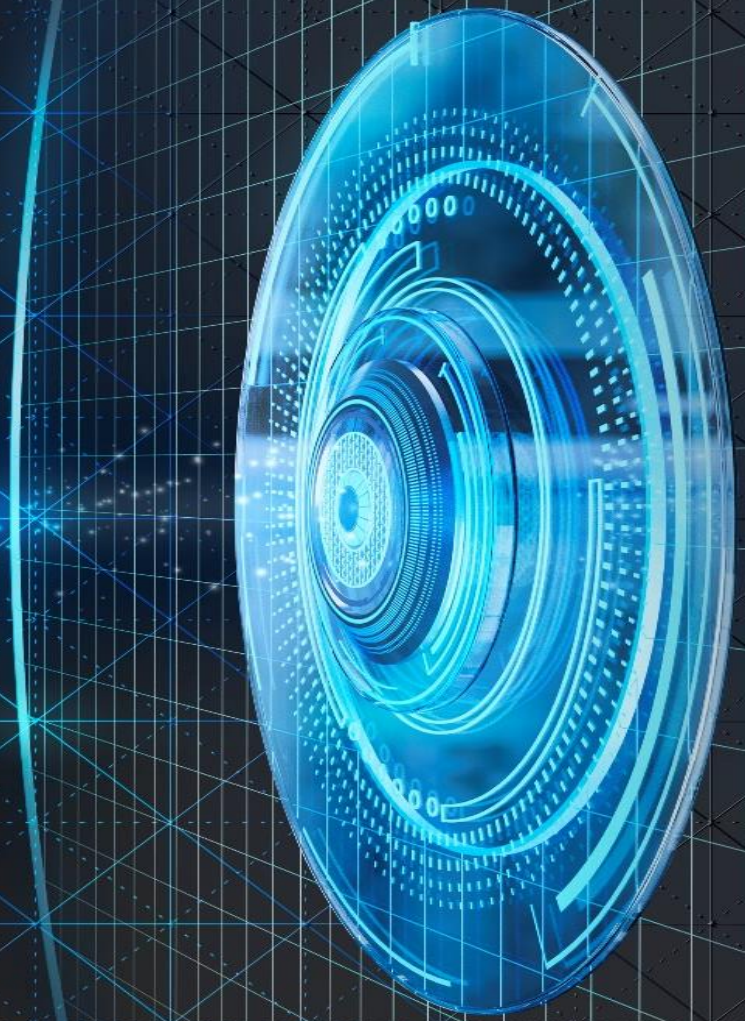
소셜 미디어 시계열 예측을 이용한 고객 니즈의 부상성 탐지

: 워드 임베딩, 네트워크 분석, LSTM 기반의 부상 키워드 탐지 방법

1. Introduction
2. Theoretical Background
3. Proposed Methodology
4. Case Study
5. Conclusion



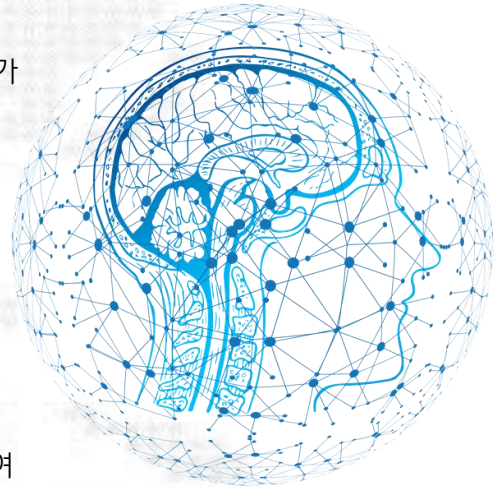
Propheta

김명섭*, 박영재, 이승주, 이권능(건국대학교 산업공학과)
최재은(건국대학교 경영학과)

*발표자 msub0310@gmail.com

1. INTRODUCTION

- 새로운 기회를 포착하는 것은 기업의 존속에 있어 가장 중요한 일 중 하나(Jeong and Yoon, 2016)이며, 연구 개발(R&D) 및 마케팅에 대한 접근 방식은 자연스럽게 고객의 요구를 분석하는 것에 중점을 두고 있음
- 기업은 고객에게 새롭거나 향상된 제품을 제공하기 위해 고객의 목소리에 특히 주의를 기울여야 함(Jeong *et al.*, 2019)
- 소셜 미디어는 제품에 대한 고객의 공개적인 의견을 교환하기 위한 매체가 되고 있으며, 소셜 미디어에 저장된 의견의 수가 폭발적으로 증가하는 추세(Wang *et al.*, 2012)
- 소셜 미디어는 사용 편의성, 속도 및 접근성으로 인해 사회의 대중적 담론을 빠르게 변화시키고 있으며, 환경 및 정치, 기술 및 엔터테인먼트 산업에 이르기까지 다양한 주제의 경향과 의제를 설정하고 있음(Brooks, 2015)
- 소셜 미디어에서 고객의 역동적인 행동 패턴, 즉 니즈(needs)를 파악하여 시장의 요구 사항을 예측하기 위해 통계학 및 기계 학습 기반의 다양한 방법들이 사용됨(Ohsawa, 2018)
- 이러한 방법을 이용하여 향후 부상할 고객의 요구 사항을 조기에 탐지하고 예측함으로써 기업은 경쟁 기업이 쉽게 모방할 수 없는 고객과의 관계를 구축할 수 있으며, 이는 기업의 경쟁력 강화로 이어짐(Jeong *et al.*, 2019)



- 본 연구에서는 소셜 미디어에 나타난 고객의 의견을 분석하고 예측하여 향후 파급을 일으킬 키워드를 탐지하는 방법을 제안
- Reddit 사이트에서 subreddit이 home automation에 속하는 328,010건의 고객 의견을 수집한 후 Jeong *et al.* (2018)의 방법 일부를 이용하여 워드 임베딩, 단어 네트워크 분석을 통해 키워드의 중요도를 산출
- 이후 LSTM, ARIMA, Prophet을 사용하여 키워드의 중요도를 예측, 키워드의 현재 중요도와 키워드의 미래 중요도를 이용하여 부상성을 계산
- 마지막으로 키워드의 현재 중요도와 부상성을 이용하여 현재 중요도가 낮으나, 향후 중요해질 고객 니즈(키워드), 현재 중요도가 높고, 향후에도 더욱 중요해질 고객 니즈를 탐지
- 본 연구는 계량적인 방법으로 향후 중요도가 더욱 높아질 고객의 니즈를 탐지하였다는 점에서 기업의 경쟁력 강화를 위한 정량적인 의사 결정을 지원하는데 기여할 것으로 기대됨
- 또한 전문가의 판단에 비해 적은 시간, 낮은 비용으로 미래에 부상할 고객 니즈를 탐지할 수 있다는 점에서 다양한 산업에서 폭넓게 활용될 수 있을 것으로 기대됨

