

『졸업작품 프로젝트』 과제제안서

프로젝트 내용	희망하는 연구개발의 제목	통합 리뷰를 기반으로 한 제품 안내 어플		
	지도교수	이기훈		
	팀원	2 명	컴퓨터공학과 2015722084 한승주 컴퓨터공학과 2015722083 김성종	
교육 신청관련 개인정보 수집, 활용 및 제공에 대한 동의				동의 여부
수집하는 개인정보 항목		• 소속, 성명, 학년, 학번, 연락처, E-mail		○
개인정보의 수집 및 이용목적		• “졸업작품 프로젝트” 안내를 위함.		○
개인정보의 보유 및 이용기간		- 수집한 고객의 개인정보를 수요조사 기간까지만 보유하며 수요조사 완료 후 관련법규에 의거하여 안전하게 파기합니다. (개인정보보호법 시행령 제 16 조) - 정보제공자가 개인정보 수집·이용에 대한 동의를 철회할 경우 수집한 개인정보를 즉시 파기합니다.		○
개인정보 제공 동의 거부 권리 및 동의 거부 따른 불이익 내용 또는 제한사항		귀하는 개인정보 제공 동의를 거부할 권리가 있으며, 동의 거부에 따른 불이익은 없음. 다만, 추가적인 교육 서비스를 받을 수 없음.		○
<p>※ 개인정보 제공자가 동의한 내용외의 다른 목적으로 활용하지 않으며, 제공된 개인정보의 이용을 거부하고자 할 때에는 개인정보 관리책임자를 통해 삭제를 요청 할 수 있음.</p> <p>「개인정보보호법」 등 관련 법규에 의거하여 상기 본인은 위와 같이 개인정보 수집 및 활용에 동의함.</p> <p style="text-align: right;">2020 년 01 월 10 일</p> <p style="text-align: right;">성명 : 한 승 주 , 김 성 종 (한승주, 김성종)</p>				

『졸업작품 프로젝트』 과제제안서

제작동기

1. 기존의 리뷰방식의 문제점

- 각기 다른 플랫폼이나 웹사이트별로 사용하는 연령대나, 나이 대 등이 달라 제품의 정보를 단 하나의 사이트만 보고는 신뢰성 있는 정보를 통한 구매가 불가능하다.
- 실제 국내 페이스북 유저 비율은 남성이 여성 대비 14% 많고, 인스타그램은 여성이 남성 대비 4% 많은 비율을 가지고 있다. 또한 페이스북은 연령대가 고른 반면, 인스타그램은 20~30 대 비율이 상대적으로 높다.

2. 통합 리뷰방식의 장점

- 정보의 신뢰성
여러 플랫폼에 퍼져 있는 제품의 리뷰를 중요 키워드를 중심으로 데이터베이스를 구축하면 이렇게 만들어진 데이터베이스를 통해 키워드만으로 사용자가 원하는 조건의 제품을 소개하고 찾아볼 수 있다. 긍정적 리뷰와 부정적 리뷰를 다 같이 한번에 보여줌으로써 제품의 장점뿐만 아니라 단점을 파악하고 같은 제품군을 비교 구매하고 싶어 하는 사용자에게 편의성을 제공할 수 있다.
- 접근의 용이성
이런 결과를 어플리케이션(혹은 웹사이트)으로 추가적인 구현을 하면 사용자가 언제 어디서든 쉽게 접근하고 사용할 수 있다.

3. 국내 현황

- 페이스북, 인스타그램, 트위터와 같은 소셜미디어에 자주 노출되는 광고와 함께 실제 사용자가 올리는 리뷰, 그리고 제품 구매처인 쇼핑몰(G마켓, 옥션 등), 제품 가격 비교 사이트(다나와, 네이버 비교 쇼핑 등)은 여러 존재하지만, 이렇게 많은 웹사이트와 플랫폼에 퍼져 있는 사용자의 구매 후기를 통합적으로 파악할 수 있는 곳은 없다.

