



Parameter Optimization & 개요

서울대학교 산업공학과

양호성, 조성준

hoseong@dm.snu.ac.kr, zoon@snu.ac.kr

1

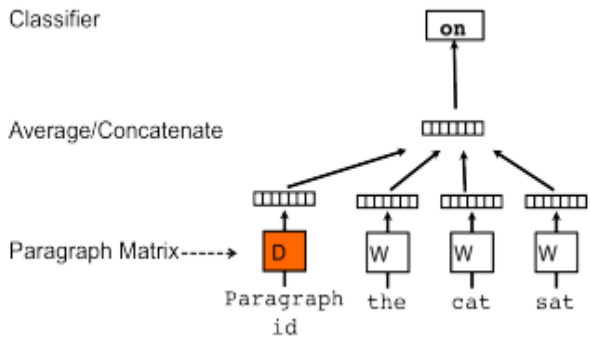
Word2vec / Doc2vec Parameters

Word2vec 방법은 다양한 parameter가 존재함

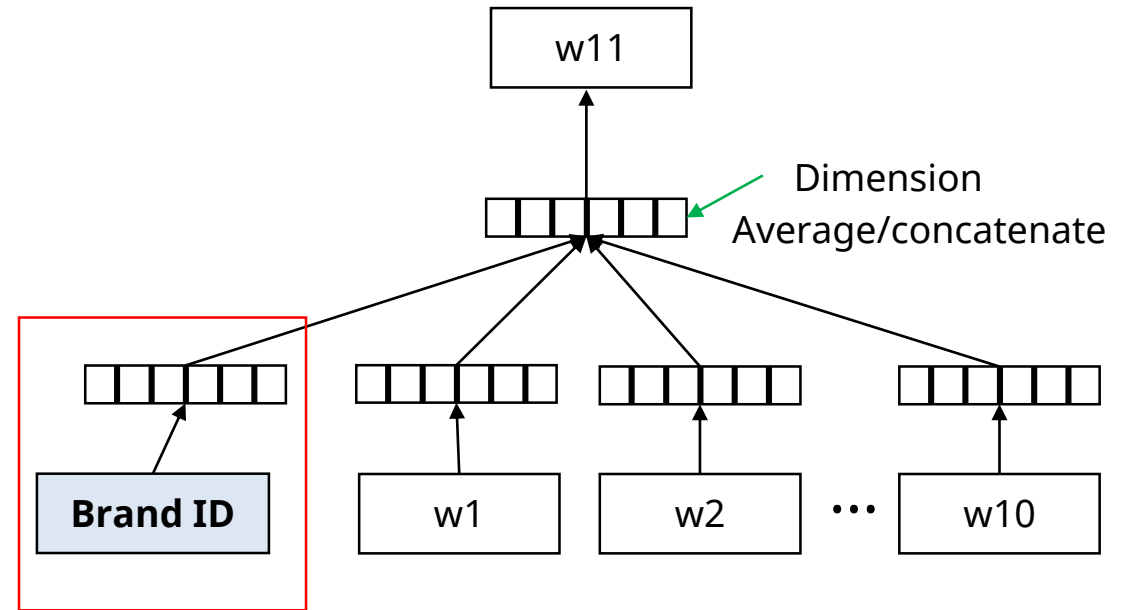
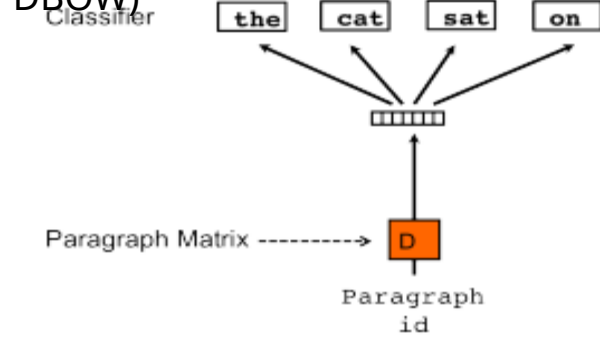
특히, word dimension, window size, training epoch 가 민감한 parameter

- **Dimension**
- **Windows**
- **Training epoch**
- Methods (PV-DM/PV-DBOW)
- Concatenate/average
- Negative Sampling, Subsampling of Frequent words 유무

distributed memory (PV-DM)



distributed bag of words (PV-DBOW)



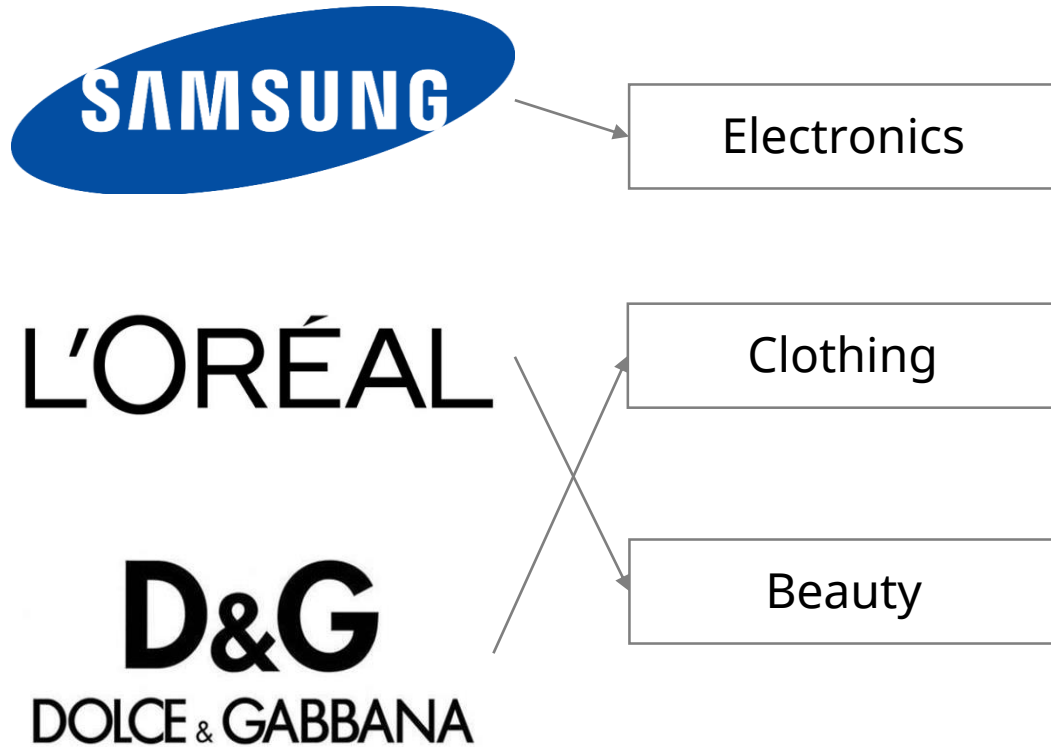
2

실험 설계

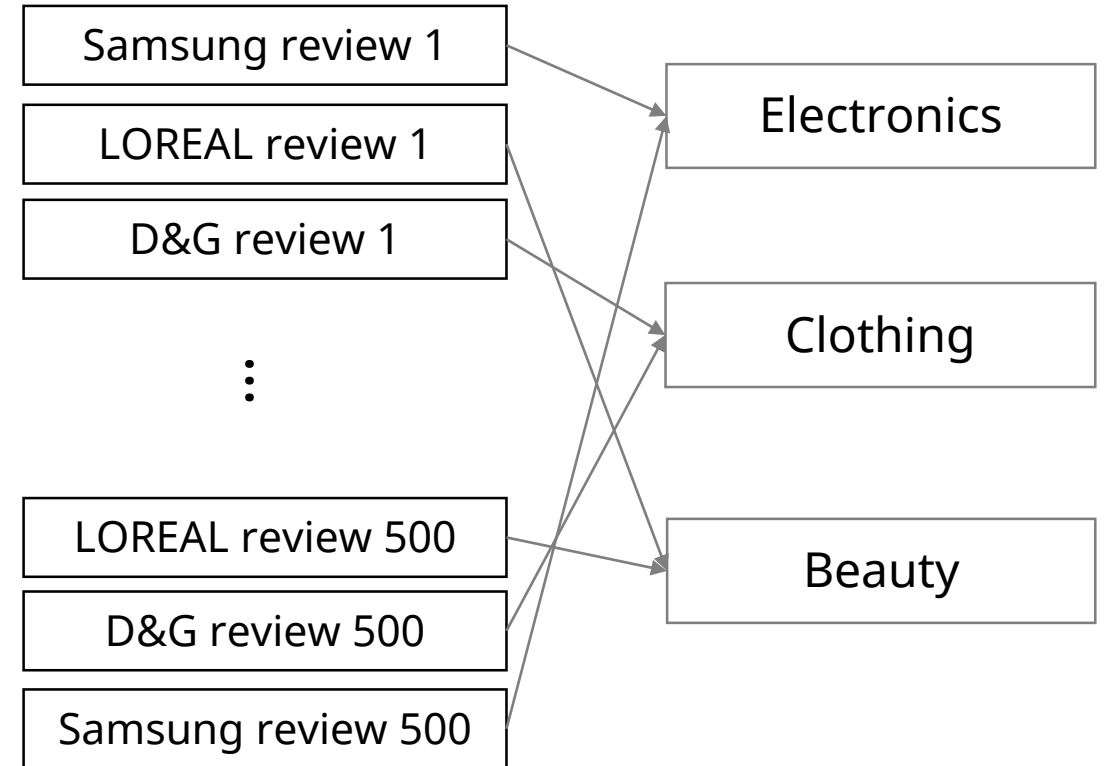
Brand Vector 가 잘 embedding 되었는지 확인하기 위해서 아래와 같이 category 를 분류하는 문제를 제안함.

그러나, Brand 개수는 전체 문서 개수나, word 개수에 비해 매우 적기 때문에 브랜드 정보를 잘 학습한다고 해서, 단어나 문서 정보를 잘 포함하고 있는지는 알 수 없음.

실험 1



실험 2



3

실험 결과

각 category 에서 review 개수가 많은 3000개의 브랜드를 선택하였다.
그 중, 50만개의 review 를 임의로 선택하였다.