2. THEORETICAL BACKGROUND

Word2Vec

- 머신러닝 모델을 적용하기 위해 단어의 순서와 의미를 내포하는 벡터의 형태로 단어를 표현하는 기법
- 특정 임베딩 공간상에서 같은 맥락을 갖는 단어들이 가까운 거리를 가진다는 전제에서 출발하여 주어진 문장에 대한 문법적 해석이 가능하며 단어의 거리를 통해 다양한 분석이 가능함

Network Analysis

- 각 개체들 사이의 관계와 구조를 Node와 Edge로 표현하여 분석
- Closeness Centrality : 전체 네트워크에서 중심에 위치되는 정도를 파악하여 단어의 구조적 위치와 관계 빈도를 분석
- Co-occurrence : 문헌 집합에서 키워드를 추출하여 각 단어 쌍의 동시 출현 빈도를 통해 단어 간의 연관도를 구하여 시각적으로 표현하는 방법

LSTM

- RNN의 한 종류, 긴 의존 기간을 필요로 하는 학습을 수행할 수 있는 구조
- 일반적인 RNN 과는 달리 히든 노드에 셀 스테이트를 추가하여 각 셀의 정보를 다음 셀로 전달하는 조건을 조절할 수 있는 구조로 설계되어 대량의 데이터를 이용하여 학습 시 기울기 소실 문제를 해결함

3. PROPOSED METHODOLOGY

STEP 1

데이터 전처리 &수집

- 데이터 전처리

Google Big Query를 이용해 Reddit 사이트에서 40개월 동안의 게시글(post)와 댓글(review)를 수집

- 데이터 수집

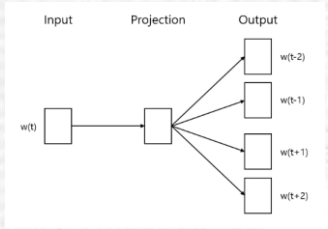
	doc_index	keyword	time
▶	87	understand	2015/12/31
	87	use	2015/12/31
	87	phillips	2015/12/31
	87	interface	2015/12/31
	87	is	2015/12/31
	87	just	2015/12/31
	87	lights	2015/12/31
	87	devices	2015/12/31
	87	directly	2015/12/31
	87	do	2015/12/31
	87	e	2015/12/31
	87	open	2015/12/31

게시글과 댓글의 텍스트를 띄어쓰기 단위로 분할 및 등록 월별로 분류

STEP 2

가중치 행렬 생성

- Word2vec 임베딩 및 가중치 적용



	ability	ac	access	...	zones
ability	1	0.87	0.73	...	0.12
ac	0.25	0.78	0.49	...	0.54
access	0.11	0.46	0.28	...	0.89
...	...	...	...	...	...
zones	0.27	0.39	0.97	...	0.08

Weight Matrix

Word2vec으로 생성된 임베딩 된 차원에서 키워드 간의 거리에 가중치를 적용하여 가중치 행렬 생성

STEP 3

키워드 중요도 산출

- 가중치와 동시 출현을 고려한 closeness centrality 계산

	ability	ac	access	...	zones
ability	1	0.87	0.73	...	0.12
ac	0.25	0.78	0.49	...	0.54
access	0.11	0.46	0.28	...	0.89
...	...	...	...	...	...
zones	0.27	0.39	0.97	...	0.08

Weight Matrix

X

	ability	ac	access	...	zones
ability	1	2	1	...	1
ac	8	1	2	...	8
access	9	2	1	...	5
...	...	...	...	...	...
zones	0	4	3	...	1

Co-occurrence Matrix

=

가중치 행렬에 동시 출현 행렬을 요소 별로 곱하여 단어 네트워크를 생성 후 계산된 closeness centrality를 키워드의 중요도로 간주

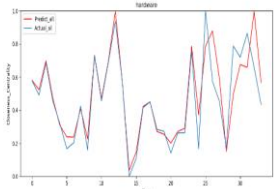
STEP 4

키워드 중요도 예측

- 중요도의 시계열 예측

keyword	2015_12	2016_1	2016_2	2016_3	2016_4
the	0.94785	0.94747	0.94544	0.94436	0.94262
i	0.90915	0.92061	0.90892	0.92035	0.90968
to	0.92611	0.93589	0.92749	0.92666	0.92595
a	0.92023	0.92486	0.92643	0.92783	0.92299
and	0.91309	0.91515	0.91328	0.906	0.91186
it	0.87854	0.8848	0.88595	0.88168	0.88831
you	0.80757	0.82749	0.81186	0.8177	0.81997
that	0.85093	0.85912	0.85421	0.86048	0.86016
is	0.85634	0.86346	0.85757	0.85986	0.86555
of	0.86851	0.8672	0.85773	0.86373	0.86398
for	0.84844	0.84282	0.84485	0.85305	0.84912

