소비자 관점에서의 유용한 리뷰 분석

tripadviser의 호텔 리뷰를 중심으로

시스템 설계 3조 / 이현찬 교수님

B511023 김민아

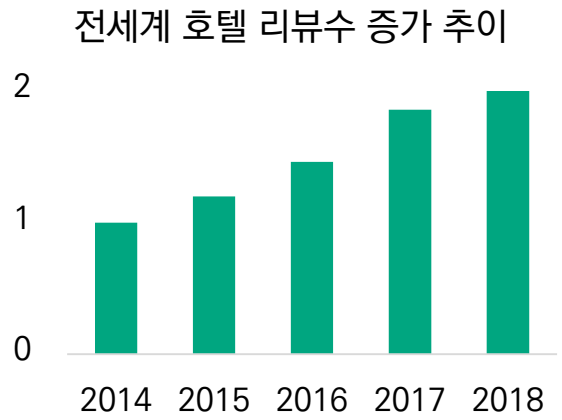
B631188 이아림

목차

1 프로젝트 배경 및 요약

리뷰의 중요성

리뷰 : 소비자에게 구매 여부 결정에 도움이 되는 중요한 정보
하지만 무성의하게 작성된 리뷰나 리뷰의 방대한 양으로 인해 소비자가 모든 리뷰를 확인하는 것은 비효율적
➔ 유용한 리뷰를 선별할 필요



