 알…	아이들이 장성하고결혼적령기에 접어들어이렇게 접하게되는
4	5	차라리 혼자 …	서로 모르는 사이였던 남녀가 만나 깊은 정과 공감을 나누고
5	5	차라리 혼자 …	원하는 이상형과 결혼해서 갈등없이 살아가는 부부가 과연
6	5	차라리 혼자 …	차라리 혼자 살걸 그랬어 - 라리루 책의 띠지는 이 책이
7	4	차라리 혼자…	차라리 혼자살걸 그랬어 가정행복코치인 저자는 이 책을 통
8	5	[서평] 차라…	=가정행복코치. 이수경의 충격고백.
9	5	차라리 혼자…	저는 결혼 5년차 전업 주부입니다.현재 5세딸과 9개월된 아들
10	5	차라리 혼자 …	인간은 참으로 이상한 존재가 아닐 수 없다.혼자일 때는 둘이
11	5	차라리 혼자…	사랑한다는 것은 기적과 같다. 특히, 남녀가 만나 사랑하는
12	5	둘이라면 꼭 …	누군가를 만나서 사랑에 빠지고, 그 사랑의 영원함을 믿으며
13	5	[책이있는마…	상처뿐인 결혼생활을 회복하는 사랑의 기술 「차라리 혼자 살
14	5	차라리 혼자 …	사랑하는 남녀가 만나 연예를 하고 결혼하는 과정을 바라보다
15	5	차라리 혼자…	'차라리 혼자 살걸 그랬어' 라는 책을 읽게 되었다. 어쩌면 저
16	5	불쑥 불쑥 찾…	행복.. 행복.. 행복한 결혼생활을꿈꾼다..내가 선택한 결혼..
17	5	이토록 따뜻…	그림의 색조, 말, 분위기가 모두 따뜻해서 차가웠던 마음이
18	5	연애도 결혼…	처음에는 페이스북 페이지를 통해서 이 만화의 한 컷 한 컷을
19	5	신혼의 달달…	신혼의 달달달한 모습을 그려주셨습니다이가 아플정도의 너
20	5	결혼준비 웹…	결혼준비란 게 생각보다 넘 복잡하고 할 것도 너무 많더라구
21	5	결혼을 준비…	여러모로 예상했던 것보다 훨씬 놀라웠던 책이라고 말씀드리
22	5	[저자 권도헌,…	결혼한지도 벌써 4년차에 접어든 부부~~ 정말 요즘 친구들

감성사전 구축에 이용할 데이터 수집

https://github.com/berry2971/scraper-yes24

YES24 리뷰 개요

전체: 751,353건



YES24 리뷰 개요

상당히 편중된 경향을 보이기에, 즉 일반적으로 후한 점수를 주는 것으로 사료되기에 다음과 같이 분류를 시도하여도 의미가 있을 것으로 판단함.

score≥10	긍정	494,226건
8≤score≤9	중립	
score≤7	부정	53,584건

YES24 리뷰 개요

상당히 편중된 경향을 보이기에, 즉 일반적으로 후한 점수를 주는 것으로 사료되기에 다음과 같이 분류를 시도하여도 의미가 있을 것으로 판단함.

Classify(S = REVIEW_STAR_CONTENT + REVIEW_STAR_EDITION)

S≥10	Positive	494,226건
8≤S≤9	Neutral	
S≤7	Negative	53,584건

실제 분석은 Positive 100,000건, Negative 53,584건을 통해 진행함

데이터 분석 _ 형태소 분석

사용한 형태소 분석기: Mecab

- khaiii, kkma, hannanum 등에 비해 월등히 빠른 속도
- 품사 태깅이 가능해야 하기에 지도 학습 기반 형태소 분석기를 선택함
- 비교적 정형화된 데이터이기에 komoran보다 나은 선택이라 판단함
- okt와 달리 VCP(지정사), ETM(관형형 전성어미) 등 분리 가능하여 데이터의 추후 사용성 증가

긍정·부정 기준단어집합(Standard Positive·Negative Words) _ VA(형용사)

1. 긍정·부정 리뷰에서의 빈출 단어 추출

2. 각 극성에 대해 상위 15개의 단어가 기준단어집합이 됨(Song et al, 2010)

3. 동일 단어에 대해서는 출현률이 높은 극성에 해당하는 단어로 분류

긍정·부정 기준단어집합(Standard Positive·Negative Words) _ VA(형용사)