Keyword Importance over Time



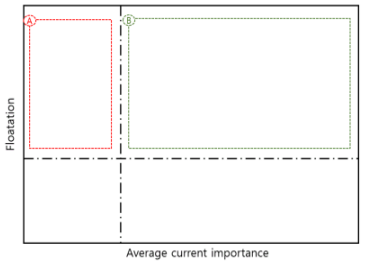
Time Series Prediction

시계열적인 키워드 중요도를 input으로 LSTM을 적용하여 키워드의 중요도를 예측

STEP 5

부상 키워드 탐지

- 미래에 파급을 일으킬 키워드 탐지



Keyword Positioning Map

키워드의 중요도와 부상성을 이용하여 keyword positioning map을 구성하고 미래에 파급을 일으킬 고객의 니즈를 탐지

사진 출처 : https://www.reddit.com/r/homeautomation/

3. PROPOSED METHODOLOGY

3.1 데이터 수집 및 전처리

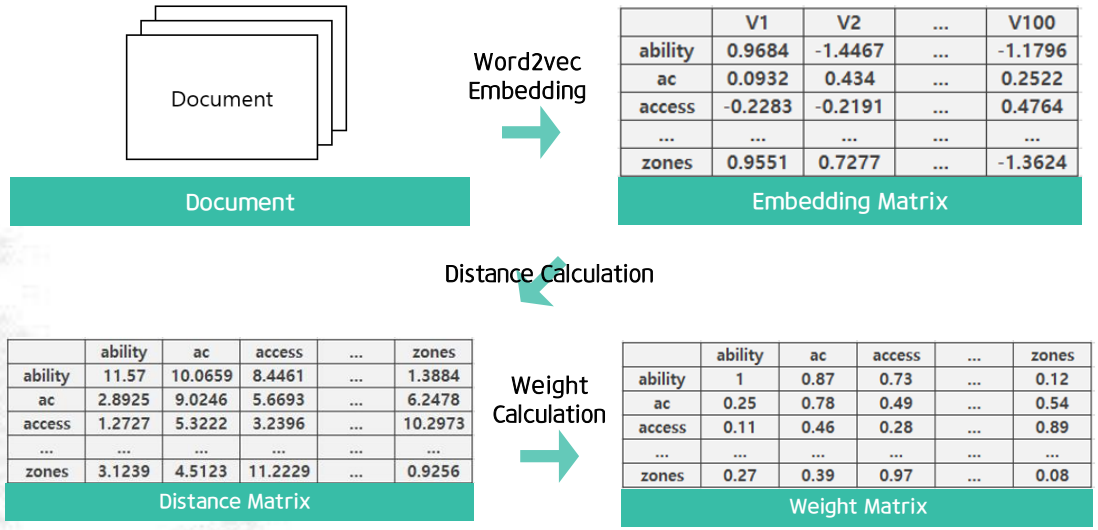
- 소셜 뉴스 웹사이트인 Reddit(https://www.reddit.com) 에서 데이터 수집
- Google Big Query를 활용하여 데이터 수집
- 게시글(post), 댓글(review), 작성일 등의 데이터 수집
- 수집된 게시글과 댓글을 띄어쓰기 단위로 분할

3.2 가중치 행렬 생성

- 수집하여 전처리 된 키워드에 Word2vec을 적용하여 저차원으로 임베딩
- 임베딩 된 차원에서 키워드 간의 거리를 계산하여 거리행렬 생성
- 계산된 거리에 가중치 식 (eq.1) 을 적용하여 키워드 간의 거리가 가까울수록 큰 값을 가지는 **가중치 행렬**을 생성(Jeong et al., 2018)

$$\omega_{ij} = \exp\left(-\frac{d(x_i, x_j)^2}{\sigma}\right) where\ node\ i\ and\ node\ j\ adjacent\ \ (eq.1)$$

$$d(x_i, x_j) : distance\ between\ node\ i\ and\ node\ j$$



	V1	V2	...	V100
ability	0.9684	-1.4467	...	-1.1796
ac	0.0932	0.434	...	0.2522
access	-0.2283	-0.2191	...	0.4764
...	...	...	...	...
zones	0.9551	0.7277	...	-1.3624

	ability	ac	access	...	zones
ability	11.57	10.0659	8.4461	...	1.3884
ac	2.8925	9.0246	5.6693	...	6.2478
access	1.2727	5.3222	3.2396	...	10.2973
...	...	...	...	...	...
zones	3.1239	4.5123	11.2229	...	0.9256

	ability	ac	access	...	zones
ability	1	0.87	0.73	...	0.12
ac	0.25	0.78	0.49	...	0.54
access	0.11	0.46	0.28	...	0.89
...	...	...	...	...	...
zones	0.27	0.39	0.97	...	0.08

3. PROPOSED METHODOLOGY

3.3 키워드 중요도 산출

- 한 문서 내에서 동시에 출현한 단어의 횟수를 세어 **동시 출현 행렬** 생성
- 3.2에서 생성한 가중치 행렬과 동시 출현 행렬을 요소 별 곱하여 **단어 네트워크** 생성
- 생성된 단어 네트워크에서 키워드의 **근접 중심성**(eq.2)을 **키워드의 유사도와 동시 출현을 고려한 중요도**로 간주(Jeong *et al.*, 2018)
- 근접 중심성은 0에서 1로 **정규화 된 값**을 사용

- $closeness\ centrality(x) = \frac{N}{\sum_y d(y,x)}\ (eq.2)$

$d(y,x)$: *shortest distance between node i and node j*

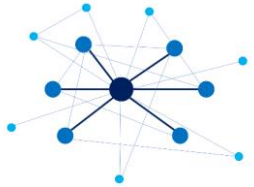
	ability	ac	access	...	zones
ability	1	0.87	0.73	...	0.12
ac	0.25	0.78	0.49	...	0.54
access	0.11	0.46	0.28	...	0.89
...	...	...	...	...	...
zones	0.27	0.39	0.97	...	0.08

Weight Matrix

element product

	ability	ac	access	...	zones
ability	1	2	1	...	1
ac	8	1	2	...	8
access	9	2	1	...	5
...	...	...	...	...	...
zones	0	4	3	...	1

Co-occurrence Matrix



Word Network

Closeness Centrality calculation

keyword	2015_12	2016_1	2016_2	2016_3	2016_4
the	0.94785	0.94747	0.94544	0.94436	0.94262
i	0.90915	0.92061	0.90892	0.92035	0.90968
to	0.92611	0.93589	0.92749	0.92666	0.92595
a	0.92023	0.92486	0.92643	0.92783	0.92299
and	0.91309	0.91515	0.91328	0.906	0.91186
it	0.87854	0.8848	0.88595	0.88168	0.88831
you	0.80757	0.82749	0.81186	0.8177	0.81997
that	0.85093	0.85912	0.85421	0.86048	0.86016
is	0.85634	0.86346	0.85757	0.85986	0.86555
of	0.86851	0.8672	0.85773	0.86373	0.86398
for	0.84844	0.84282	0.84485	0.85305	0.84912

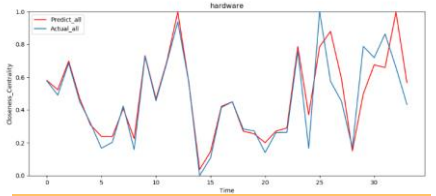
Keyword Importance

3.4 키워드 중요도 예측

- 시계열 예측 모델 **LSTM, ARIMA, Prophet**을 사용하여 keyword 중요도를 예측
- **Keyword 예측 값의 평균 MSE**(Mean squared error)**를 비교**하여 우수한 모델을 채택

keyword	2015_12	2016_1	2016_2	2016_3	2016_4
the	0.94785	0.94747	0.94544	0.94436	0.94262
i	0.90915	0.92061	0.90892	0.92035	0.90968
to	0.92611	0.93589	0.92749	0.92666	0.92595
a	0.92023	0.92486	0.92643	0.92783	0.92299
and	0.91309	0.91515	0.91328	0.906	0.91186
it	0.87854	0.8848	0.88595	0.88168	0.88831
you	0.80757	0.82749	0.81186	0.8177	0.81997
that	0.85093	0.85912	0.85421	0.86048	0.86016
is	0.85634	0.86346	0.85757	0.85986	0.86555
of	0.86851	0.8672	0.85773	0.86373	0.86398
for	0.84844	0.84282	0.84485	0.85305	0.84912