Dataset

Electronics, Clothing, Beauty 각 category 에서 review 수가 많은 3000 개의 브랜드 선택

- Electronics : 2,902,574
- Beauty : 1,207,767
- Clothing : 796,366

Category 별 review 개수가 다르므로, 50만개씩 sampling 총 150만개 dataset 사용

실험 1 Doc2vec을 이용한 Review dataset 에 맞는 parameter setting

실험 2 Brand2vec을 이용한 Brand Vector 에 맞는 parameter setting

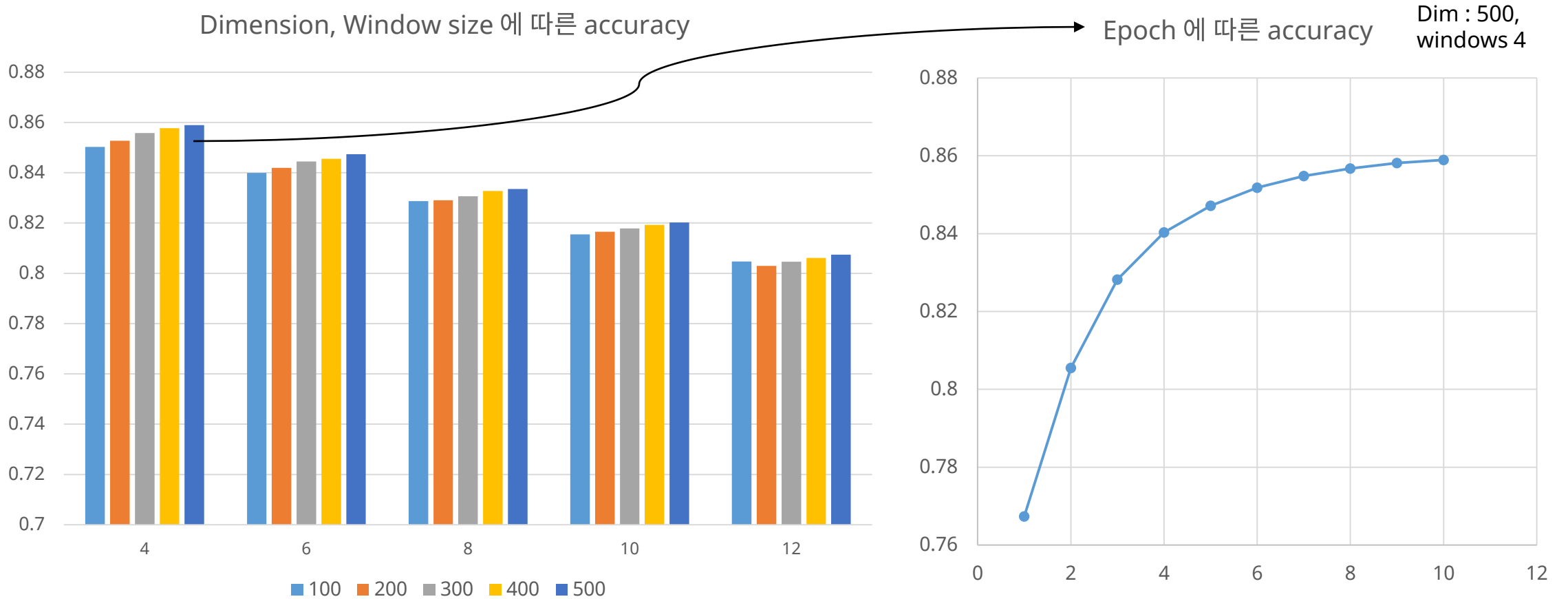
3

실험 1 결과

실험 1은 기존의 Doc2vec 방법을 이용하여 각 review 문서를 학습시킨다.

학습된 document vector 들을 training/test set으로 나눠서 training set으로 logistic regression 모델을 학습한다.

각 window size, dimension, epoch 에 따라 test accuracy 를 확인한다.



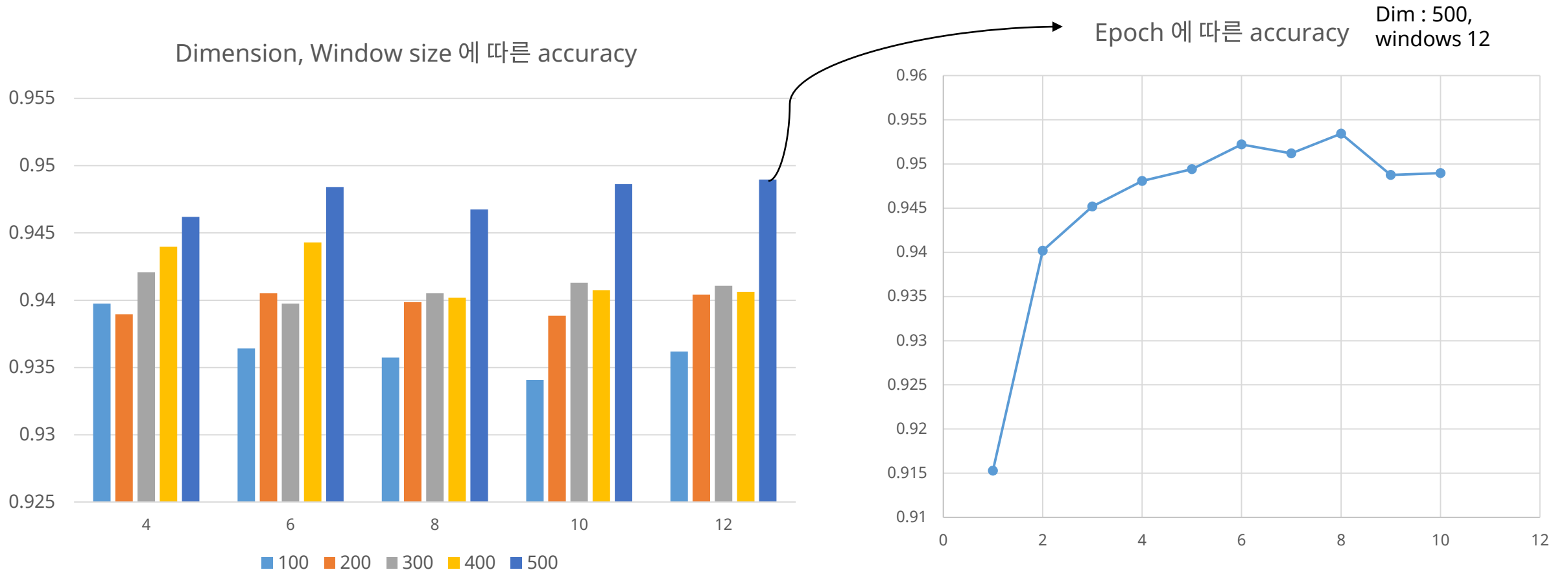
3

실험 2 결과

실험 2는 제안한 Brand2Vec 방법을 이용하여 각 Brand Vector를 학습시킨다.

학습된 Brand Vector 들을 training/test set으로 나눠서 training set으로 logistic regression 모델을 학습한다.

각 window size, dimension, epoch 에 따라 test accuracy 를 확인한다.

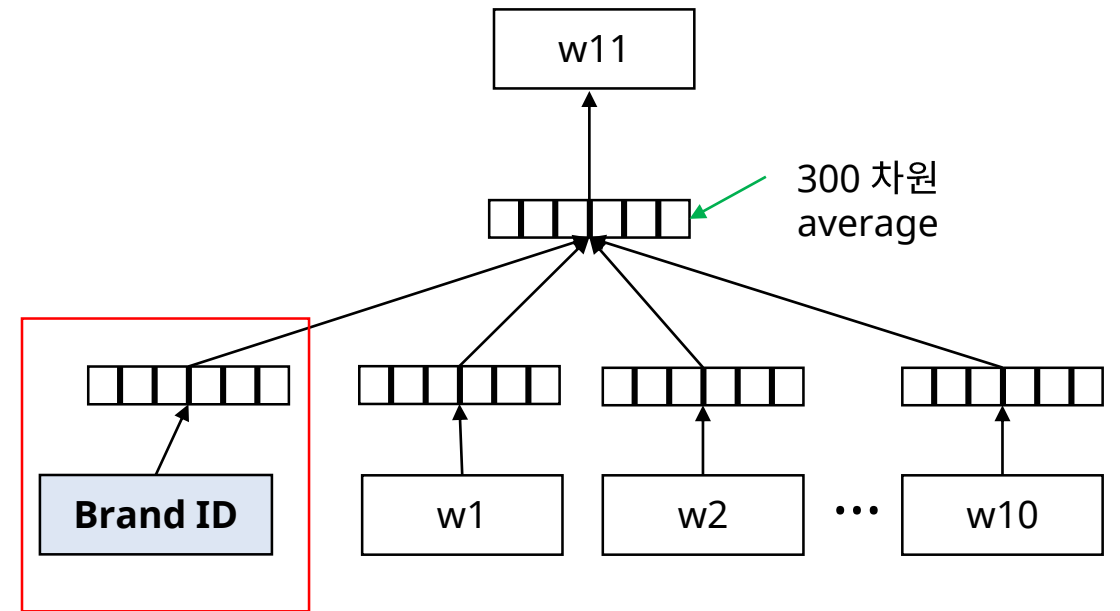


3

실험 1 & 2 를 종합한 결과

1, 2 모두 dimension 이 500일 경우 category 분류 성능이 가장 좋았으나, 반대로, Document 단위로 학습할 경우 window size가 작을 경우 성능이 좋았으며, Brand vector 의 경우는 클 수록 학습이 잘 되었다. 따라서, 두 가지를 고려하여 Dimension 500, window 8, epoch 10

- Parameter setting
 - 1) Brand vector, Word vector dimension : 500
 - 2) Window : 8
 - 3) Epoch : 10
 - 4) Model : Distributed Memory model
 - 5) 300번 이하로 등장한 단어는 제거
 - 6) Average / concatenation 방법 중 average





Brand2vec : Distributed representation of Brands and Applications

서울대학교 산업공학과

양호성, 조성준

hoseong@dm.snu.ac.kr, zoon@snu.ac.kr

0 목차

1. 서론

- UGC(User generated content)의 중요성, 기존 설문조사와의 차이점
- Counting-based 방법의 장, 단점
(전처리 \uparrow , 자동화 \times , 객관성 \times , reproducible)
- Counting-based 방법의 단점을 극복하기 위해 Word2vec 을 응용한 Brand2vec 방법을 제안

2. 관련 연구

- UGC를 활용한 사례
- Word representation

3. 제안하는 방법 : **Brand2Vec**

- Brand2Vec 설명
- Parameter search (문제 제기, 실험 설계)

4. 실험 결과 및 **application**

- Dataset description (Amazon dataset에 대한 설명)
- Parameter search 결과
- Application (1. 제품 속성에 기반한 Positioning, 2. Keyword extraction)

5. Discussion

- 기존 방법과의 비교
- 추가 확장 가능성 (연예인, 정치인, 영화 등)