개념설계

1. 플랫폼의 웹사이트 코드 분석 및 크롤링(Beautiful soup)

- DB 를 구축하기 위해 해당 웹사이트 접속
- 제품 정보와 사용자 및 구매자의 리뷰가 담긴 웹사이트 코드 분석
- BS4 를 이용해 페이지 데이터 호출
- 제품의 기본 정보(이름, 가격, 리뷰 등) tag 를 찾아 추출
- 각 사이트와 페이지별로 링크를 재귀적으로 검색하여 데이터를 추출

Ex)

... 생략 ...

```
def get_product_info(box):
```

```
    ...
```

```
def get_page_products(url):
```

```
    ...
```

```
urls = [
```

```
    "http://joise.com/category/tonermist/43/?page=1",
```

```
    "http://joise.com/category/tonermist/43/?page=2"
```

```
    "http://joise.com/category/tonermist/43/?page=3"
```

```
    "http://joise.com/category/tonermist/43/?page=4"
```

```
    "http://joise.com/category/tonermist/43/?page=5"
```

```
]
```

```
for page_number in range(0,5):
```

```
    page_products= get_page_products(urls[page_number])
```

```
    print(page_products)
```

결과

```
[{'name': "It's skin Cera Routine Moisturizer 150ml", pri ... 생략 ...
```

```
[{'name': 'may COOP Raw Sauce 150ml', 'price': 'USD 37.65', ... 생략 ...
```

```
[{'name': 'The FACE SHOP Dr.Belmeur Daily Repair Toner 200ml', ... 생략 ...
```

```
...
```

2. 크롤링된 정보를 이용한 DB 구축(MySql)

- MySQL 서버에 접속하여 데이터베이스 생성
- Cursor 를 추출하여 execute 메서드로 SQL 을 실행, 테이블 생성
- Execute 메서드에 데이터를 계속 확장

Ex)

```
import MySQLdb
```

```
#데이터베이스 연결하기
```

```
connection = MySQLdb.connect(
```

```
    user="scrapingman",
```

```
    passwd="myPassword-1",
```

```
    host="localhost",
```

```
    db="scrapingdata",
```

```
    charset="utf8")
```

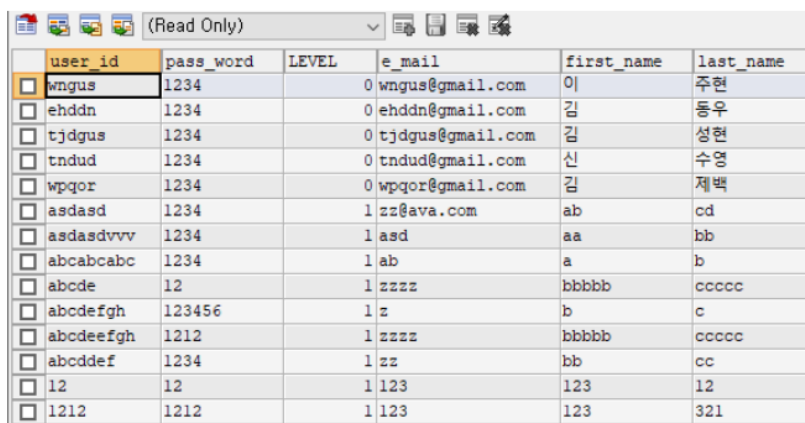
```
# 커서 생성 및 실행할 때마다 다른결과가 나오지 않게 테이블 제거
```

```
cursor=connection.cursor()
```

```
cursor.execute("DROP TABLE IF EXISTS user_id")
```

```
... 생략 ...
```

Ex)



user_id	pass_word	LEVEL	e_mail	first_name	last_name
wngus	1234	0	wngus@gmail.com	이	주현
ehddn	1234	0	ehddn@gmail.com	김	동우
tjdgus	1234	0	tjdgus@gmail.com	김	성현
tndud	1234	0	tndud@gmail.com	신	수영
wpqor	1234	0	wpqor@gmail.com	김	제백
asdasd	1234	1	zz@ava.com	ab	cd
asdasdvvv	1234	1	asd	aa	bb
abcabcabc	1234	1	ab	a	b
abcde	12	1	zzzz	bbbbbb	ccccc
abcdefgh	123456	1	z	b	c
abcdeefgh	1212	1	zzzz	bbbbbb	ccccc
abcddef	1234	1	zz	bb	cc
12	12	1	123	123	12
1212	1212	1	123	123	321