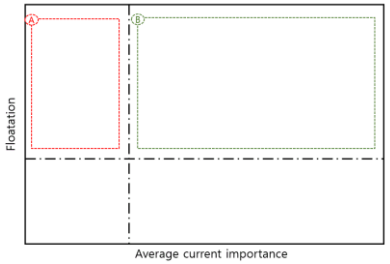
Keyword Importance



Keyword Importance Prediction

3.5 부상 키워드 탐지

- 수집하여 계산된 현시점까지 중요도의 평균을 **현재 중요도(Current importance)**로 간주
- 예측된 중요도의 평균을 **미래 중요도(Future importance)**로 간주
- **미래 중요도를 현재 중요도로 나누어 부상성을 계산**(eq.3)
- 현재 중요도와 부상성을 각 축으로 positioning
- 부상성이 1보다 높은 키워드를 대상으로 현재 중요도가 낮으나 향후 중요해질 키워드, 현재 중요도가 높으며 향후에 더욱 중요해질 키워드를 탐지



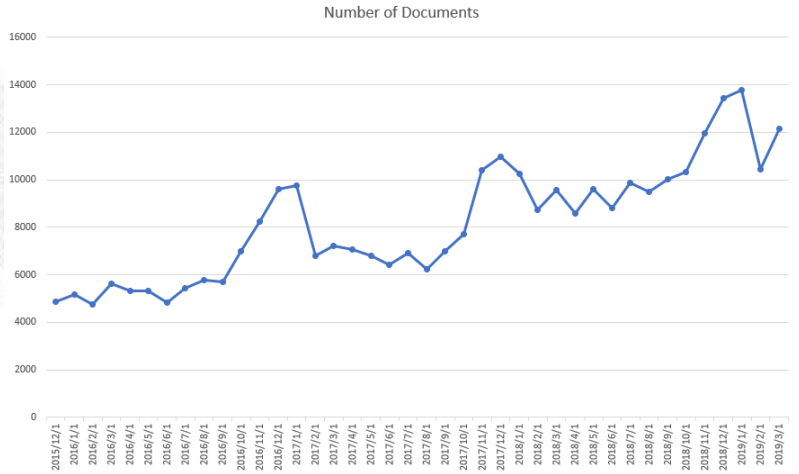
Keyword Positioning Map

- $floatation(keyword) = \frac{avg(current\ importance)}{avg(predicted\ importance)}\ (eq.3)$

4. CASE STUDY

4.1 데이터 수집 및 전처리

- 소셜 뉴스 웹사이트인 Reddit은 관심 영역인 subreddit으로 구성
- 본 연구에서는 subreddit 'Home automation'의 데이터를 사용
- 4차 산업혁명과 IOT(Internet of Things)에 대한 관심이 형성되고 고조되는 현시점에 Home automation(가정 자동화)에 관한 소비자의 의견을 분석하는 것은 시의성 있는 이슈
- 2015년 12월부터 2019년 3월까지 40개월 동안의 post 31,430개와 review 296,580개 총 328,010개의 데이터 수집
- 수집된 문서를 월별로 분류 후 띄어쓰기 단위로 분할
- 총 113,398개 키워드 사용



Number of Documents

4.2 가중치 행렬 생성

- 4.1에서 수집하여 전처리 된 데이터에 Python의 Gensim 모듈의 Word2Vec을 사용
- 100차원 공간에 임베딩
- Skip-gram 방식 사용
- 입력 단어의 앞뒤로 출력층의 개수를 의미하는 윈도우 크기 5로 설정
- 신경망이 충분히 수렴할 수 있도록 신경망 반복 횟수(epoch)를 1000번 적용
- 이후 분산 표상으로 100차원으로 나타낸 키워드 벡터들에 eq.1을 적용하여 가중치 행렬을 생성
- eq.1의 σ값은 키워드 간의 거리의 총 분산을 사용

word	D1	D2	D3	D4	...	D100
the	0.18221	-0.3244	0.20727	0.00164	...	0.24718
i	0.34751	-0.5808	0.18869	0.02078	...	0.09315
to	0.23468	-0.3072	0.08783	0.29094	...	-0.0356
a	0.08364	-0.3601	0.1453	0.02795	...	0.23075
and	0.09061	-0.2682	0.06932	0.01708	...	0.11882
it	0.26019	-0.4296	-0.0919	0.15387	...	0.23221
you	0.47966	-0.1108	-0.0606	0.27633	...	0.18317
that	0.46601	-0.3605	0.18801	-0.0008	...	0.07774
is	0.13819	0.04394	0.05231	-0.0494	...	0.26503
of	0.47528	-0.3041	0.45397	-0.1088	...	0.09361
for	0.30109	-0.2771	0.26444	-0.0819	...	-0.0638
in	0.45441	-0.5144	0.09808	-0.2102	...	-0.0449
on	0.15362	-0.0278	0.10363	0.17139	...	-0.2109
have	0.03611	0.01049	0.18929	-0.1902	...	0.15196
with	0.32786	-0.1338	0.0511	0.1514	...	-0.045
t	-0.0131	0.14472	0.32753	0.01785	...	0.42593
...	...	...	...	...	...	...

Part of Embedding Table

4. CASE STUDY

4.3 키워드 중요도 산출

- 4.1에서 전처리한 키워드를 이용해서 문서에서 동시 출현 행렬을 생성
- 4.2에서 생성한 가중치 행렬과 동시 출현 행렬을 요소 별 곱하여 단어 네트워크를 생성
- 생성된 단어 네트워크에서 키워드의 근접 중심성을 계산
- 근접 중심성은 중요한 노드일수록 다른 노드에 이르는 거리가 짧을 것이라는 가정을 기저에 두고 있음
- 계산된 근접 중심성을 키워드의 중요도로 간주
- 수집 기간인 2015년 12월부터 2019년 3월까지 1개월별로 모든 키워드에 대해 중요도를 계산
- 이후 “the”, “I”, “things” 등 기술적으로 유의하지 않다고 판단되는 단어를 연구자가 정성적으로 판단하여 제거
- 총 613개 선택된 키워드의 1개월 별 중요도를 4.4의 input으로 사용