5. 결론

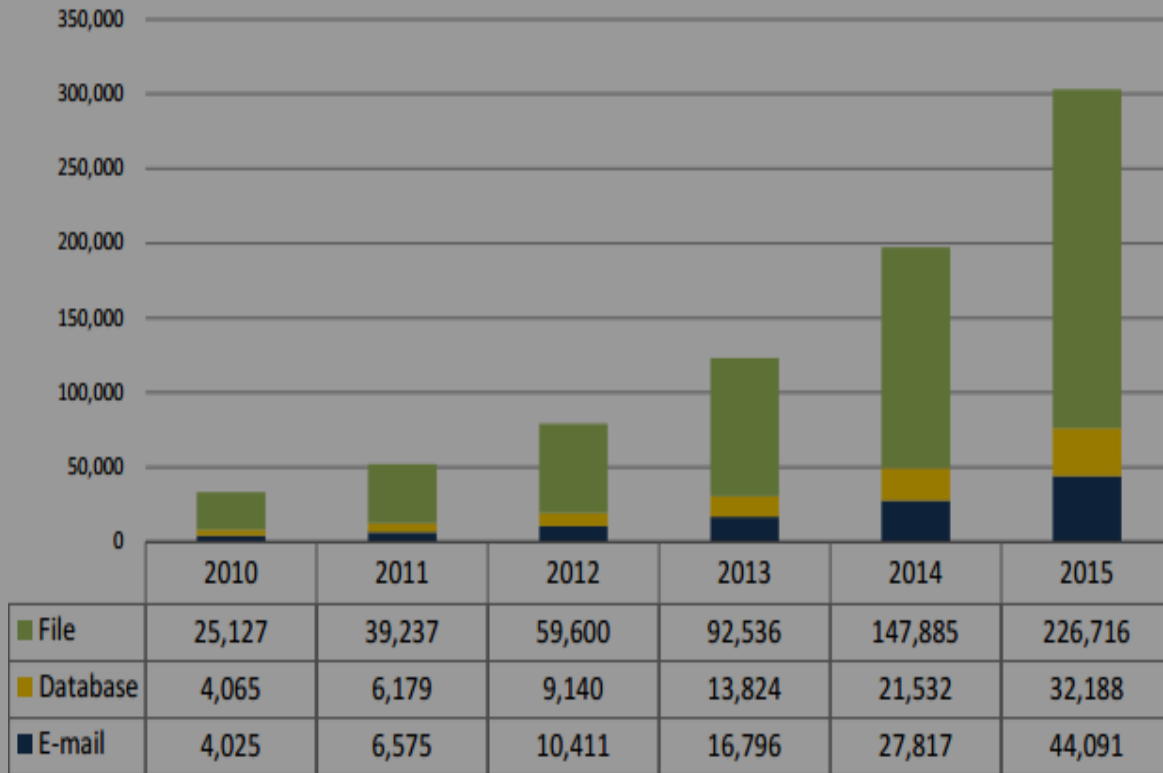
- Summary
- Limitation & Future work

1

서론

- UGC(User generated content)의 중요성, 기존 설문조사와의 차이점
- Counting-based 방법의 장, 단점 (장: 직관적/ 단: 빈도가 적은 단어는 무시, 객관성이 떨어지고, reproducible 하지 않음)
- Counting-based 방법의 단점을 극복하기 위해 Word2vec 을 응용한 Brand2vec 방법을 제안

Total Worldwide Digital Archive Capacity, by Content Type, 2010-2015 (Petabytes)



*File = File-based or Unstructured data

Source : Enterprise Strategy Group, 2010

설문조사

- 통계적인 분석이 쉽고, 원하는 질문에 대한 결과를 얻기 용이함
- 표본 수 제한 및 많은 비용 발생
- 특정 설문 환경에서 실시하기 때문에 조사자의 주관의 개입

Review / Social Media

- 표본 수가 매우 많아 bias를 줄일 수 있음
- 소비자들의 실제 목소리를 반영
- 적은 비용으로 데이터 수집이 가능함
- Noise가 심해 분석이 어려움
- 비정형 데이터인 텍스트를 분석하기 어려움

2

관련 연구 - 1) UGC를 활용한 사례

Text mining 을 활용해 마케팅, 브랜드 전략에 활용한 사례, 연구들

관련 연구

- 다양한 Text mining 기법을 활용하여 Business 분야에 적용하는 연구가 이뤄짐.
- 특정 브랜드의 Social Media 페이지 정보를 활용하거나[2], 트위터에서 브랜드의 sentiment 정보를 이용하거나[4], LDA(Latent Dirichlet Allocation)을 이용한 방법[7] 등을 통해 보다 나은 의사결정을 위한 방법을 제시

관련 실제 적용 사례

- SKT에서는 Social Media Buzz량을 통하여 광고 효과를 분석¹⁾
- LG 전자는 Social Media 데이터를 이용해 제품 개발 프로세스에 적용함.²⁾
- 다음소프트, 솔트룩스, LG CNS 등은 자체 Social Media Text 분석 시스템을 구축하여 다양한 산업군에 적용하고 있음.

1) <http://www.bizwatch.co.kr/pages/view.php?uid=8966>

2) www.etnews.com/20141124000268

2

관련 연구 - 2) Word Representation

Text를 분석하기 위해서는 Text를 '숫자'로 변환하는 과정이 필요함.

단어를 숫자로 표현하는 방법은 크게 단어의 등장 빈도를 계산하는 Discrete 한 방법과, Neural network 등을 통해 Distributed 하게 나타내는 방법이 있다.

Discrete representation

예) One-hot vector
Word-word / Word-Document Matrix

$$dog = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad cat = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad pig = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

- 전체 문서 혹은 특정 범위에서 등장한 단어의 비율 / 확률을 계산하는 방법
- 단어와 단어 사이의 'similarity' 비교가 불가능
- 문맥 상 중요한 단어더라도 '빈도'가 낮으면 무의미한 단어로 분석

Distributed Representation

예) Word2vec, Glove

$$dog = \begin{bmatrix} 1.5 \\ 0.3 \\ 0.8 \end{bmatrix} \quad cat = \begin{bmatrix} 1.8 \\ 1.1 \\ 0.2 \end{bmatrix} \quad pig = \begin{bmatrix} 1.6 \\ -2.3 \\ -1.5 \end{bmatrix}$$

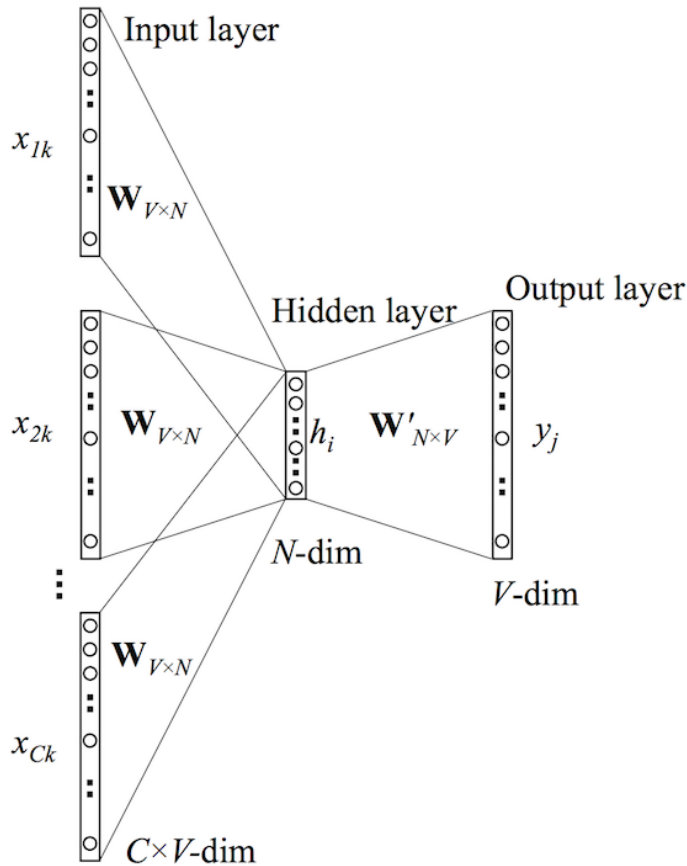
- Neural Network 를 통해 단어를 continuous 한 vector 로 변환
- Bengio[1] 가 처음 제안해서 최근 Mikolov[5] 의 방법(Word2vec)이 주목받음
- 단어 별 similarity 계산이 가능하며, 비슷한 문맥에서 사용되는 단어들이 similarity 가 높음

2

관련 연구 - 2) Word representation

Word2Vec은 주어진 C 개의 window 사이즈 만큼의 단어가 있을 때, 그 다음에 등장할 단어의 확률인 $p(w_o|w_I)$ 를 Maximize 하는 $W_{V \times N}$ 과 $W'_{N \times V}$ 을 Neural Network 를 이용하여 구하는 방법이다.

Gradient Descent 방법을 이용해 학습된 $W_{V \times N}$, $W'_{N \times V}$ 의 평균값을 해당 단어의 representation 이라 한다.



그림출처 : Reference

[6]

- 초기화
 - Vocabulary size V , Hidden layer size N (=Word Vector Size)
 - Input node 값들은 각 단어들의 One-hot encoding 값으로 $x_1=[1,0,\dots,0], x_2=[0,1,\dots,0], x_V=[0,0,\dots,1]$
 - $W_{V \times N}$ 과 $W'_{N \times V}$ 는 random 하게 초기화

x_k 에 해당하는 단어가 나왔다면, $x_k^T \cdot W = W_k = v_{wI}$
지정된 window 를 돌면서 각 단어마다 나오는 v_{wI} 값을 평균하여 h 를 구한다.

최종적으로 Input 단어가 주어졌을 때 다음 단어를 예측하게 될 확률은

$$p(w_o|w_I) = \frac{\exp(v'_{w_o} \cdot v_{wI})}{\sum_{j=1}^V \exp(v'_{w_j} \cdot v_{wI})}$$

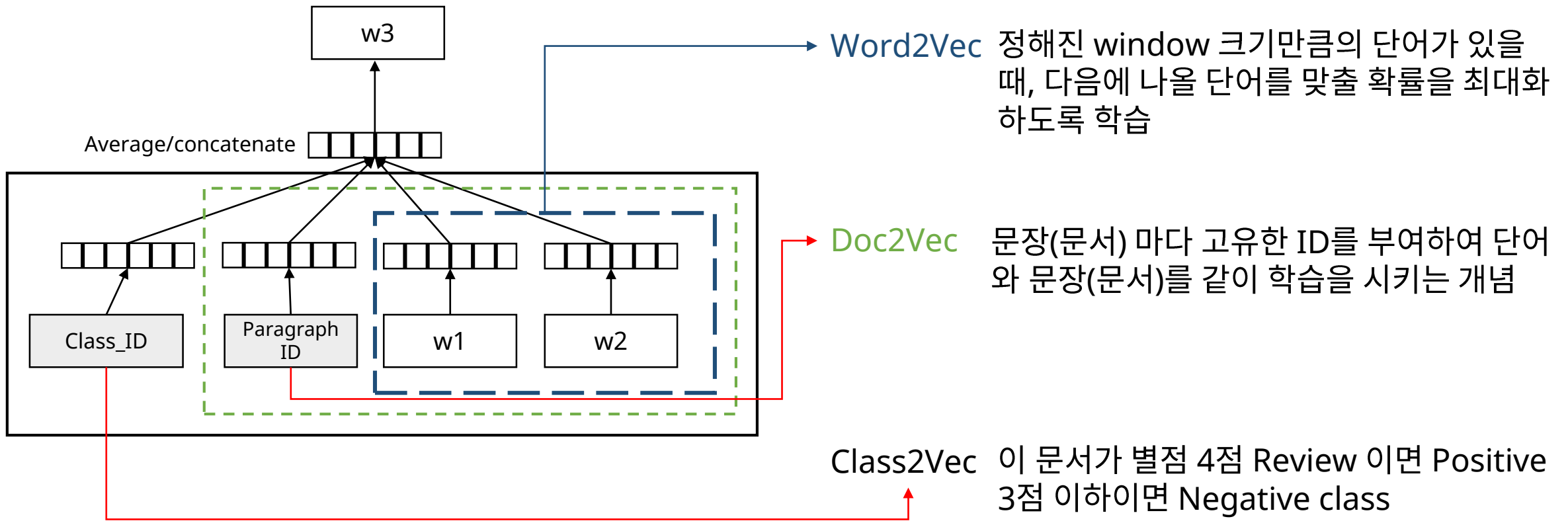
- Training
 - $\max p(w_o|w_I)$
 - Gradient Descent (+Negative Sampling)

2

관련 연구 - 2) Word representation

기존의 Word2Vec 방법을 응용하여 문장(혹은 문서나 class)를 단어와 함께 학습을 시키는 방법이 등장하였다.