긍정 단어			부정 단어		
리뷰 개수	100000		리뷰 개수	53584	
토큰 개수	1106161		토큰 개수	525768	
단어	빈도	확률	단어	빈도	확률
좋	178168	0.161069	없	83548	0.158907
같	160827	0.145392	같	72904	0.138662
없	123268	0.111438	있	61217	0.116433
있	119705	0.108217	좋	60341	0.114767
많	108427	0.098021	많	45397	0.086344
쉽	55904	0.050539	쉽	17249	0.032807
어렵	19455	0.017588	그렇	14760	0.028073
그렇	19389	0.017528	어렵	7915	0.015054
맛있	18301	0.016545	재미있	7794	0.014824
높	16779	0.015169	크	7607	0.014468
재미있	16342	0.014774	작	6759	0.012855
크	15935	0.014406	다르	6544	0.012447
작	15319	0.013849	높	5807	0.011045
힘들	13801	0.012476	깊	5803	0.011037
다르	12019	0.010866	힘들	5732	0.010902
적	6998	0.006326	괜찮	4854	0.009232
재밌	6570	0.005939	짧	4124	0.007844
깊	6402	0.005788	적	3827	0.007279
괜찮	6242	0.005643	수많	3524	0.006703
싫	6240	0.005641	싫	3488	0.006634
편하	6153	0.005562	아쉽	3096	0.005889
즐겁	6035	0.005456	어떻	2834	0.00539
수많	5864	0.005301	재밌	2693	0.005122
짧	5841	0.00528	젊	2384	0.004534
빠르	5667	0.005123	가볍	2224	0.00423
어떻	5453	0.00493	옳	2107	0.004007
예쁘	4812	0.00435	길	2073	0.003943
늦	4516	0.004083	낮	1976	0.003758
낮	4278	0.003867	새롭	1927	0.003665
새롭	4152	0.003754	편하	1887	0.003589

긍정 단어			부정 단어		
리뷰 개수	100000		리뷰 개수	53584	
토큰 개수	1106161		토큰 개수	525768	
단어	빈도	확률	단어	빈도	확률
좋	178168	0.161069	없	83548	0.158907
같	160827	0.145392	있	61217	0.116433
많	108427	0.098021	그렇	14760	0.028073
쉽	55904	0.050539	재미있	7794	0.014824
어렵	19455	0.017588	크	7607	0.014468
맛있	18301	0.016545	다르	6544	0.012447
높	16779	0.015169	적	3827	0.007279
작	15319	0.013849	깊	5803	0.011037
힘들	13801	0.012476	괜찮	4854	0.009232
재밌	6570	0.005939	싫	3488	0.006634
편하	6153	0.005562	수많	3524	0.006703
즐겁	6035	0.005456	짧	4124	0.007844
빠르	5667	0.005123	어떻	2834	0.00539
예쁘	4812	0.00435	아쉽	3096	0.005889
늦	4516	0.004083	젊	2384	0.004534

긍정·부정 기준단어집합(Standard Positive·Negative Words) _ MAG(부사)

긍정 단어			부정 단어		
리뷰 개수	100000		리뷰 개수	53584	
토큰 개수	1106161		토큰 개수	525768	
단어	빈도	확률	단어	빈도	확률
더	100451	0.09081	더	46091	0.087664
잘	100055	0.090452	잘	32824	0.062431
많이	61741	0.055816	많이	21493	0.040879
정말	59586	0.053867	안	20987	0.039917
함께	50094	0.045286	다	19616	0.037309
안	44614	0.040332	좀	19580	0.037241
가장	43628	0.039441	너무	19280	0.03667
다	42554	0.03847	가장	16525	0.03143
어떻게	40884	0.03696	또	15799	0.030049
너무	40772	0.036859	다시	15117	0.028752
바로	36727	0.033202	어떻게	14993	0.028516
이렇게	33764	0.030524	정말	14716	0.02799
또	33482	0.030269	함께	13457	0.025595
꼭	31365	0.028355	왜	12411	0.023605
좀	26848	0.024271	바로	12126	0.023063
다시	26764	0.024195	조금	11939	0.022708
참	22130	0.020006	이렇게	11782	0.022409
특히	21910	0.019807	물론	11053	0.021023
왜	21462	0.019402	사실	9745	0.018535
없이	20257	0.018313	지금	9693	0.018436
지금	19746	0.017851	그냥	9618	0.018293
먼저	19394	0.017533	없이	9560	0.018183
못	19320	0.017466	그렇게	9476	0.018023
조금	18624	0.016837	아직	8865	0.016861
사실	18217	0.016469	못	8478	0.016125
물론	17524	0.015842	특히	7974	0.015166
아주	17507	0.015827	아주	7616	0.014485
제대로	17345	0.01568	모두	7460	0.014189
아직	16913	0.01529	얼마나	7202	0.013698
매일	16815	0.015201	꼭	6960	0.013238
모두	16635	0.015038	같이	6648	0.012644
같이	16530	0.014944	이미	6564	0.012485
직접	16156	0.014605	먼저	6504	0.01237
더욱	15614	0.014115	참	6490	0.012344
얼마나	14648	0.013242	제대로	6026	0.011461
이제	14562	0.013164	스스로	5603	0.010657
그냥	14038	0.012691	거의	5268	0.01002
열심히	14017	0.012672	그대로	5251	0.009987
스스로	13631	0.012323	전혀	5110	0.009719
자주	12963	0.011719	더욱	5060	0.009624