keyword	2015_12	2016_1	2016_2	2016_3	2016_4	2016_5	2016_6	2016_7	...	2019_3
home	0.68936	0.69007	0.68524	0.69242	0.69812	0.6878	0.69717	0.68798	...	0.71096
switch	0.63194	0.63493	0.62876	0.62455	0.62047	0.60383	0.61283	0.6264	...	0.64768
hub	0.63531	0.63863	0.62211	0.63055	0.6352	0.60976	0.62276	0.61597	...	0.61595
control	0.65323	0.65906	0.65168	0.65026	0.6479	0.64725	0.65204	0.64604	...	0.63388
switches	0.61179	0.62358	0.61269	0.60916	0.61226	0.59073	0.59886	0.59648	...	0.62134
lights	0.63259	0.63785	0.61556	0.61744	0.62462	0.60904	0.61936	0.61129	...	0.62384
devices	0.64088	0.65722	0.63331	0.64466	0.64439	0.62318	0.61068	0.61702	...	0.64335
amp	0.58976	0.58282	0.58453	0.59212	0.5796	0.57012	0.58022	0.6025	...	0.65231
light	0.63754	0.62845	0.6042	0.59971	0.60831	0.60578	0.60555	0.60736	...	0.61942
house	0.64146	0.6456	0.63474	0.63891	0.63882	0.61881	0.62343	0.64186	...	0.65317
time	0.65466	0.65925	0.64407	0.64644	0.6642	0.63968	0.63348	0.64339	...	0.65956
system	0.65595	0.64568	0.64505	0.65766	0.65144	0.61989	0.63206	0.64356	...	0.62843
app	0.62657	0.61591	0.60087	0.61851	0.61392	0.61074	0.61482	0.60501	...	0.60872
...	...	...	...	...	...	...	...	...	...	...
deployed	0.49979	0.5024	0.50329	0.50209	0.49672	0.50085	0.50265	0.50175	...	0.50794

Part of Closeness Centrality of Keywords Over Time

home	automation	remote	connect	software	systems
switch	smartthings	wire	connected	voice	options
hub	bulbs	wall	button	wireless	installed
control	device	phone	link	internet	code
switches	assistant	echo	plug	wires	line
lights	wifi	tv	install	access	wired
devices	alexa	zigbee	homeseer	option	bridge
amp	sensors	wink	led	alarm	wiring
light	sensor	network	camera	homekit	tell
house	room	controller	bulb	price	trigger
time	run	lutron	garage	cheap	outlet
system	setup	security	cameras	battery	cost
app	motion	thermostat	lighting	video	caseta

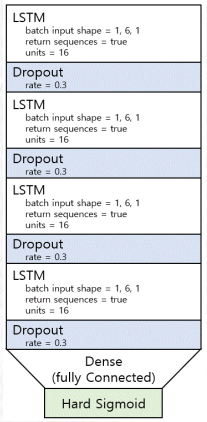
Part of Selected Keyword

4. CASE STUDY 4.4 키워드 중요도 예측

(Long Short-Term Memory) (Autoregressive Integrated Moving Average)

LSTM | ARIMA | Prophet

4.3에서 계산한 키워드의 중요도를 이용하여 1개월 단위 40개 과거 데이터로 총 12개월의 1개월 단위 키워드의 중요도를 예측



- Keras Sequential의 LSTM 이용
- Sequence의 길이를 의미하는 Time step = 6, batch size 1로 설정
- Train : Test = 75 : 25
- LSTM layer 4층, Optimizer adam, 총 epoch 1800으로 설정
- 각 layer마다 존재하는 Hidden node의 개수 16개로 설정, Loss MSE로 설정
- 각 키워드마다 모델 초기화

keyword	MSE
home	0.0178
switch	0.15143
hub	0.07157
control	0.04481
switches	0.0922
...	...
deployed	0.03786
average	0.07437

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Where:
- $g(t) = trend$
- $s(t) = seasonality$
- $h(t) = holiday$

- 시계열 분석 및 예측에서 가장 널리 사용되는 모델
- Autoregressive는 자기 회귀(p), Moving Average는 이동 평균(q)을 의미
- Non-Stationary 데이터를 Stationary 상태로 만들기 위해 관측치에 Difference(d)를 활용
- ARIMA 모형은 대체로 ARIMA(p, d, q)로 표현
- PYRAMID ARIMA Package의 Auto_ARIMA 활용
  : 하이퍼 파라메터 p, d, q를 데이터의 Stationary 등을 고려

Keyword	MSE
home	0.20215
switch	0.53033
hub	0.23493
control	0.19974
switches	0.15046
...	...
deployed	0.15671
average	0.1961

$$Y_t = c + \emptyset_1 Y_{t-1} + \cdots + \emptyset_p Y_{t-p} + \Theta_1 e_{t-1} + \cdots \Theta_q e_{t-q} + e_t$$

Where:
- $Y_t = the\ variable\ that\ will\ be\ explained\ in\ time\ t;$
- $c = constant\ or\ intercept;$
- $\emptyset = coefficient\ of\ each\ parameter\ p;$
- $\Theta = coefficient\ of\ each\ parameter\ q;$
- $e_t = Residuals\ or\ errors\ in\ time\ t.$

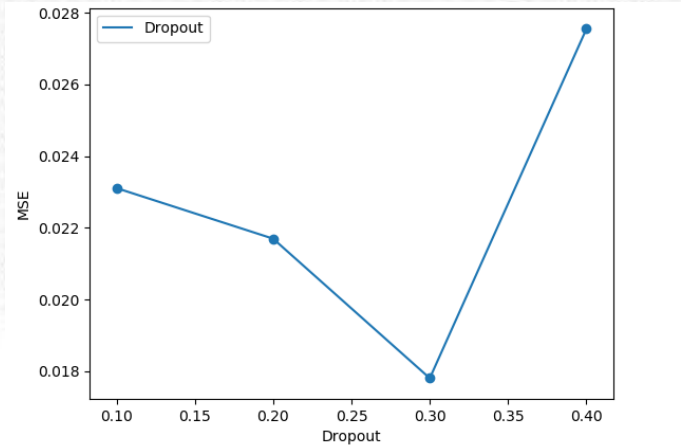
- 2017년 2월 23일 공개된 페이스북의 오픈소스 시계열 예측 라이브러리
- ARIMA 모델과의 차이점 : 차분 및 정규화 과정 필요 없음
- GAM Model(Generalized Additive Model) 기반
- 기본적으로 3개의 요소와 잔차로 구성
  1. g(t) : Trend(일정 기간 동안 유지되는 비 반복적 요소)
  2. s(t) : Seasonality(일, 월, 연 단위의 반복적 요소)
  3. h(t) : holiday(가끔 불규칙하게 영향 주는 요소)

Keyword	MSE
home	0.25267
switch	0.20366
hub	0.1855
control	0.21102
switches	0.19074
...	...
deployed	0.12078
average	0.14049