Doc2vec [3]은 단어와 Paragraph 정보를 통해 다음 단어를 예측하는 모델이며, Class2vec[8]은 긍정/부정과 같은 class 부여하여 단어와 함께 학습시키는 모델이다.



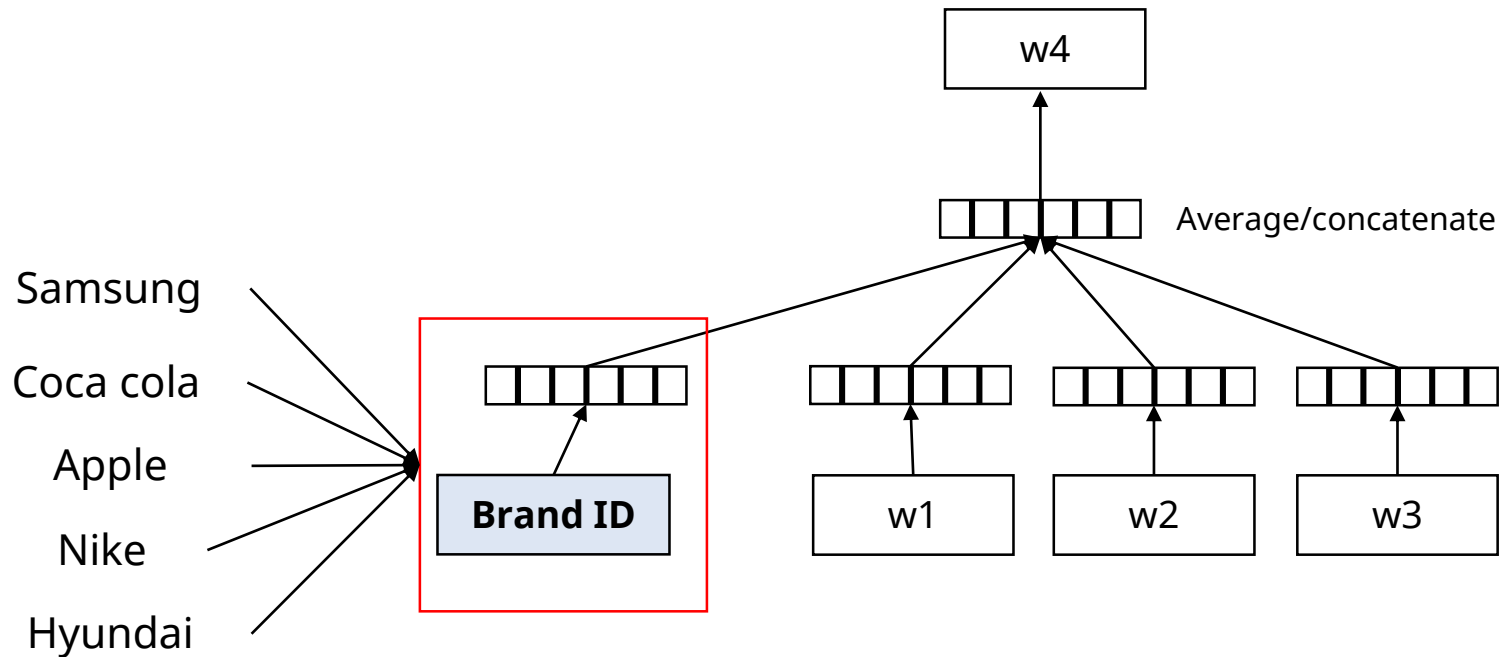
3

Proposed Method – 1) Brand2vec

Distributed Memory Model[3] & Class2vec[8] 의 아이디어를 접목

특정 브랜드에 대해 Social media 혹은 product review data에서 말하는 모든 정보들을 모아
특정 브랜드를 '하나'의 vector로 표현 하면?

예) Samsung 제품에 대한 Review 인 경우 Brand_ID = 'Class_brands_Samsung' 과 같이 고유한 ID를 부여하여
모든 Samsung 제품의 Review를 반영한 하나의 Vector를 생성할 수 있음



3

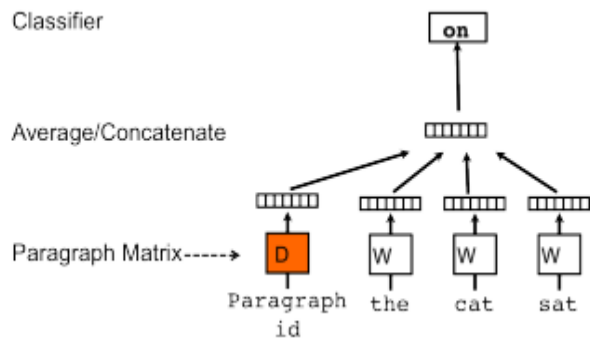
Proposed Method – 2) Brand2vec Parameter Optimization

Word2vec 방법은 다양한 parameter가 존재함

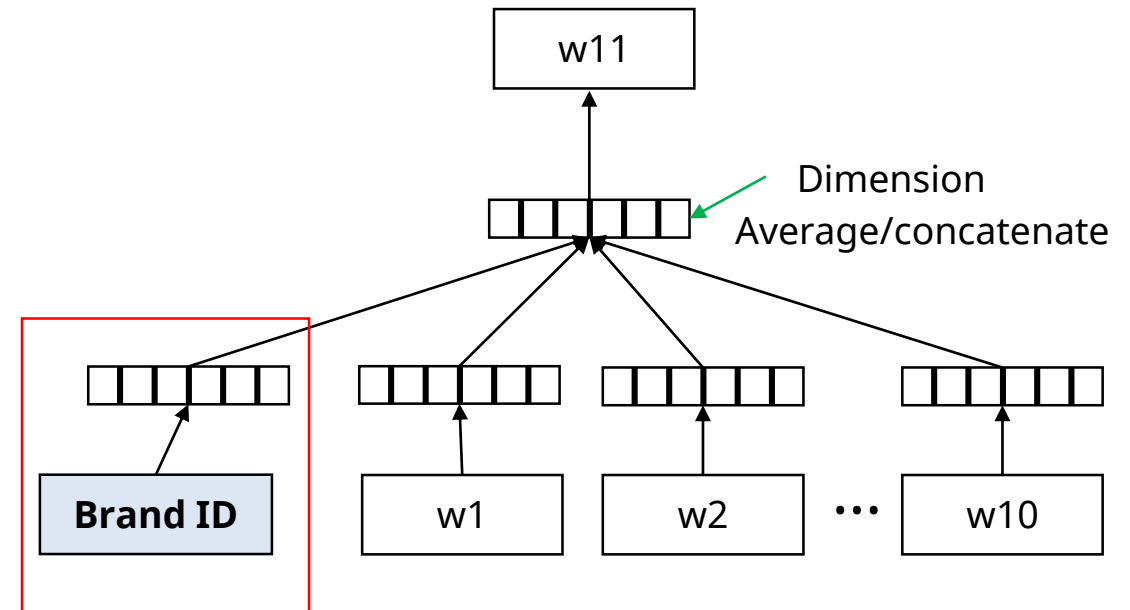
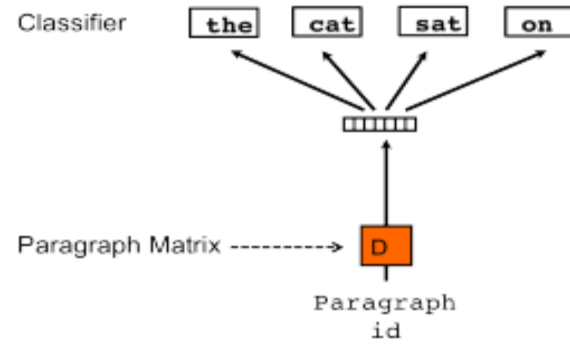
특히, word dimension, window size, training epoch 가 민감한 parameter

- **Dimension**
- **Windows**
- **Training epoch**
- Methods (PV-DM/PV-DBOW)
- Concatenate/average
- Negative Sampling, Subsampling of Frequent words 유무

distributed memory (PV-DM)



distributed bag of words (PV-DBOW)



3

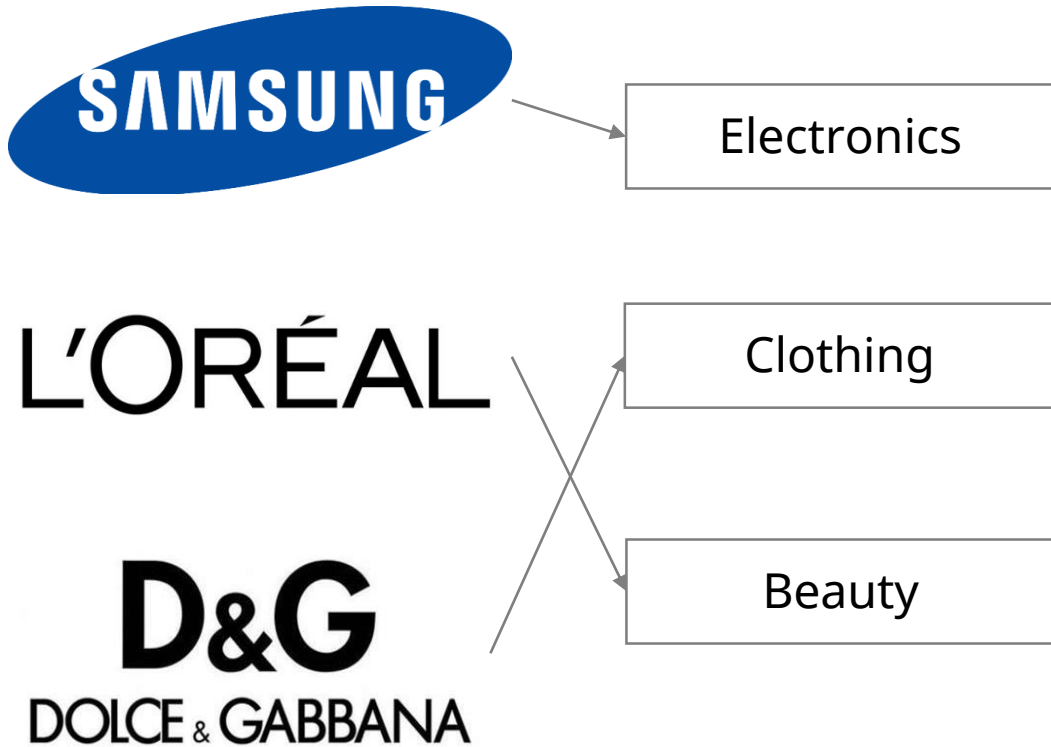
Proposed Method – 2) Brand2vec Parameter Optimization

Brand Vector 가 잘 embedding 되었는지 확인하기 위해서 아래와 같이 category 를 분류하는 문제를 제안함.