긍정 단어			부정 단어		
리뷰 개수	100000		리뷰 개수	53584	
토큰 개수	1106161		토큰 개수	525768	
단어	빈도	확률	단어	빈도	확률
더	100451	0.09081	좀	19580	0.037241
잘	100055	0.090452	다시	15117	0.028752
많이	61741	0.055816	왜	12411	0.023605
정말	59586	0.053867	지금	9693	0.018436
함께	50094	0.045286	조금	11939	0.022708
안	44614	0.040332	사실	9745	0.018535
가장	43628	0.039441	물론	11053	0.021023
다	42554	0.03847	아직	8865	0.016861
어떻게	40884	0.03696	얼마나	7202	0.013698
너무	40772	0.036859	그냥	9618	0.018293
바로	36727	0.033202	그렇게	9476	0.018023
이렇게	33764	0.030524	이미	6564	0.012485
또	33482	0.030269	거의	5268	0.01002
꼭	31365	0.028355	그대로	5251	0.009987
참	22130	0.020006	전혀	5110	0.009719

단어 임베딩

```
from gensim.models import Word2Vec
model = Word2Vec(posv_sentences + negv_sentences,
                 size=100, window=2, min_count=10, workers=4, sg=1, iter=25, sample=0.00001)
```

Set parameters for genism Word2Vec model

- size: Dimension of literal vector
- window: Number of types algorithm considers
- min_count: Types only appeared more than min_count are considered.
  By setting, we can avoid removing all book titles from review text.
- workers: Number of cores for multi-processing.
- sg: 1 for Skip-gram, other for CBOW
- iter: Number of iteration
- sample: The threshold for configuring which higher-frequency words are randomly downsampled