기존 연구의 한계

기존 리뷰에 대한 연구는 대부분 소비자가 아닌 경영자를 위한 관점
: 서비스 품질 관련하여 고객만족 요인, 개선점, 마케팅 전략

➔ 소비자 관점에서 유용한 리뷰를 제공하는 시스템 필요

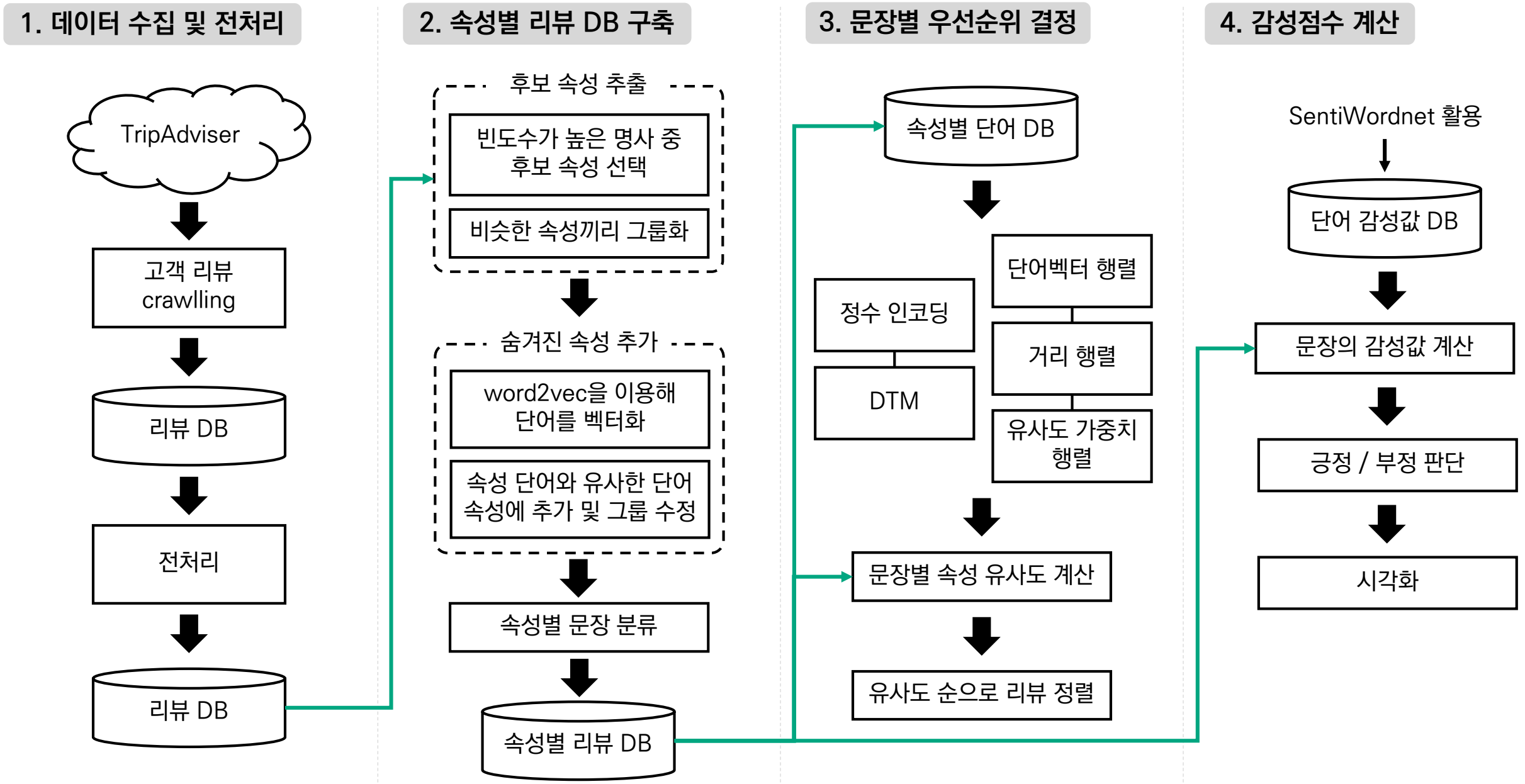
프로젝트 방향성

오피니언 마이닝 기술을 통해
사용자 리뷰를 자동으로 분류, 요약하여
잠재적 사용자에게 유용한 정보로 표현

소비자 관점에서 유용성이란?
: 방대한 리뷰 중에서 원하는 내용에 관련한 리뷰를 빠르게 얻는 것

- 원하는 내용 = 서비스 품질 속성과 관련
- 속성과 연관성이 높을수록 유용한 리뷰
- 속성별 리뷰의 긍정/부정 지표가 한 눈에 보이게 표시

2 프로세스


1. 데이터 수집 및 전처리
TripAdviser
고객 리뷰 crawlling
리뷰 DB
전처리
리뷰 DB
2. 속성별 리뷰 DB 구축
후보 속성 추출
빈도수가 높은 명사 중 후보 속성 선택
비슷한 속성끼리 그룹화
숨겨진 속성 추가
word2vec을 이용해 단어를 벡터화
속성 단어와 유사한 단어 속성에 추가 및 그룹 수정
속성별 문장 분류
속성별 리뷰 DB
3. 문장별 우선순위 결정
속성별 단어 DB
정수 인코딩
DTM
단어벡터 행렬
거리 행렬
유사도 가중치 행렬
문장별 속성 유사도 계산
유사도 순으로 리뷰 정렬
4. 감성점수 계산
SentiWordnet 활용
단어 감성값 DB
문장의 감성값 계산
긍정 / 부정 판단
시각화

3-1 데이터 수집

데이터 선택

- 외국인 방문 비율이 가장 높은 지역인 서울 〉 공항까지의 편리한 교통망, 우수한 문화관광자원으로 외국인들이 많이 방문하는 **마포구**
- 대표적인 숙박시설인 **호텔**
- 방한하는 외국인 대상이므로 가장 많이 사용하는 언어인 **영어 리뷰**

최종데이터

Index	Type	Size	Value
0	list	4	['location' , 'near' , 'station' , 'convenience']
1	list	2	['nice' , 'service'

내부 list의 각 요소에 하나의 문장을 이루는 단어들이 있는 2중 list 형태

분석 시 대부분 문장 단위로 시행하기 때문에 하나의 리뷰가 아닌

하나의 문장을 기준으로 전처리 (총 6187개의 리뷰 문장)

전처리 후 단어의 총 개수가 84451 -〉 38740 개로 축소

리뷰 데이터 web crawling

trip adviser에서 마포구에 위치한 13개 호텔 1570개의 영어 리뷰 크롤링

1. 리뷰 페이지 변화에 따라 url에서 변화하는 부분 확인

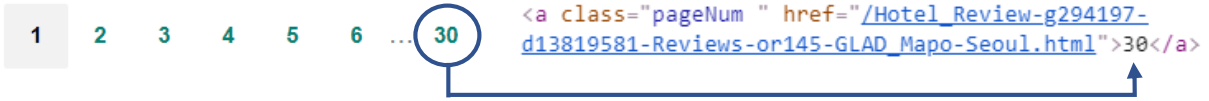
: 한 페이지에 5씩 증가 -〉 반복문 사용

tripadvisor.co.kr/Hotel_Review-g294197-d13819581-Reviews-or5-GLAD_Mapo-Seoul.html#REVIEWS

tripadvisor.co.kr/Hotel_Review-g294197-d13819581-Reviews-or10-GLAD_Mapo-Seoul.html#REVIEWS