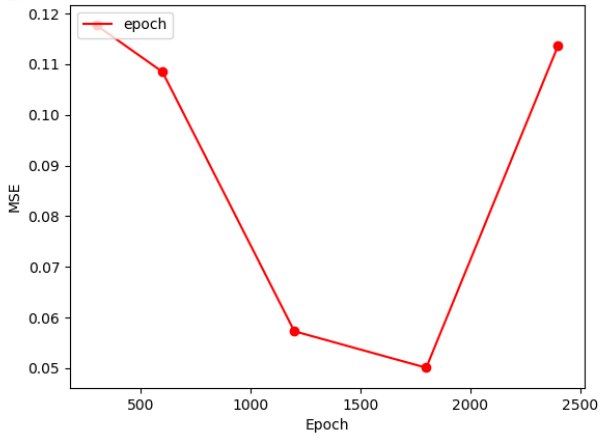
- 각 모델 별로 선택된 키워드에 대해 Test set의 MSE를 계산 후 평균 산출 -> LSTM의 평균 MSE가 0.07437로 가장 낮음 -> LSTM을 예측 모델로 채택

4. CASE STUDY 4.4 키워드 중요도 예측

- **하이퍼 파라메터**(Hyper parameter) : 신경망 학습을 통해서 튜닝 또는 최적화해야 하는 주 변수가 아니라 사람들이 선험적 지식으로 설정하는 변수
- LSTM 모델에서 하이퍼 파라메터를 변경해가며 최적의 하이퍼 파라메터를 탐지



MSE per Dropout of Keyword "Home"



MSE per Epoch of Keyword "Installed"

- Drop out : layer가 많아질 때 발생하는 과적합 문제를 방지하기 위해 layer에 포함된 weight 중 일부만 학습에 참여

- Epoch : 인공 신경망에서 전체 데이터 셋에 대해 한번 학습을 완료한 단위

본 연구에서 사용한 LSTM 모델에서는 Drop out이 0.3 일 때, Epoch이 1800일 때 가장 낮은 MSE를 보이므로 해당 하이퍼 파라메터를 채택

4. CASE STUDY

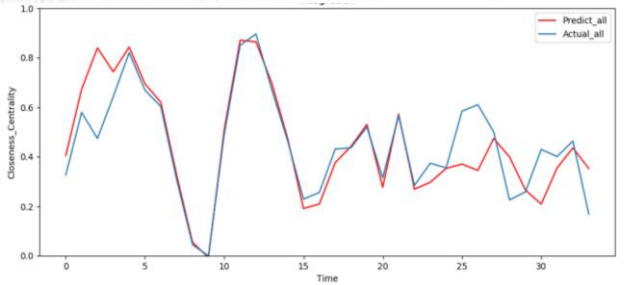
4.4 키워드 중요도 예측 - LTMS 학습 결과

keyword	MSE	keyword	MSE	keyword	MSE	keyword	MSE
home	0.01780486	bulbs	0.03483944	echo	0.11418891	homeseer	0.00351
switch	0.15143166	device	0.09239039	tv	0.03777805	led	0.08538214
hub	0.07156621	assistant	0.01824716	zigbee	0.23025021	camera	0.2750185
control	0.04481109	wifi	0.01883095	wink	0.03868524	bulb	0.15190964
switches	0.09220461	alexa	0.0148151	network	0.20186313	garage	0.03847465
lights	0.01761589	sensors	0.05531822	controller	0.03924454	cameras	0.02523675
devices	0.11463641	sensor	0.09377562	lutron	0.07459786	lighting	0.01822131
amp	0.48454216	room	0.09399323	security	0.00792748	software	0.01741684
light	0.08948141	run	0.10212176	thermostat	0.02401461	voice	0.04282109
house	0.09409	setup	0.23618026	connect	0.19099626	wireless	0.0764681
time	0.10329459	motion	0.01986042	connected	0.07008757	internet	0.06249675
system	0.02642409	remote	0.10352606	button	0.11828338	wires	0.02698357
app	0.17067042	wire	0.05477272	link	0.04471545	access	0.08771578
automation	0.03255284	wall	0.07895047	plug	0.15373916	option	0.05020492
smartthings	0.05255295	phone	0.12385262	install	0.10546882	alarm	0.03037074

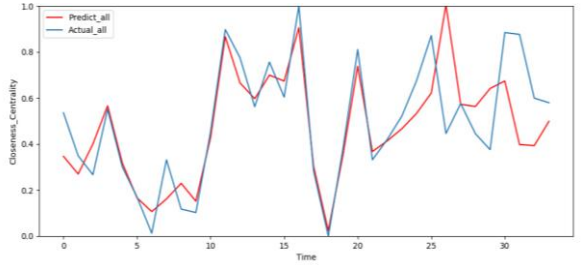
Part of MSE of Keyword

Predict
Actual

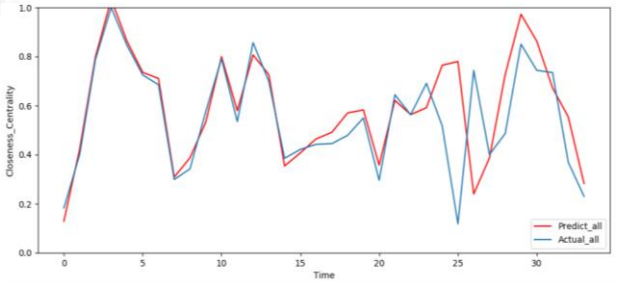
세로축: 중요도(Closeness centrality)
가로축: 시간(개월)



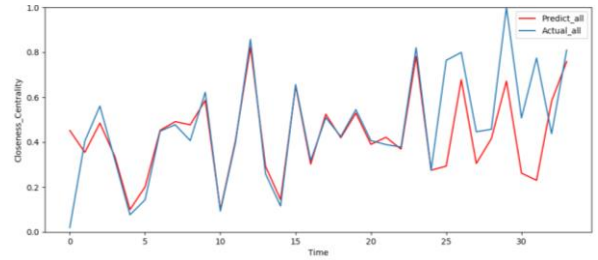
keyword "Integration"



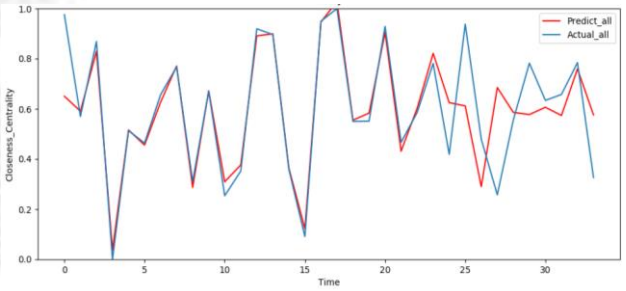
keyword "Display"