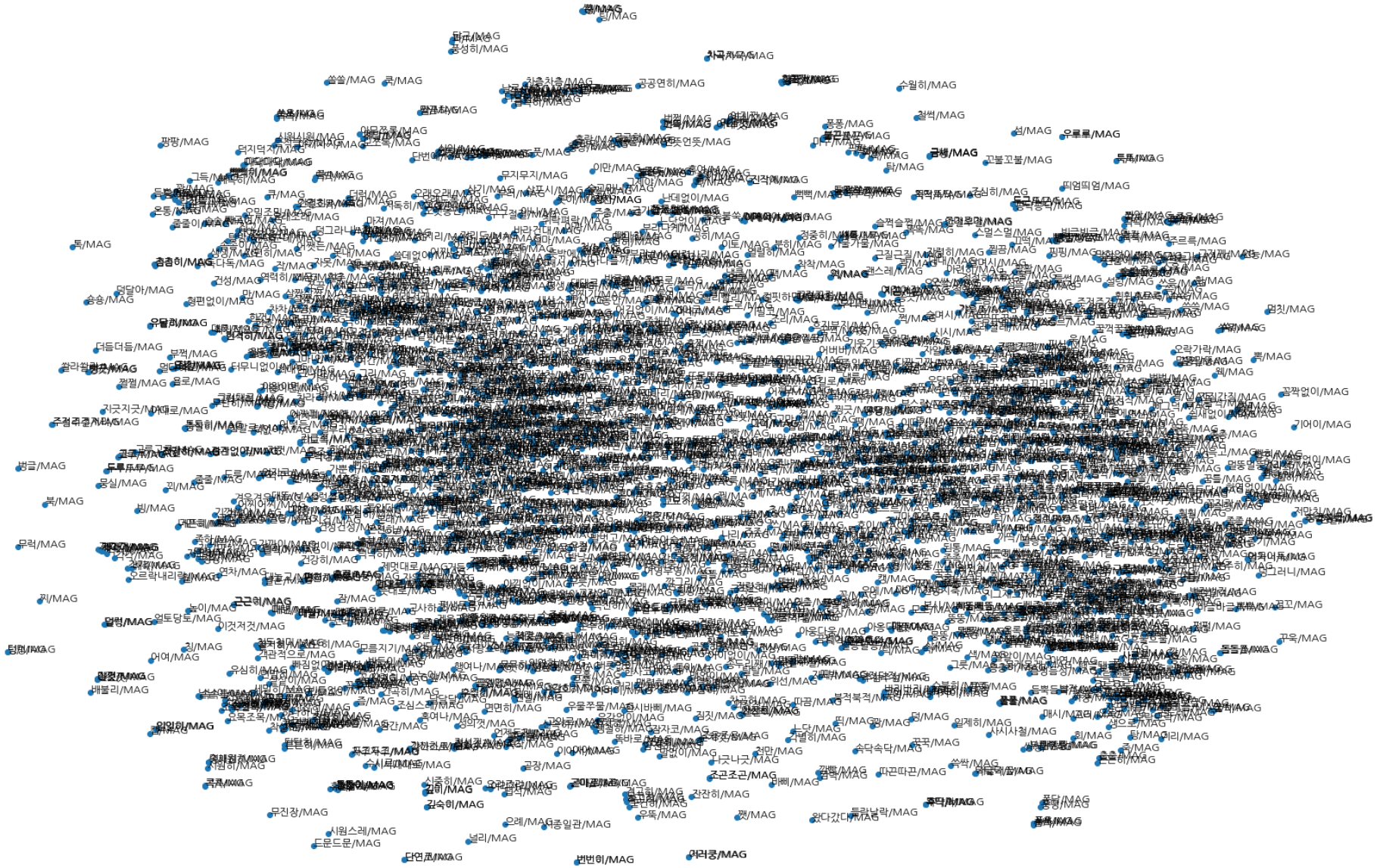
단어 임베딩 결과 시각화

100차원 → 2차원의 차원 축소(dimension reduction)을 위해
t-SNE(t-Stochastic Neighbor Embedding) 알고리즘 사용

- PCA(Principal component analysis), LDA(Linear discriminant analysis), NMF(None-negative matrix factorization) 등의 기존 차원 축소 방식에 비해 군집성 등을 유지시키므로 시각화에 비교적 적합함
- Van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing Data Using t-SNE". *Journal of Machine Learning Research*. 9: 2579–2605.

단어 임베딩 결과 시각화 _ VA(형용사)

단어 임베딩 결과 시각화 _ MAG(부사)



결과물 _ 후처리

이상훈 외(2016)에서는 단어를
strong positive/positive/neutral/weak negative/strong negative로 나누었으나,
이러한 방식은 예를 들어 감성 사전을 기반으로 감성을 예측하는 모듈을 만들 때,
해당 단어가 갖는 긍정 및 부정의 정도성을 불연속적으로 변환시켜 버린다.

이 과제에서는 긍정 형용사 목록과 부정 형용사 목록을 join시키면서,
단어의 극성 처리를 아래와 같이 한다.

positiveness of word = max_similarity(word, SP)
negativeness of word = max_similarity(word, SN)
(SP = list of standard positive words, SN = list of standard negative words)

결과물

형용사			부사		
단어	긍정도	부정도	단어	긍정도	부정도
있	0.66096	1	조금	0.743067	1
같	1	0.66096	더	1	0.847858
없	0.538521	1	덜	0.749048	0.718608
많	1	0.675437	보다	0.824843	0.697935
적	0.659974	1	약간	0.626829	0.786943
좋	1	0.754254	너무	1	0.687111
그렇	0.556243	1	훨씬	0.839599	0.682007
이렇	0.342276	0.696704	확실히	0.626409	0.633932
나쁘	0.624553	0.702556	다소	0.645027	0.708398
다르	0.454324	1	아무래도	0.615668	0.691423
어렵	1	0.560617	상당히	0.679434	0.622466
상관없	0.399659	0.578255	살짝	0.556808	0.693433
옳	0.356826	0.529297	아주	0.741847	0.600496
미덥	0.425734	0.528994	그나마	0.55391	0.628074
별다르	0.455126	0.527998	한층	0.737364	0.597245
	...			...	