그러나, Brand 개수는 전체 문서 개수나, word 개수에 비해 매우 적기 때문에 브랜드 정보를 잘 학습한다고 해서, 단어나 문서 정보를 잘 포함하고 있는지는 알 수 없음. 따라서, 아래 2 가지 경우에 대해 실험함.

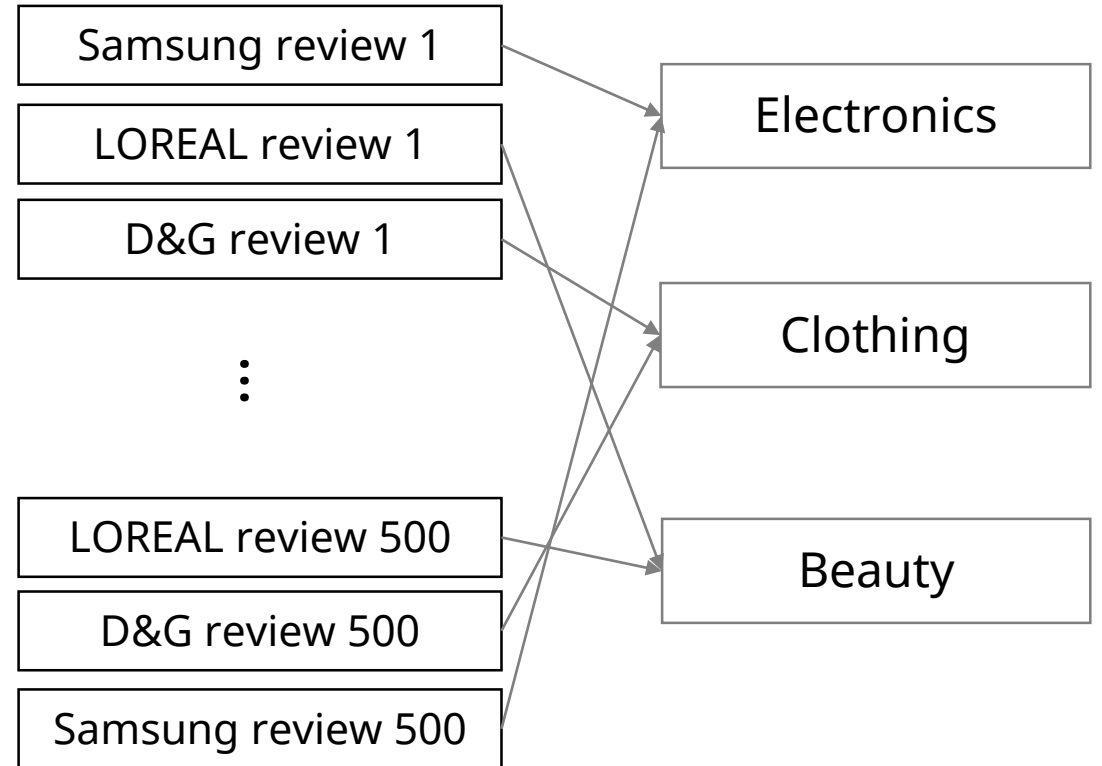
실험 1

Brand Vector 를 기준



실험 2

Document Vector 를 기준



4

실험 결과 - 1) Dataset Description

Amazon Review Data[4] 중 Electronics 카테고리의 2012년도 이후 약 3백만개의 review, 3억개에 달하는 token 들에 대해 분석을 실시함. 총 9,557개의 브랜드 중 10개 미만의 review 가 있는 브랜드는 제외하였으며,

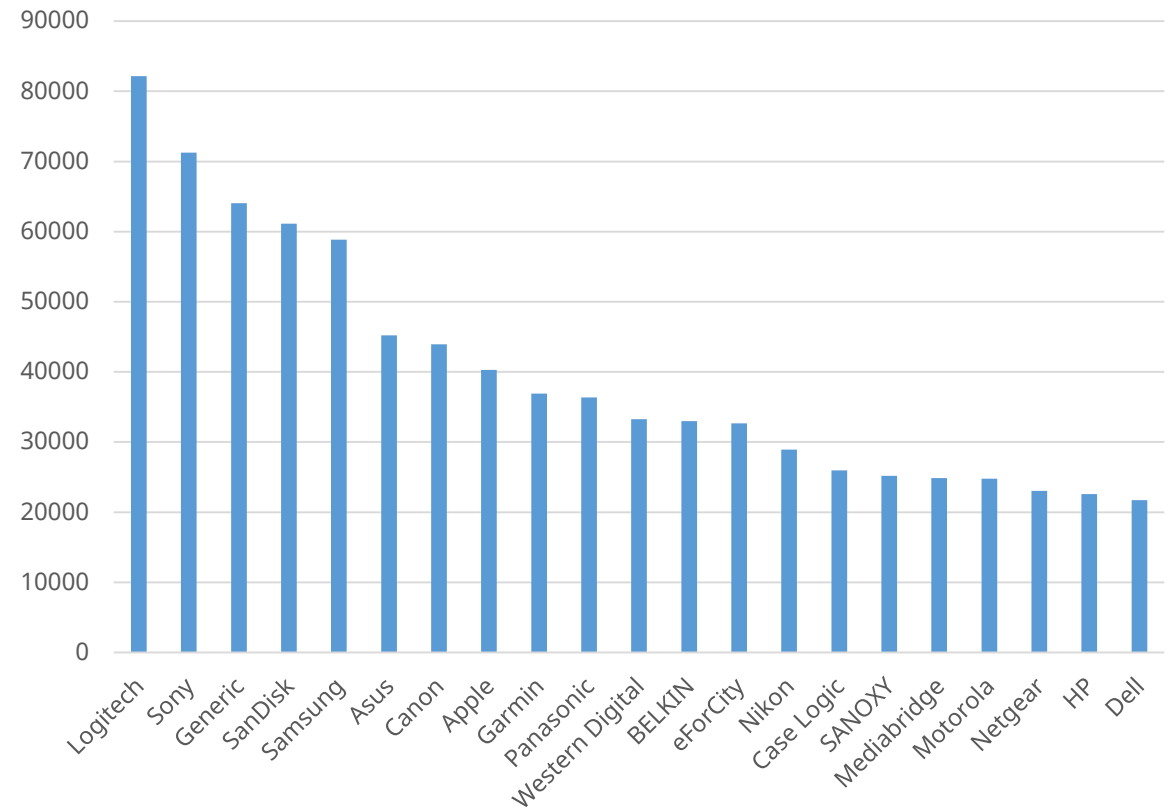
Logitech, Sony, Generic 브랜드 순으로 review가 많이 달려 있는 것을 알 수 있다.

구분	
2012년 이후 Electronics 카테고리의 review	5,566,912 reviews
Review가 없거나, brand 정보가 없어서 제외	2,959,904 reviews
Review가 10개 이상인 Brand 만 선택	2,944,904 reviews
Total tokens	218,288,099 token
Number of unique words	886,768 words

Year	Num of Reviews
2012	611,669
2013	1,405,001
2014	954,708

총 Brands	9,557
Review 10개 이상인 Brands	<u>5,079</u>

Number of Reviews by Brands



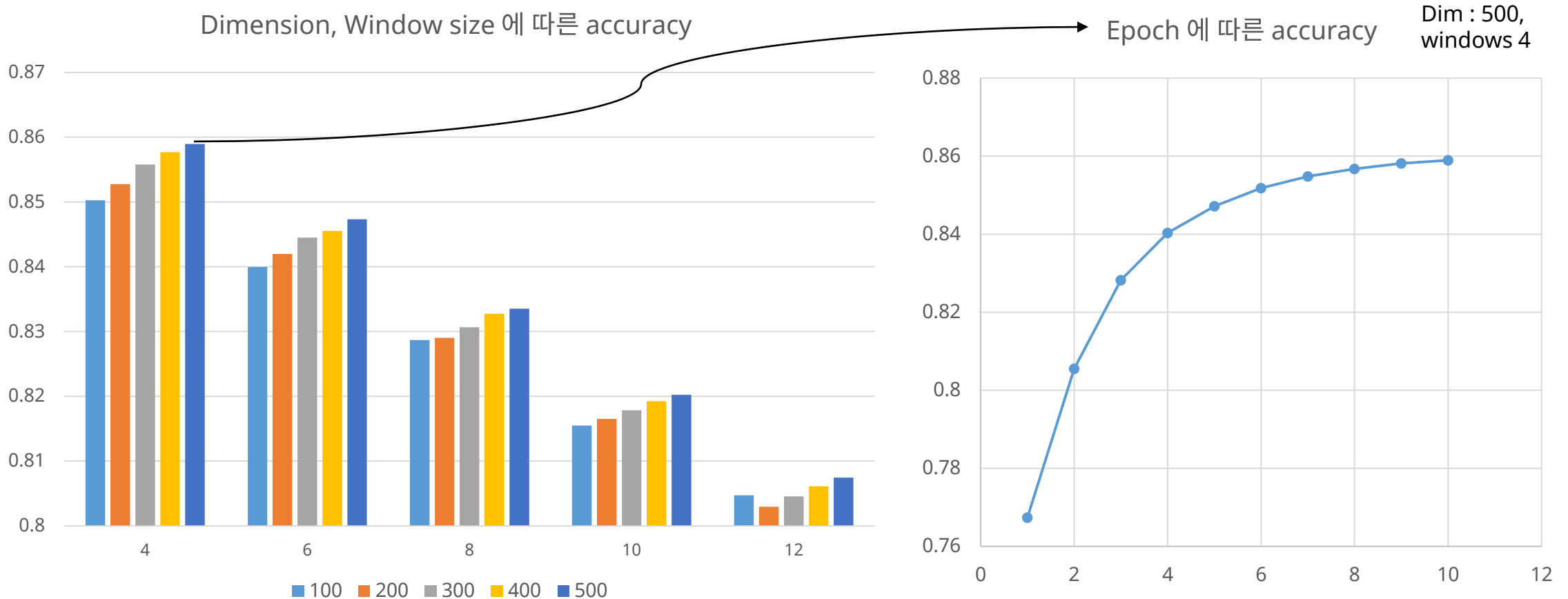
4

실험 결과 - 2) Parameter search

실험 1은 기존의 Doc2vec 방법을 이용하여 각 review 문서를 학습시킨다.

학습된 document vector 들을 training/test set으로 나눠서 training set으로 logistic regression 모델을 학습한다.

각 window size, dimension, epoch 에 따라 test accuracy 를 확인한다.