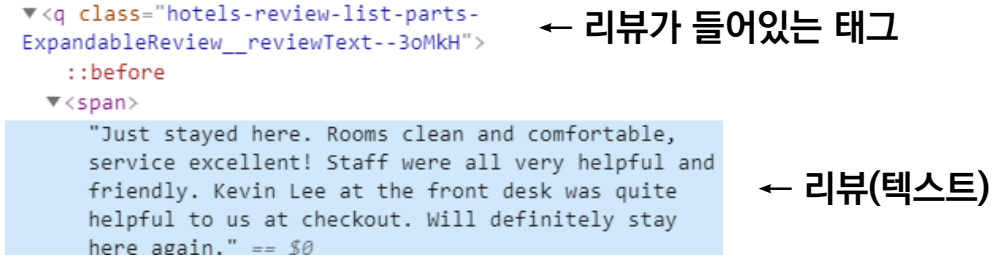
2. maxpage 추출하여 리뷰 데이터 크롤링에 필요한 url 목록 만들기

: maxpage 까지 5씩 증가하는 url 목록



3. 각 url에서 리뷰 데이터 추출

: url 별로 findAll 함수를 이용하여 해당 class 값을 가진 태그 q를 모두 뽑고 반복문을 이용하여 span 태그에 들어있는 텍스트 추출



← 리뷰가 들어있는 태그

← 리뷰(텍스트)

3-2 데이터 전처리

전처리 과정 현재 가지고 있는 데이터에서 노이즈 데이터를 제거하여 분석결과의 정확도를 높이기 위하여 전처리가 필요

토큰화	한 사용자가 쓴 리뷰 전체 〉 문장으로 나누기 〉 단어로 나누기
영어만 남기기	영어 외의 다른 언어 및 숫자, 문장부호 제거
소문자로 변경	대문자 〉 소문자로 통일
축약어 대체	isn't 〉 is not과 같이 축약된 단어를 대체
불용어 제거	분석에 무의미한 대명사, be동사, 전치사 등의 속성과 관련 없는 불용어 제거
짧은 단어 제거	2글자 이하의 단어들 또한 대부분 불용어에 해당
표제어 추출	단어 원형 복원

" Bed and pillows were very comfortable! "

↓

['Bed' , 'and' , 'pillows' , 'were' , 'very' , 'comfortable' , '!']

↓

['Bed' , 'and' , 'pillows' , 'were' , 'very' , 'comfortable' , ' ']

↓

['bed' , 'and' , 'pillows' , 'were' , 'very' , 'comfortable' , ' ']

↓

['bed' , 'and' , 'pillows' , 'were' , 'very' , 'comfortable' , ' ']

↓

['bed' , 'pillows' , 'comfortable' , ' ']

↓

['bed' , 'pillow' , 'comfortable']

4 속성별 문장 분류

: 기존의 호텔경영에서 사용하는 서비스 품질 속성 대신 소비자 관점에서의 호텔 서비스품질 속성 정의 필요

후보 속성 추출

리뷰에서 많이 언급할 수록 소비자가 관심 있는 속성
➔ 리뷰에서 자주 출현하는 명사 단어

숨겨진 속성 추가

비정형 자연어 데이터 특성상
같은 속성을 지칭하는 여러 단어들이 존재
➔ 동의어나 유의어에 대한 처리 필요

분석 대상이 되는 품사

리뷰 문장 = 속성 + 감성단어

속성: 명사(NN)
감성단어: 형용사(JJ), 부사(RB), 동사(VB)

ex) Room is clean and tidy, 5 mins walk to Hongdae main street.