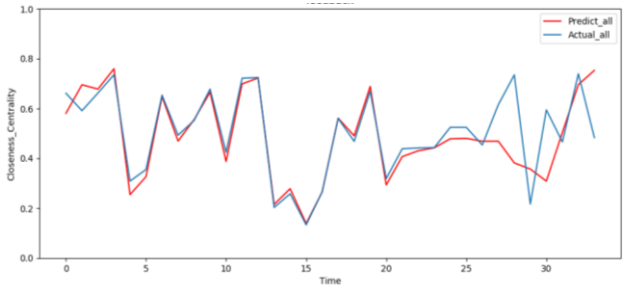
keyword "Bluetooth"



keyword "Network"

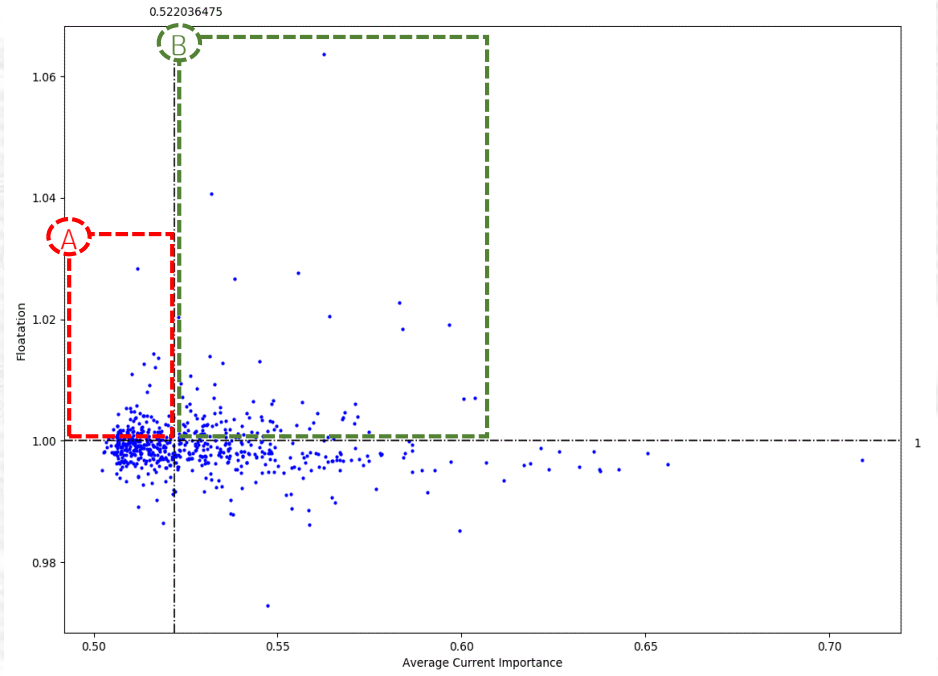


keyword "Delay"



keyword "Feedback"

4. CASE STUDY 4.5 부상 키워드 탐지



Keyword Positioning Map

- **X축** : 1개월 단위로 수집되어 계산된 총 40개월의 키워드 중요도 평균**(현재 중요도)**
- **Y축** : 예측된 12개월의 키워드 중요도 평균을 미래 중요도로 간주,
  미래 중요도를 현재 중요도로 나눈 값인 **부상성(Floatation)**

- 각 키워드의 **현재 중요도**와 **부상성**을 이용하여 **키워드 포지셔닝 맵**을 구성,
  x축의 경우 경계 값을 중앙값(median), y축의 경우 경계 값을 1로 선정

- 좌측 상단에 붉은색으로 표시된 A 구역의 키워드
  - 현재 낮은 중요도를 가지나 미래에 큰 폭으로 중요도가 상승할 키워드
  - **현재 큰 주목을 받지 않으나 향후 부상할 고객 니즈**로 판단

- 우측 상단에 녹색으로 표시된 B 구역의 키워드
  - 현재 높은 중요도를 가지며, 미래에 중요도가 더욱 상승할 키워드
  - **현재에도 주목받고 있으며 향후 더욱 부상할 고객 니즈**로 판단(Yoon, 2012)

- 총 114개 키워드가 A 구역에 위치
  총 106개 키워드가 B 구역에 위치

4. CASE STUDY 4.5 부상 키워드 탐지

keyword	floatation	keyword	floatation	keyword	floatation	keyword	floatation
customization	1.02831502	error	1.00418259	programmed	1.00250865	linking	1.00173135
lan	1.01433034	bug	1.00416566	communicates	1.00242058	notified	1.00167877
circuits	1.01363671	recognition	1.00407463	microphone	1.00240875	latency	1.0016484
maintenance	1.01263222	coverage	1.00406242	reconnect	1.0023799	smartthings	1.00158705
arlo	1.01214871	repeater	1.00404148	sw	1.00235627	cabinet	1.00158658
protection	1.01099093	traffic	1.00381006	homebridge	1.00222735	netflix	1.0015266
bedrooms	1.00909272	projector	1.00379006	siren	1.00220738	availability	1.00152305
bandwidth	1.00797692	adapters	1.00368134	micro	1.00214595	smartapp	1.00150937
warranty	1.00576672	cars	1.00356644	consumers	1.0021202	heaters	1.00150703
cabinets	1.00537371	charging	1.00345734	minimote	1.00193154	spotify	1.00150392
chromecasts	1.00535521	decora	1.003232	speeds	1.00192895	detecting	1.00145649
charger	1.00487636	versions	1.00314513	stereo	1.00191302	robust	1.00141589
isp	1.00484173	damage	1.00314076	upload	1.00188367	construction	1.00141114
bell	1.00460183	centralized	1.00291827	listening	1.00187256	engine	1.00139358
aux	1.00438445	package	1.00265246	itunes	1.00177302	rental	1.00130759

Keywords of Area A

- A 구역의 키워드를 부상성이 큰 순서로 나열
- “customization”, “lan”, “circuits”, “maintenance”, “protection” 등
- 해당 키워드들은 현재는 낮은 중요도를 가지나 향후 파급을 일으킬 것으로 예상되는 높은 부상성을 가짐

“I recently got a Phillips Hue Motion Sensor, and while I'm impressed by how well it works, I'm rather disappointed by the level of **customization**/automation. ”

“I’ve been using HS3 since 2014 and love it for it’s stability and **customization**.”

Customer Needs of Keyword “Customization”

- “최근에 Phillips Hue Motion Sensor를 구매했는데 잘 작동되어 좋은 인상을 받았지만 사용자 지정(customization)/자동화 수준에 다소 실망했다.”