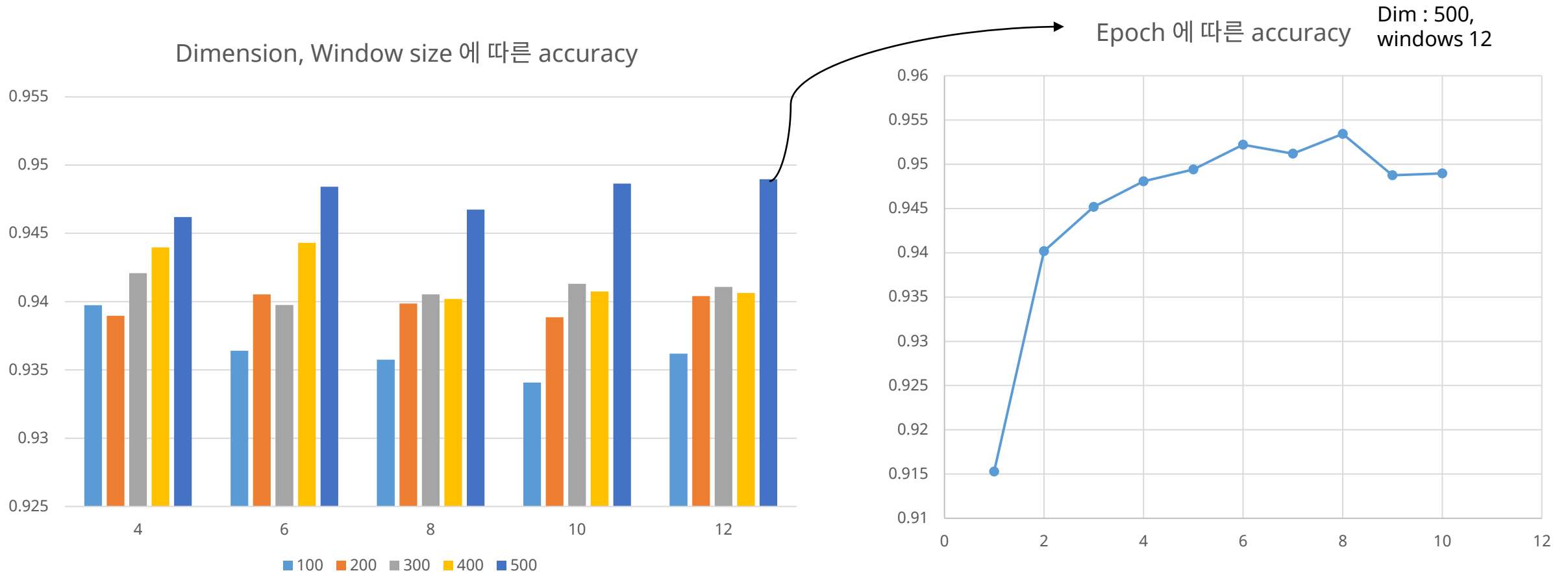
4

실험 결과 - 2) Parameter search

실험 2는 제안한 Brand2Vec 방법을 이용하여 각 Brand Vector를 학습시킨다.

학습된 Brand Vector 들을 training/test set으로 나눠서 training set으로 logistic regression 모델을 학습한다.

각 window size, dimension, epoch 에 따라 test accuracy 를 확인한다.



4

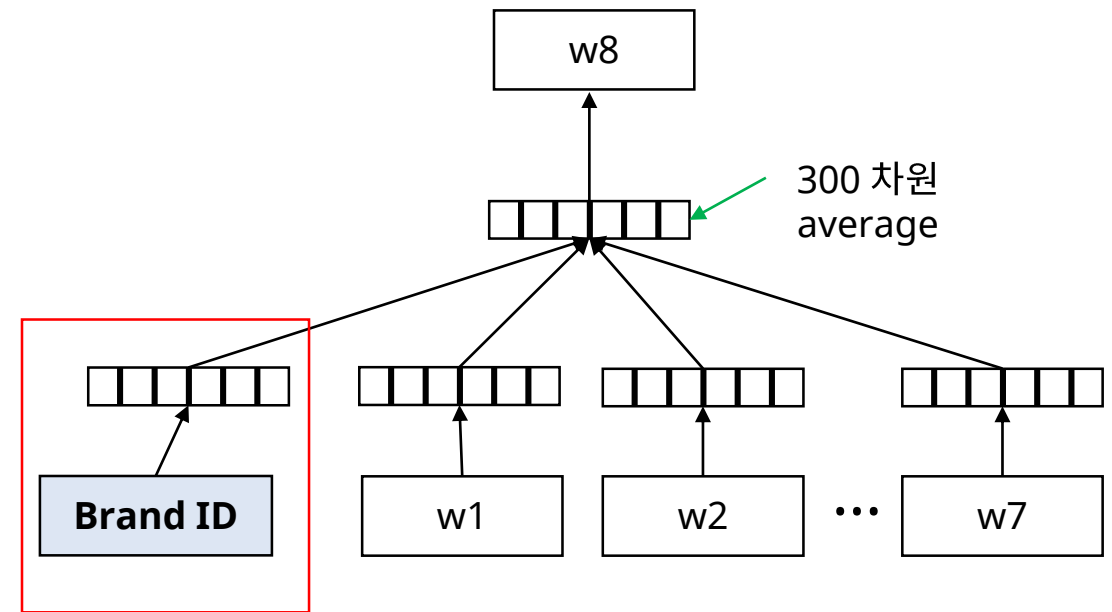
실험 결과 - 2) Parameter search

1, 2 모두 동일하게 dimension 이 500일 경우 category 분류 성능이 가장 좋았다.

그러나, Document 단위로 학습할 경우 window size가 작을 경우 성능이 좋았으며,

Brand vector 의 경우는 클 수록 학습이 잘 되었다. 따라서, 두 가지를 고려하여 Dimension 500, window 8, epoch 10

- Parameter setting
 - 1) Brand vector, Word vector dimension : 500
 - 2) Window : 8
 - 3) Epoch : 10
 - 4) Model : Distributed Memory model
 - 5) 300번 이하로 등장한 단어는 제거
 - 6) Average / concatenation 방법 중 average
- Preprocessing
 - 1) #,?,@ 등 특수문자 제거
 - 2) 숫자 제거
 - 3) 소문자



4

실험 결과 - 3) Application #1 Product positioning

Modeling 결과를 확인하기 위해 대표적으로 Samsung, Apple 의 Brands vector에 대해서 다른 Brand Vector, 그리고 단어 Vector 들을 각각 cosine similarity를 기준으로 정렬하면 아래와 같다.

Brand Vector 끼리 비교했을 때는 비슷한 브랜드가 등장하고, 단어와 비교했을 때는 각 Brand의 제품들이 등장함.

Samsung Brand Vector vs 다른 Brand Vector			Samsung Brand Vector vs Word Vectors			Canon Brand Vector vs 다른 Brand Vector			Canon Brand Vector vs Word Vectors		
Rank	Brand Vector	Similarity	Rank	Word Vectors	Similarity	Rank	Brand Vector	Similarity	Rank	Word Vectors	Similarity
1	brand_Acer	0.599127412	1	samsung	0.311653763	1	brand_Nikon	0.837055743	1	canon	0.429237217
2	brand_Toshiba	0.554059386	2	samsung's	0.283994943	2	brand_Focus Camera	0.794900835	2	zoom	0.407715142
3	brand_Proscan	0.502482474	3	smart	0.251906753	3	brand_Fujifilm	0.777616978	3	dslr	0.406417727
4	brand_Lenovo	0.496127099	4	google	0.249304563	4	brand_Sigma	0.764016926	4	telephoto	0.389285982
5	brand_Sharp	0.493932307	5	-inch	0.234910056	5	brand_Pentax	0.762909949	5	nikon	0.376861036
6	brand_HP	0.493476391	6	ativ	0.231575415	6	brand_Tamron	0.761495411	6	slr	0.376529872
7	brand_TCL	0.466564476	7	lg	0.229704186	7	brand_Tokina	0.709402978	7	bodies	0.362603426
8	brand_Dell	0.465381593	8	series	0.227229744	8	brand_Rokinon	0.704099953	8	megapixel	0.35245049
9	brand_VIZIO	0.463649154	9	plasma	0.225709692	9	brand_SSE	0.690469563	9	macro	0.347912639
10	brand_Kocaso	0.456790328	10	smarttv	0.224222615	10	brand_Vivitar	0.674572706	10	point-and-shoot	0.339259535
11	brand_Seiki	0.451489329	11	viera	0.218730122	11	brand_Olympus	0.671126723	11	minolta	0.337206662
12	brand_Hannspree	0.439872414	12	dlp	0.213553518	12	brand_Bower	0.670254111	12	full-frame	0.334055722
13	brand_Le Pan	0.434715688	13	microsdxc	0.205569863	13	brand_Leica	0.663152575	13	low-light	0.330319881
14	brand_Pipo	0.430985391	14	multitasking	0.202972621	14	brand_Zeikos	0.660996616	14	cannon	0.327950805
15	brand_Asus	0.427316636	15	dual-core	0.20019044	15	brand_Bower Camera	0.654028356	15	-mm	0.326378137
16	brand_OCZ	0.426803976	16	snappy	0.200056091	16	brand_Holga	0.638572991	16	wide-angle	0.324263781
17	brand_TabletExpress	0.4194749	17	quad-core	0.19897674	17	brand_Ricoh	0.6347543	17	panoramic	0.318859756
18	brand_Marquis	0.416073591	18	random	0.194532454	18	brand_Raynox	0.632844329	18	lens	0.316940278
19	brand_Double Power	0.409678906	19	bd	0.194449708	19	brand_Elite Brands Inc	0.622957945	19	body	0.312601537
20	brand_Kaser	0.407244563	20	android	0.191939622	20	brand_HeroFiber	0.615471661	20	nex	0.309334666

4

실험 결과 - 3) Application #1 Product positioning

Brand2Vec 방법을 통해 representation이 잘 되었다면 각각의 Vector는 각 Brand가 생산하는 제품들의 Property를 포함하고 있을 것이다.