단어	room	clean	tidy	min	walk	hongdae	main	street
tag	NN	JJ	NN	NN	VBP	NN	JJ	NN
품사	명사	형용사	명사	명사	동사	명사	형용사	명사

명사 중 빈도수로 후보속성 추출

: 토큰화된 단어 38749개 중 19888개의 명사, 2070개의 명사단어

〈출현 빈도가 높은 상위 토큰 30개 출력〉

```
In [64]: print(text.vocab().most_common(30))
[('hotel', 1818), ('room', 1227), ('location', 555), ('staff', 475),
('station', 467), ('airport', 310), ('night', 287), ('seoul', 283),
('time', 240), ('restaurant', 231), ('area', 225), ('service', 218),
('stay', 217), ('city', 207), ('subway', 202), ('line', 170),
('place', 170), ('breakfast', 161), ('floor', 159), ('convenient',
157), ('bus', 136), ('business', 128), ('train', 127), ('bed', 126),
('hongdae', 125), ('clean', 125), ('bathroom', 122), ('everything',
118), ('lot', 117), ('day', 117)]
```

이 중 호텔 서비스 속성과 관련 있다고 생각되는 단어에 표시

〈유사한 속성 구분〉

우선적으로 관련 논문을 참고하여 직관적으로 유사한 속성끼리 그룹

1 | room, bed, bathroom, clean, floor
2 | location, station, airport, restaurant, area, Hongdae, convenient
3 | subway, line, bus, train
4 | staff, service, breakfast

4 속성별 문장 분류

단어벡터 생성

리뷰에 사용된 단어들의 의미를 고려하여
밀집벡터로 변환하는 word embedding 기법 사용

단어 유사도 계산

두 단어벡터 간의 코사인 유사도

숨겨진 속성 찾기

1. 후보 속성 별로 가장 유사한 30개의 단어 중 구분

- 같은 속성그룹에 있는 단어
- 다른 속성그룹에 있는 단어
- 추가 속성 후보가 될 수 있는 단어

2. 유사도 상위 3개의 단어는 빨간색으로 표시

속성 추가

사전 분류된 속성그룹을 평가하고
그룹의 속성 중 50% 초과의 속성과 유사한 단어 추가

문장 분류

문장에 속성 키워드들 중 하나가 존재하면
해당 속성의 리뷰로 분류

[group1] 객실

대부분의 속성이 서로 유사함, 다른 그룹에 유사한 속성이 거의 없음 〉 good

➔ 추가 : 5개 속성 중 3개 이상의 속성과 유사한 size, shower, design

[group2] 근처 지역

airport : 유사한 상위 3개의 속성 모두 group3이므로 group3으로 분류

ㄴstation도 유사한 상위 2개의 속성이 group3이 되므로 group3으로 분류

➔ 추가 : 5개 속성 중 4개 이상과 유사한 shopping

restaurant과 매우 유사한 supermarket, store

*convenient(편의점)의 경우 유사한 속성이 모두 group3이지만
그 경우 형용사로 '편리한'이라는 뜻을 가지므로 해당되지 않음

[group3] 교통수단

airport, station을 추가한 결과 대부분의 속성이 서로 유사함 〉 good

➔ 추가 : 6개의 속성 전체와 유사한 exit, metro

[group4] 서비스

➔ 추가 : 3개 속성 중 2개 이상과 유사한 english, desk, pool, check

4 속성별 문장 분류

속성	같은 그룹 내 유사한 속성	다른 그룹 내 유사한 속성	새로 언급된 속성
room	bed / bathroom / clean	breakfast / staff	shower / pool / design / size
bed	room / clean / bathroom		shower / facility / size / view / design
bathroom	room / clean / bed / floor		size / shower / view / design
clean	bed / bathroom / room	staff	size
floor	bathroom / room		pool / size
location	station / hongdae	subway / line	exit / shopping / metro
restaurant	hongdae / convenient	subway	shopping / supermarket / store / located / shop
area	hongdae / restaurant / station	subway	shopping / exit / located
hongdae	area / restaurant / location / station	subway	shopping / exit
convenient	station / airport	line / subway / train	metro / shop / supermarket
station	airport / location / convenient	line / subway / bus / train	exit / metro / shopping / located
airport	convenient	line / bus / train / subway	access / metro / exit
subway	line / train / bus	airport / station	metro / exit / convenient / restaurant
line	train / subway / bus	station / airport	exit / metro / access
bus	train / line / subway	airport	exit / metro
train	line / bus / subway	station / airport	exit / metro / access
staff	service	clean	english / reception / desk / pool / check
service	breakfast	clean / location	english / desk / front
breakfast	service / staff	room / bed	pool / check