- “ 2014년부터 HS3을 사용해 왔으며 안정성과 사용자 지정(customization)이 마음에 든다.”

keyword	floatation	keyword	floatation	keyword	floatation	keyword	floatation
camera	1.063719354	chromecast	1.008557094	option	1.004616325	settings	1.003310842
batteries	1.040615991	mac	1.007268724	wemo	1.004591425	outlet	1.003296849
replace	1.027635834	bathroom	1.007116059	wired	1.004580374	radio	1.00315914
remotely	1.026632446	setup	1.0070579	computer	1.00452459	cheaper	1.003118392
assistant	1.02268818	room	1.006947609	car	1.004281208	toggle	1.003051328
cameras	1.020446257	router	1.00662251	speed	1.004141929	wiring	1.002969196
card	1.02036392	ecobee	1.006532205	cable	1.003978392	tv	1.002828434
wifi	1.019047867	wife	1.006386376	echo	1.003926455	siri	1.002728739
alexa	1.018444186	reliable	1.006059207	tell	1.003797766	replacing	1.002711603
automations	1.013887802	connecting	1.006020214	sound	1.003791696	manual	1.002702266
area	1.01311721	installed	1.006013098	address	1.003693641	ios	1.002643448
moving	1.012776049	cables	1.005618274	action	1.003635734	wires	1.002550054
gateway	1.010658058	page	1.005517622	module	1.003605629	receiver	1.002545614
storage	1.009498815	ecosystem	1.004962104	wireless	1.003516689	firmware	1.0025041
mini	1.009333461	plugs	1.00464113	ceiling	1.003428765	bridge	1.002489427

Keywords of Area B

- B 구역의 키워드를 부상성이 큰 순서로 나열
- “camera”, “batteries”, “replace”, “remotely”, “wifi”등
- 해당 키워드들은 현재 높은 중요도를 가지며 향후 더욱 중요해질 것으로 예상되는 키워드

“Might need **camera**s after all but I wonder how much a camera costs that can see well enough in the dark.”

“I really love the Floodlight **camera**. I'd probably get one in a heartbeat if it had an available local video stream.”

Customer Needs of keyword "Camera”

- “결국 카메라(camera)가 필요할지 모르지만 어둠 속에서 충분히 볼 수 있는 카메라 비용이 얼마인지 궁금하다.”

- “플러드 라이트 카메라(camera)가 정말 마음에 든다. 로컬 비디오 스트림을 사용할 수 있다면 경우 기꺼이 구매할 것이다.”

해당 키워드들은 고객 니즈를 조기에 파악하고 예측하여 경쟁사 대비 우위를 점하기 위해 주시가 필요한 키워드일 것으로 판단

5. CONCLUSION

5.1 결론

- 연구 개발(R&D) 및 마케팅에 대한 접근 방식은 고객의 요구를 분석하는 것에 중점을 두고 있으며, 고객의 요구를 분석하기 위해 고객들이 자발적으로 활발한 담론을 형성하는 소셜 미디어를 이용할 수 있음

- 소셜 미디어에 나타난 고객의 니즈를 분석하고 예측함으로써 기업의 경쟁력 강화에 기여할 수 있음

- **본 연구에서는 정량적이고 시스템적인 방법으로 향후 부상할 고객의 니즈를 파악하는 방법을 제안**

- 소셜 뉴스 웹사이트인 Reddit에서 2015년 12월부터 2019년 3월까지 subreddit이 home automation에 속하는 총 328,010개의 데이터를 수집

- 이후 띄어쓰기 단위로 분할한 키워드를 이용하여 워드 임베딩을 수행하여 워드 임베딩 결과로부터 가중치 행렬을 산출한 뒤 문서로부터 생성된 동시 출현 행렬과 요소 별 곱

- 생성된 단어 네트워크의 근접 중심성을 키워드의 중요도로 간주 후 1개월을 단위로 40개월의 키워드 중요도를 이용하여 향후 12개월의 키워드 중요도를 예측

- 현재 키워드 중요도 평균과 부상성을 이용하여 키워드 포지셔닝 맵을 구성

- 부상성이 1보다 큰 키워드에 대해 현재 중요도가 낮으나 향후 파급을 일으킬 키워드(A 구역 키워드)와 현재 중요도가 높으며 향후에도 더욱 중요해질 키워드(B 구역 키워드)를 탐지

- **본 연구는 계량적인 방법으로 향후 중요도가 더욱 높아질 고객의 니즈를 탐지**하였다는 점에서 기업의 경쟁력 강화를 위한 **정량적인 의사 결정을 지원**하는데 기여할 것으로 기대됨

- 또한 전문가의 판단에 비해 **적은 시간, 낮은 비용으로 미래에 부상할 고객 니즈를 탐지**할 수 있다는 점에서 의의가 있음

5.2 향후 과제

- 본 연구에서 중요도 예측에 사용된 키워드는 연구자가 정성적으로 선택한 키워드이므로 향후 해당 분야의 전문가가 키워드를 판단하는 등의 작업을 거치면 더 정확한 결과를 얻을 수 있을 것으로 생각됨

- 본 연구에서 사용된 중요도는 0에서 1사이로 정규화 된 값으로 향후 발생할 부상 키워드의 **절대적인 파급력을 예측하는 데 한계**가 있음

- 마지막으로 본 연구는 특정한 기술 분야를 대상으로 시행되었으나 **향후 다양한 영역으로 확대될 잠재력**을 가지고 있음, 이는 향후 과제에서 다루어야 할 문제

REFERENCES

Brooks, S. (2015). Does personal social media usage affect efficiency and well-being?. Computer in Human Behavior. 46. 26-37

Jeong, J., Mo, K., Seo, S., Kim, C., Kim, H., Kang, P. (2018). Unsupervised document multi-category weight extraction based on word embedding and word network analysis: A case study on mobile phone reviews. Korean Institute of Industrial Engineers. 44(6). 442-451

Jeong, B., Yoon, J. (2016). Identifying product development opportunity using topic modeling and sentiment analysis. The Korean Operations Research and Management Science Society. 724-728

Jeong, B., Yoon, J., Lee, J. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. International Journal of Information Management. 48. 280-290

Kwak, D., Oh, S., Nasridinov, A., Park, Y. (2019). Tourist Data Analysis using ARIMA Model and LSTM. Proceedings of KIIT Conference. 605-607.

Ohsawa, Y. (2018). Graph-based entropy for detecting explanatory signs of change in market. The Review of Socionetwork Strategies. 12(2). 183-203₩

Oo, Z. Z., Phyu, S. (2019 February). Microclimate Prediction Using Cloud Centric Model Based on IoT Technology for Sustainable Agriculture. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 660-663). IEEE.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45.

Wang, X., Yu, C., Wei, Y. (2012). Social media peer communication and impacts on purchase intentions: A consumer socialization framework. Journal of Interactive
Marketing. 26(4), 198-208

Yoon, J. H. (2012). Detecting weak signals for long-term business opportunities using text mining of Web news. Expert Systems with Applications. 39(16). 12543-12550

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.