아래와 같이 Computer, Earphone, Camera 라는 단어 Vector에 대해 각각 가장 가까운 Brand Vector를 찾아보았다.

$$(\overrightarrow{computer} + \overrightarrow{desktop})/2$$

Rank	Brand	Cosine Sim
1	Dell	0.141820
2	StarTech	0.122915
3	SIB	0.103958
4	HP	0.103505
5	Cooler Master	0.099601

$$(\overrightarrow{earphone} + \overrightarrow{headphone})/2$$


Rank	Brand	Cosine Sim
1	Sennheiser	0.211692
2	Monster	0.187543
3	JVC	0.145613
4	Monoprice	0.134583
5	Bose	0.123318

$$(\overrightarrow{camera} + \overrightarrow{cameras})/2$$


Rank	Brand	Cosine Sim
1	Canon	0.199109
2	Nikon	0.187253
3	Neewer	0.141008
4	Case Logic	0.077866
5	Panasonic	0.074816



Sennheiser HD 598
by Sennheiser
\$161.85 ~~\$349.00~~ Prime
Get it by **Friday, Nov 20**
More Buying Choices
\$156.45 new (76 offers)
\$100.91 used (7 offers)



Monster Diamond Tears
by Monster
\$219.99 ~~\$349.95~~ Prime
Only 1 left in stock - order soon.
More Buying Choices
\$194.00 new (11 offers)
\$130.99 used (5 offers)
See newer model of this item



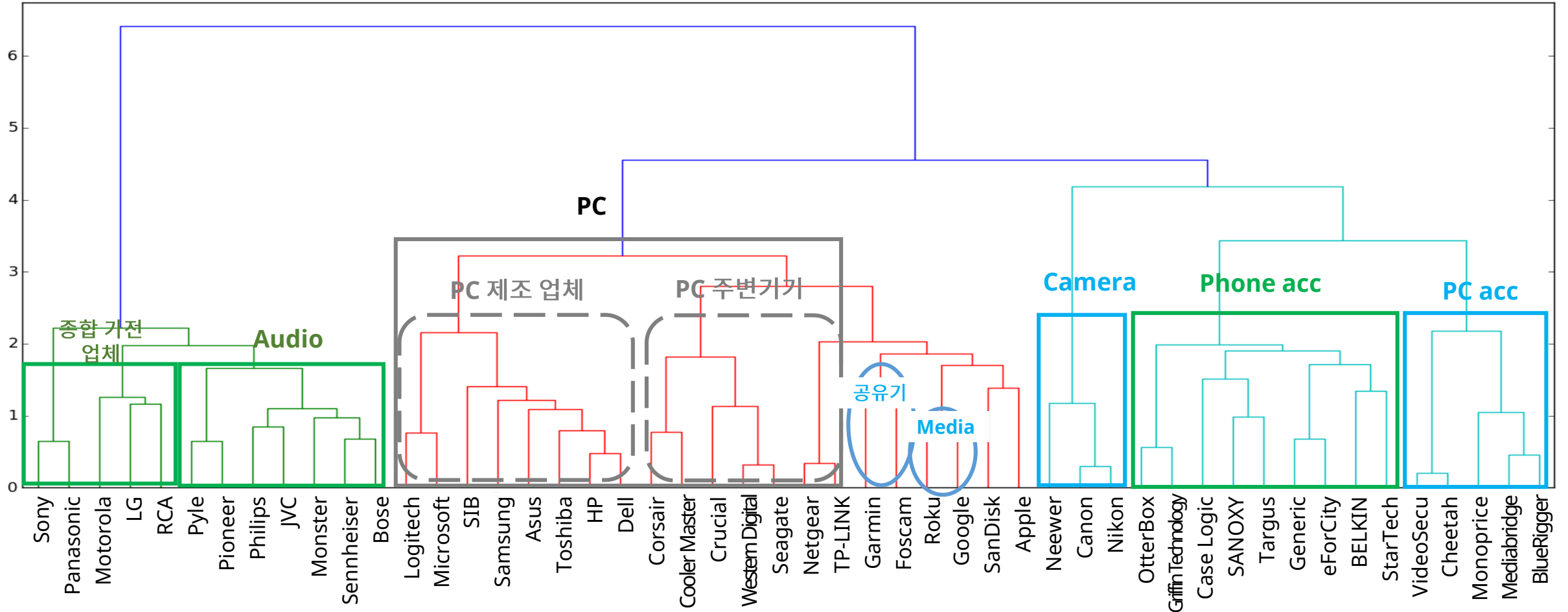
See more choices
Neewer TT560 Flash Speedlite

4

실험 결과 - 3) Application #1 Product positioning

Review 수가 많은 상위 50개 브랜드를 선정하여 Agglomerative hierarchical clustering 실시함.

각 브랜드 vector의 cosine similarity를 계산해 pair-distance matrix를 만들고, Ward's minimum variance method 사용하였다. Canon, Nikon과 같이 비슷한 제품군의 브랜드끼리 가깝게 위치한 것을 확인할 수 있다.

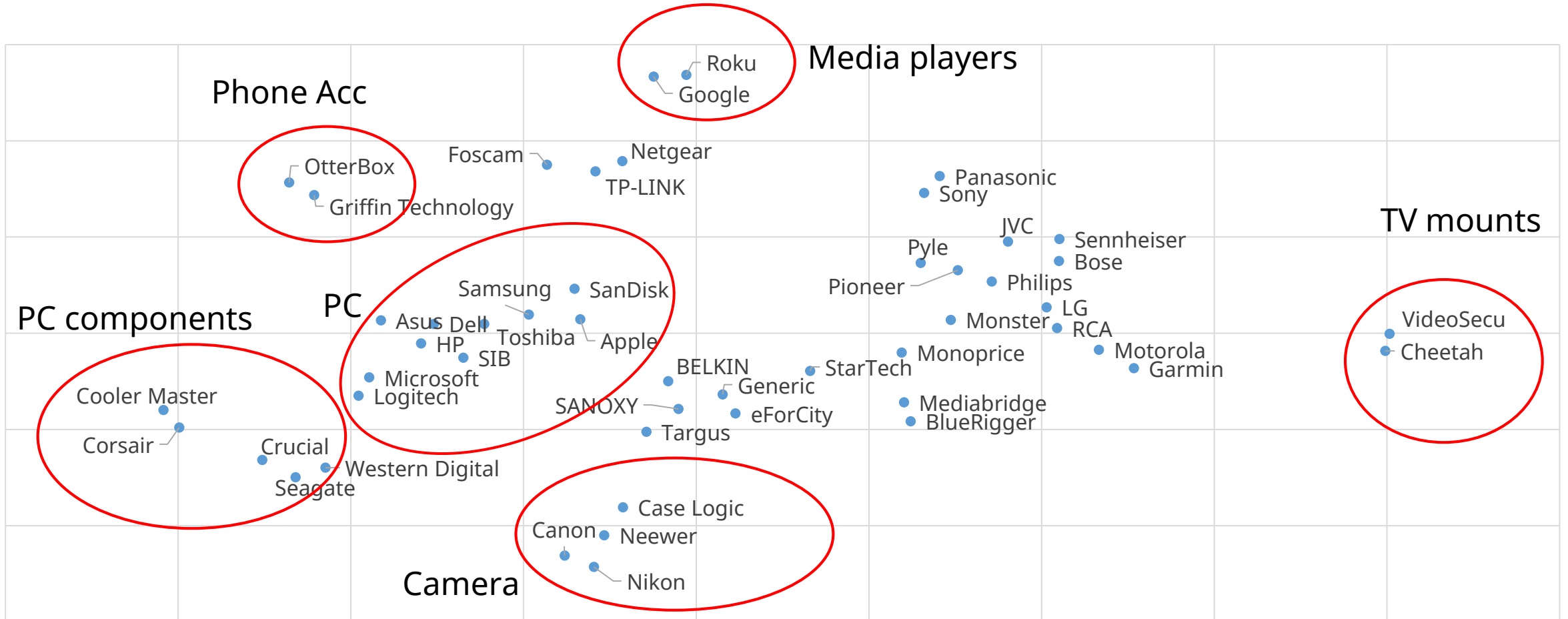


4

실험 결과 - 3) Application #1 Product positioning

아래는 50개 Brands 의 Brands Vector를 T-SNE를 이용하여 시각화 한 것이다

Dendrogram 결과와 같이 비슷한 브랜드들이 묶이는 것을 통해 Brand2Vec 가 Brand 의 특성을 반영했다고 볼 수 있다.



4

실험 결과 - 3) Application #2 Keyword extraction

빈도 기준의 방법은 주관성이 많이 개입되기 때문에 재현이 어렵다.

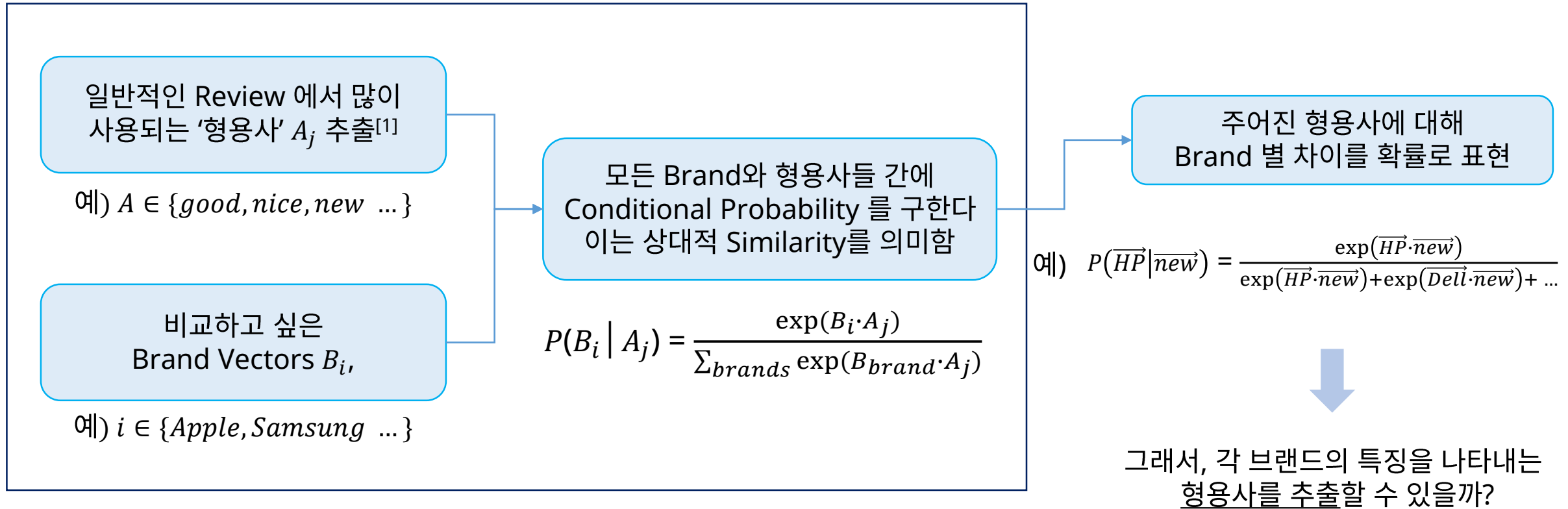
아래 표는 Samsung review 에서 많이 나온 단어들이는데, 각각이 무의미한 단어들이 뿐만 아니라 브랜드 간의 차이도 알 수가 없다.

Samsung		Apple		Microsoft	
word	freq	word	freq	word	freq
'the'	248605	'the'	126988	'the'	91585
'i'	176480	'i'	102026	'i'	62185
'and'	143593	'it'	85896	'it'	51693
'to'	139232	'and'	78219	'and'	48569
'it'	138997	'to'	75939	'a'	47302
'a'	127862	'a'	63404	'to'	46632
'is'	90798	'is'	47089	'is'	34210
'this'	77955	'my'	38751	'this'	25846
'for'	68564	'this'	38393	'for'	23818
'of'	60744	'for'	37519	'of'	22477
'my'	60073	'of'	32447	'my'	18969
'with'	53027	'with'	28079	'that'	18556
'that'	51316	'that'	27412	'on'	17431
'in'	48394	'you'	25156	'with'	16796
'on'	46018	'in'	24011	'mouse'	16560
'you'	43621	'on'	23682	'keyboard'	16430
'have'	41533	'have'	23463	'you'	15231
'was'	40660	'was'	22945	'in'	15154
'not'	40227	'apple'	22576	'but'	14456

4

실험 결과 - 3) Application #2 Keyword extraction

같은 단어라도 각 '브랜드' 에서 쓰이는 문맥이 다르기 때문이기 때문에 특정 단어의 브랜드에 대한 설명력을 $P(Brands|word)$ 를 통해 알아 볼 수 있다. 이 값이 크면, 해당 브랜드와 상대적으로 거리가 가깝다고 해석할 수 있다.