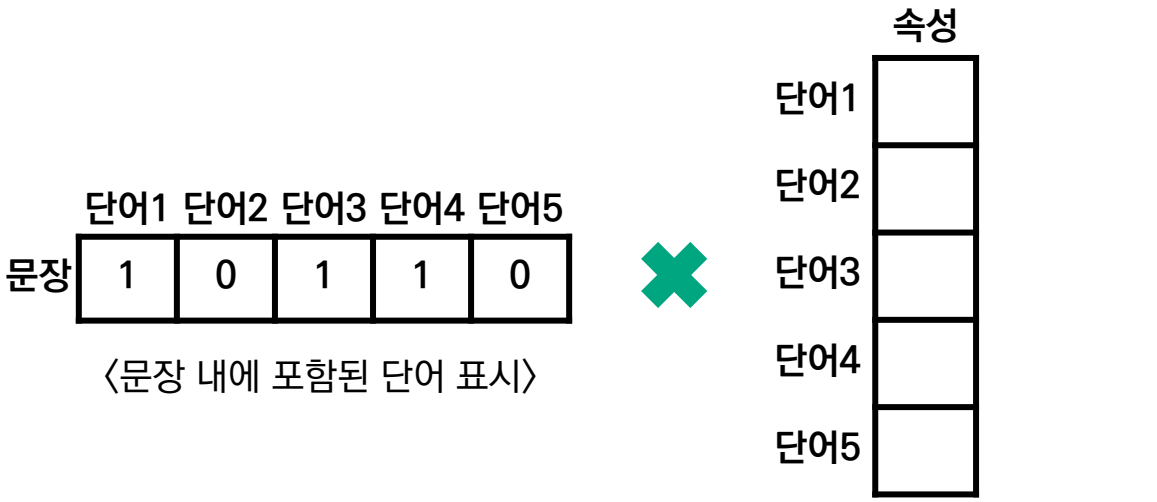
5 리뷰의 속성 연관정도에 따른 우선순위

: 해당 속성과 연관성이 높을 리뷰일수록 더 소비자가 원하는 정보이므로
소비자에게 유용한 리뷰

- 속성과 리뷰의 연관성이 높다 = 리뷰 내의 단어들과 속성 단어가 유사하다
- 단어 간의 유사도 = 단어 벡터 사이의 거리를 이용하여 계산
- 리뷰 문장의 연관 점수 = 문장 내 단어들의 연관 점수(유사도)의 합

➔ 각 문장별로 연관 점수를 구하여 연관성이 높은 순으로 정렬

〈속성과 문장의 연관 점수 구하는 법〉



〈문장 내에 포함된 단어 표시〉

〈속성과 단어의 유사도 가중치〉

DTM : Document-Term Matrix

: 다수의 문서에서 등장하는 각 단어들의 빈도를 행렬로 표현한 것

1. 해당 속성 리뷰에 포함되어 있는 단어 집합 만들기 (중복제거)

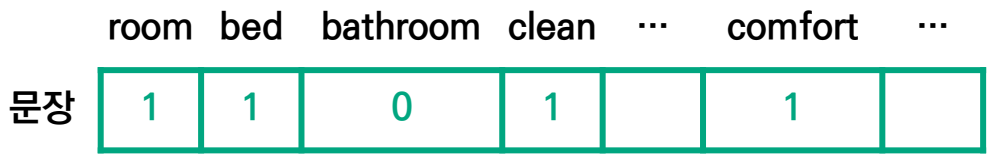
속성 그룹1의 경우 1549개의 리뷰 문장과 2008개의 단어

2. 단어에 0부터 고유한 정수 index 부여

room : 0 / bed : 1 / bathroom : 2 / clean : 3 / floor : 4 …

3. 리뷰 문장 내 단어의 등장 횟수를 index에 맞춰 기록한 벡터

ex) Room is clean and bed is comfort.



5 리뷰의 속성 연관정도에 따른 우선순위

4. 단어가 리뷰 문장에 등장한 빈도수를 나타내는 DTM

	단어 1	단어 2	단어 3	단어 4	단어 5	…	단어 2008
문장 1	1	0	1	1	0		0
…							
문장 1549	0	1	1	0	0		0

유사도 가중치 행렬

: 단어 간의 의미가 유사할수록 단어벡터 간의 거리가 가까우므로
속성과 거리가 가까울수록 큰 가중치를 갖도록 단어의 가중치 값을 계산

1. 단어 벡터 행렬

word2vec로 만든 단어벡터의 좌표 ➔ 행 : 단어 / 열 : 차원 수 (2008x100)

	1	2	3	4	5	…	100
단어 1	1	0	1	1	0		0
…							
단어 2008	0	1	1	0	0		0

2. 단어 간 거리 행렬

단어벡터 사이의 거리 ➔ 유클리디안 거리 활용

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + +(x_{i100} - x_{j100})^2}$$

	단어 1	단어 2	단어 3	단어 4	단어 5	…	단어 2008
단어 1	0	5.60	6.35	4.27	6.60		3.40
단어 2	5.60	0	7.35	6.00	7.87		5.44
…							
단어 2008	3.40	5.44	5.90	4.12	6.28		0

행과 열 모두 단어를 나타내는 정사각행렬
주대각선은 자신과의 거리이므로 0

5 리뷰의 속성 연관정도에 따른 우선순위

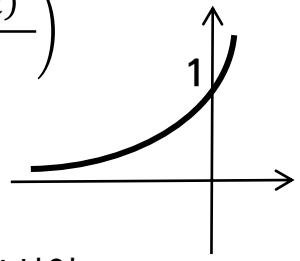
3. 유사도 가중치 행렬

속성 단어와 거리가 가까운(의미가 유사한) 단어는 높은 가중치를,
거리가 먼 단어는 낮은 가중치를 가지도록 단어 간의 거리를 이용하여 가중치 계산

	room	bed	bathroom	clean	floor	size	shower	design
단어 1	1							
단어 2		1						
...								
단어 2008								

word2vec 의 뉴럴 네트워크의 역전파 과정에서 학습된 가중치 공식 사용

$$W_{i,j} = \exp\left(-\frac{d(x_i, x_j)^2}{2\sigma^2}\right) \approx \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



- 정규분포 확률밀도식과 유사
- exp(x)에서 x가 항상 음의 값을 가지므로 가중치는 0과 1사이
- d(거리)가 커질수록 x의 절대값이 커지므로 가중치는 더 작아짐

리뷰 문장별 우선순위

DTM과 유사도 가중치 행렬을 내적하여 문장별로 속성 단어와의 유사도를
구한 뒤, 속성 단어들의 점수를 합하여 구한 속성과의 유사도에 따라 정렬

〈두 리뷰 문장의 점수 비교〉

Room is clean and bed is comfort.

The hotel is amazing, it's very clean and really well located in Hongdae.

	room	bed	bathroom	clean	floor	size	shower	design	속성
1	1.02	1.00	0.00	1.02	0.00	0.00	0.00	0.00	3.04
2	0.05	0.00	0.00	1.01	0.00	0.00	0.00	0.00	1.06

➔ 첫번째 리뷰 문장의 속성과의 유사도가 더 높음

➔ 소비자에게 두번째 리뷰보다 더 먼저 보일 수 있게 배치

6 속성별 감성점수 계산

단어 감성점수 → 문장 감성점수

Sentiwordnet을 활용하여 각 단어의 감성 점수 도출

단어 감성점수의 평균으로 문장 감성점수 계산

단어 별 긍정/부정 점수 출력

단어	품사	긍정점수	부정점수
clean	명사	0.0	0.0
clean	형용사	0.5	0.375
dark	명사	0.125	0.5
uncomfortable	형용사	0.125	1.25

-〉 synset 집단의 경우, 각 단어의 품사에 따라 점수의 차이 발생
앞서 진행한 품사가 부착된 단어들을 활용하여 각 단어 별로 같은 품사를 지닌 경우의 긍정 / 부정 점수를 추출하도록 구현