[1] 3백만개의 Review 중 3만개만 Random Sampling 하였으며, NLTK를 이용해 Part of Speech tagging을 실시

4

실험 결과 - 3) Application #2 Keyword extraction

예를 들어, 위에서 실시한 Desktop 제조업체 9개를 이용하여 $P(B_i | A_j)$ 를 구하면 각 브랜드 별로 특징을 나타내는 단어들을 뽑을 수 있다. 그 중 대표적으로 Samsung, Apple, Microsoft 의 예시가 아래와 같다.

그 단어들의 문맥상 쓰임새를 파악하기 위해 해당 브랜드의 Review에 대해서 PMI^[1]가 높은 단어들 혹은 원문을 확인하였다

Samsung		
word	Distance	freq
sharp	1	881
direct	1	362
interested	1	237
local	1	640
glad	1	1035
forth	0.999997	183
slim	0.999996	709
enable	0.999995	198
interesting	0.999986	146
final	0.999986	195
want	0.999981	5575
pull	0.999973	331
crucial	0.999921	226
confident	0.999455	80
live	0.999167	731
thin	0.998927	1060
available	0.998615	1893
important	0.998227	728
electronic	0.996423	263

Apple		
word	Distance	freq
air	1	2144
classic	1	706
cute	1	118
magnetic	1	89
proprietary	1	121
popular	1	80
compatible	1	400
magic	1	392
versatile	1	95
similar	1	384
white	1	587
handy	1	324
short	1	524
sturdy	1	221
stronger	1	60
substantial	1	38
heavier	1	126
impossible	1	143
protective	1	158

Microsoft		
word	Distance	freq
vertical	1	99
stiff	1	165
closer	1	112
comfortable	1	2034
traditional	1	195
key	1	2669
natural	1	786
mechanical	1	225
couch	1	139
ergonomic	1	1403
responsive	1	558
soft	1	360
harder	1	158
smooth	1	581
sensitive	1	308
uncomfortable	1	235
regular	1	558
love	1	2991
easier	1	508

[1] window size = 3, 10번 미만으로 나온 word는 제거함

4

실험 결과 - 3) Application #2 Keyword extraction : Samsung

Samsung Brand에 대해 나온 키워드를 살펴보면, Samsung 3d Active Glasses 와 같은 Samsung 고유한 제품을 나타내는 단어가 높게 나오는 것을 알 수 있다.

이를 통해 브랜드 벡터가 제품의 특징을 포함하고 있음을 알 수 있다.

Samsung	
word	Distance
active	1
vibrant	0.9994
dynamic	0.99793
picky	0.99677
fat	0.99618
higher	0.99408
pic	0.99394
song	0.99333
popular	0.99311
blurry	0.99226
pull	0.99085
write	0.98579
different	0.98198
special	0.9786
upper	0.97353
dish	0.96545
music	0.96043
better	0.94718
worse	0.94411
ridiculous	0.9336
professional	0.8956

glasses, 3d

Samsung 3d Active Glasses



much cheaper than any other active 3d glasses

Network, TV

Dish Network



I should mention that my signal provider is dish network

Photographer, reviews

4

실험 결과 - 3) Application #2 Keyword extraction : Apple

'Extra' 라는 단어가 높게 나온 것을 통해, 사용자가 다양한 옵션에 대해 많이 고민하고 있음을 알 수 있었다.

비슷한 맥락으로 'Protective' 라는 단어가 높게 나온 것을 보면, Apple 제품의 내구성에 대해 issue 가 있음을 알 수 있었다.

Apple	
word	Distance
air	1
previous	0.9999636
adhesive	0.9993592
magic	0.9992156
white	0.9987635
greatest	0.9968496
latest	0.995961
larger	0.9953194
sorry	0.9914955
dry	0.9900548
heavier	0.9893667
protective	0.987897
recent	0.9759585
monthly	0.9743298
tired	0.9676014
stronger	0.9642975
substantial	0.9606566
clearer	0.953744
extra	0.9512706
safe	0.9501784
thermal	0.9471483

Trackpad, mouse ..

Case, plastic, cover

cash, spend

Extra (다양한 옵션)
 2 Choose a capacity:
 Now available with up to 128GB of storage¹

16GB ¹ \$199.00 Available to ship: In Stock	32GB ¹ \$249.00 Available to ship: In Stock	64GB ¹ \$299.00 Available to ship: In Stock	128GB ¹ \$399.00
---	---	---	--------------------------------

I recommend buying the or gb and spending the extra money gb

Protective case, cover(내구성)

recommend a protective carry case



4

실험 결과 - 3) Application #2 Keyword extraction : Microsoft

Microsoft 브랜드는 타 브랜드에 비해 keyboard 제품의 특징이 두드러지며, XBOX 라는 게임기를 만드는데 그러한 특징을 잘 나타내는 키워드가 추출 됨을 확인 할 수 있다.

Microsoft	
word	Distance
ergonomic	1
flat	0.9999853
harder	0.9999626
natural	0.9999315
neat	0.9998434
visual	0.9998108
solar	0.9997959
traditional	0.9996345
sweet	0.9995185
comfortable	0.9968041
uncomfortable	0.9967529
yellow	0.9964272
magnetic	0.9957358
optional	0.9948526
included	0.9942972
full	0.9924303
unreliable	0.992372
usable	0.9891648
virtual	0.9870451
hot	0.9856131
convenient	0.9855615

Sculpt, Keyboard

Elite, positioning

Hot keys

Ergonomic Products



Microsoft Sculpt Ergonomic Mouse
by Microsoft

\$33.80 ~~\$59.95~~ Prime
Get it by **Friday, Nov 20**

More Buying Choices
\$32.25 new (135 offers)
\$15.89 used (21 offers)

Xbox Controller 'Elite'



5

Discussion - 1) 빈도 기반 방법과 비교

기존에 마케팅에 활용되던 대부분의 Text mining 기법들은 Frequency 기반의 방법임. 그러나, 이러한 방법은 빈도수가 적은 단어는 무시하거나, 혹은 무의미한 단어들 많이 등장하게 됨.

반면, Brand Vector 방법은 빈도가 적더라도 유의미한 단어를 추출할 수 있고, 객관성을 확보함.

	장점	단점
Frequency based	<ul style="list-style-type: none">문서 수가 상대적으로 적은 text 에 대해서도 분석 가능직관적	<ul style="list-style-type: none">무의미한 단어들 많이 등장브랜드의 특징을 나타내는 단어를 추출하기 어려움많은 수작업이 필요하기 때문에 주관성이 개입될 가능성이 높음
Brand Vector approach	<ul style="list-style-type: none">브랜드 간의 유사도를 파악할 수 있음등장 빈도가 적더라도 Brand의 특징을 구분 짓는 keyword를 추출할 수 있음수작업을 최소화하여 Reproducible 하기 때문에 객관성을 확보	<ul style="list-style-type: none">직관적이지 않음한 브랜드에 대한 Review 가 많은 경우에 유용함

5

Discussion - 2) 확장 가능성

Brand Vector 방법을 활용하면, 제품 뿐만 아니라

연예인, 정치인, 스포츠 스타 등의 Social Media를 활용한 vector representation 도 가능할 것이다.

이를 통해 비슷한 인물을 clustering 하거나, 유의미한 keyword를 추출하는 등에 적용할 수 있을 것으로 기대한다.

데이터 수집



Brand2vec 적용

[0.28, 0.56, 0.13, ... 0.91]



[0.32, 0.65, 0.18, ... 0.25]



[0.32, 0.65, 0.18, ... 0.25]

시각화 혹은 Keyword 추출

Brand2Vec 인물 적용 예시



헤리(예시)	
word	distance
단발	0.xx
애교	0.yy
이쁜	0.zz
귀여운	0.ww

김연아(예시)	
word	distance
홍보	0.aa
피겨	0.bb
아름다운	0.cc
응원	0.dd

6

결론 - 3) Summary & Future work

Word2Vec을 확장하여, 각 Brand를 하나의 Vector로 표현하였다. 이를 통해 브랜드 간의 상대적 거리가 유지되는지 확인하였고, conditional probability를 활용하여 Brand 별 특징적인 keyword를 추출할 수 있었다.

추가적으로 최적화된 parameter search 방법과, 시간에 따른 Brand의 변화 추이를 분석할 수도 있을 것이다.

- Summary

- 각 브랜드별 Review 정보를 모두 반영하여 각 브랜드를 하나의 Vector로 representation 함
- Brand Vector의 다양한 활용방안을 제시함
 - Hierarchical Clustering, Positioning Map을 통해 유사 브랜드끼리 묶이는지 확인함
 - Brand의 Product Property가 유지되는지 확인함
 - Conditional Probability를 활용한 Brand 별 유의미한 단어 추출함
- 이러한 방법을 통해 Brand2Vec의 유용성을 확인하였으며, 향후 연예인, 정치인 등 다양한 분야에도 활용할 수 있을 것으로 기대

- Future work

- Electronics 카테고리 이외에 패션, 식품 등 다양한 Brand에 대한 실험을 통해 정성적 검증
- 시간에 따른 review data를 활용하여 브랜드의 변화 추이를 분석
- Model에 Sentiment 정보를 반영하면 Sentiment analysis도 가능할 듯

7 Reference

-
- [1] Bengio, Yoshua, et al. "A neural probabilistic language model." *The Journal of Machine Learning Research* 3 (2003): 1137-1155.
 - [2] He, Wu, Shenghua Zha, and Ling Li. "Social media competitive analysis and text mining: A case study in the pizza industry." *International Journal of Information Management* 33.3 (2013): 464-472.
 - [3] Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." *arXiv preprint arXiv:1405.4053* (2014).
 - [4] McAuley, Julian, et al. "Image-based recommendations on styles and substitutes." *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015.
 - [5] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
 - [6] Mostafa, Mohamed M. "More than words: Social networks' text mining for consumer brand sentiments." *Expert Systems with Applications* 40.10 (2013): 4241-4251.
 - [7] Rong, Xin. "word2vec Parameter Learning Explained." *arXiv preprint arXiv:1411.2738* (2014).
 - [8] Sachan, Devendra Singh, and Shailesh Kumar. "Class Vectors: Embedding representation of Document Classes." *arXiv preprint arXiv:1508.00189* (2015).
 - [9] Tirunillai, Seshadri, and Gerard J. Tellis. "Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation." *Journal of Marketing Research* 51.4 (2014): 463-479.