단어의 감성점수 계산

한 단어에 긍정 점수와 부정 점수 모두 존재 가능

➔ 감성 점수 = 긍정 점수 – 부정점수

단어	긍정 점수	부정 점수	감성 점수
clean (a)	0.5	0.375	0.125
dark	0.125	0.5	−0.375
Uncomfortable	0.125	1.25	−1.125

리뷰의 감성점수 계산

The room light was very very very dark, which made me very uncomfortable, not recommended at all

['room', 'light', 'dark', 'made', 'uncomfortable', 'recommended']
0 0 −0.375 −0.375 −1.125 0.525

➔ 각 단어 별 감성점수의 합은 −1.35, 이를 단어 수인 6으로 나누면 리뷰의 감성 점수는 −0.225 로 산출

6 속성별 감성점수 계산

속성별 리뷰의 극성비율 계산

[room 속성의 경우]

속성에 해당하는 리뷰	감성 점수
Room is clean and tidy, 5 mins walk to Hongdae main street.	0.071
Also thanks the housekeeping team to make the room tidy and clean.	0.078
The room light was very very very dark, which made me very uncomfortable, not recommended at all.	-0.225
I had a nice view from my room and it was located in a good area.	0.271
The staff was quick to get us into another room and compensated us with free breakfast.	0.054

감성점수의 계산 시 '긍정 점수 - 부정 점수'로 계산을 하였기에,
감성 점수가 〉 0 이면 긍정 리뷰, 감성 점수가 〈 0 이면 부정 리뷰로 판단

이 외에도 room 속성에 해당하는 총 **1487**개의 review가 존재
〉 전체 리뷰 1487 중 긍정 리뷰 수, 부정 리뷰 수의 비율을 구함

room 속성의 경우, 긍정 리뷰는 1053개, 부정 리뷰는 434개 존재
전체 1487개 중 **71%가 긍정, 39%가 부정**임을 확인 가능
이와 같은 방법으로, 각각의 속성에 대해 진행하여 비율 계산

7 결론

향후 연구방향

현재 구축한 서비스에서 고객에게 더 큰 유용성을 제공할 수 있도록
'정확도 향상'이 필요

1. 전용 감성사전의 구축

: 속성에 대한 감성을 나타내는 단어는 '어떠한 속성과 함께 쓰이냐'에 따라 다른 감성 값을 가질 수 있으므로 속성별로 감성사전을 구축한다면 리뷰의 극성 판단 시 더욱 정확한 결과를 얻을 수 있을 것이다.

2. 리뷰의 문맥을 고려한 감성 점수의 산출

: ' ~도 좋고 ~도 좋고 다 좋았지만, 전반적인 서비스가 별로 였다.'
같은 리뷰의 경우 앞 문장으로 인하여 긍정 점수가 더 높게 산출되어
긍정 리뷰로 분류된 가능성이 존재하지만 문맥상 부정 리뷰에 해당

➔ 이러한 문맥까지 고려한 리뷰의 극성 분석이 필요

또한, 'very', 'much'등의 정도 부사에 대한 가중치를 부여하여
감성 점수를 산출할 수 있다면, 더욱 정확하게 우선순위 선정이 가능할 것

참고문헌

강인수(2013),『영어 트위터 감성 분석을 위한 Sentiwordnet 활용 기법 비교, 한국 지능시스템학회 논문지, 23(4), 317-324

김유영, 송민(2016) ,『영화 리뷰 감성분석을 위한 텍스트 마이닝 기반 감성 분류기 구축』, 연세대학교, 1-4

딥 러닝을 이용한 자연어 처리 입문, 위키독스

문화체육관광부, 외래관광객조사, 2018

연종흠, 이동주, 심준호, 이상구(2011),『전자 상거래의 온라인 감성분석처리가 필요한 이유』, 한국전자거래학회 학술대회 발표집, 139-142

이현애, 정남호, 구철모(2017) ,『호텔 등급에 따른 온라인 리뷰 유형과 유용성의 관계 분석』, 경영학연구, 46(1), 137-156

조선배(1994),『호텔 서비스 평가 척도의 개발』, 호텔경영학연구, 2, 167-188

카토 코타,『파이썬을 이용한 머신러닝, 딥러닝 실전 개발 입문』, 위키북스, 2017

카토 코타,『파이썬을 이용한 웹 크롤링과 스크레이핑』, 위키북스, 2018