3. Kkma, Hannanum 을 이용한 KoNLP(키워드 생성)

- Kkma 나 Hannanum 모듈을 이용하여, 해당 모듈에 맞추어 입력된 문자열에서 키워드로 표현할 품사 추출
- 가장 빈도수가 높은 단어(키워드로 설정할 단어)를 DB 에 저장

- 문자열의 형태소 분류(Hannanum)

N : 체언 (고유명사 대명사 의존명사 수사 등이 포함)

J : 관계언 (격조사가 포함)

X : 접사 (접두사 접미사 등이 포함)

E : 어미 (연결어미 종결어미 등이 포함)

P : 용언 (동사 형용사 등이 포함)

S : 기호

- Pos 를 이용한 문장 분할 및 추출 및 def 를 이용한 단어 카운트

Ex)

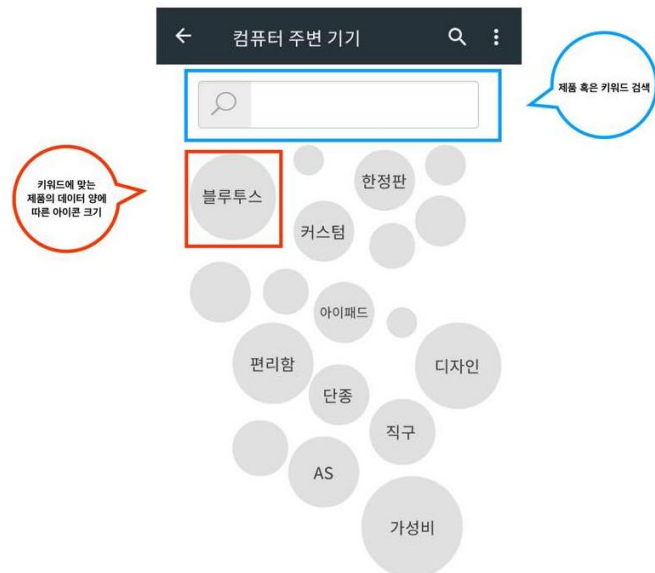
0 대한민국헌법 유구한 역사와
9 대한민국은 3·1 운동으로 건립된 대한민국임시정부의 법통과 불의에
98 총강 제 1 조 ① 대한민국은 민주공화국이다. ②대한민국의
100 ① 대한민국은 민주공화국이다. ②대한민국의 주권은 국민에게
110 나온다. 제 2 조 ① 대한민국의 국민이 되는
126 의무를 진다. 제 3 조 대한민국의 영토는 한반도와
133 부속도서로 한다. 제 4 조 대한민국은 통일을 지향하며,
147 추진한다. 제 5 조 ① 대한민국은 국제평화의 유지에
787 군무원이 아닌 국민은 대한민국의 영역안에서는 중대한
1836 파견 또는 외국군대의 대한민국 영역안에서의 주류에
3620 경제 제 119 조 ① 대한민국의 경제질서는 개인과
[0, 9, 98, 100, 110, 126, 133, 147, 787, 1836, 3620]

가장 빈도수 높은 단어 10개 가져오기 (키워드 생성)

[('저', 155),
('법률', 121),
('수', 88),
('대통령', 84),
('국가', 73),
('국민', 69),
('헌법', 69),
('조', 58),
('국회', 55),
('때', 55)]

4. Android UI 제작

- 자신이 구매하고자 하는 제품의 리뷰를 보기 위한 제품의 검색창과 위에서 나타낸 키워드 중 전체 제품에서 가장 많은 비중을 차지하는 몇 개의 키워드를 다음과 같이 제품 검색창 아래에 사용자가 보기 편하도록 UI로 구현
- 검색 시, 제품 검색 및 키워드(조건)를 검색할 수 있게 구현



- 제품 검색 시, 제품의 가격 정보와 리뷰의 중요 키워드를 노출
- 실제 리뷰를 사이트 별로 노출
- 어플 자체 리뷰를 남겨 사용자가 어플을 더 다양하게 사용하도록 유도



기대효과	<p>[기업적 측면]</p> <ul style="list-style-type: none"> - 제품에 대한 긍정적인 리뷰를 많이 보이는 제품의 경우, 구매자로 하여금 다양한 웹사이트의 공통적인 리뷰를 확인가능함으로써 리뷰의 신뢰성이 높아져 구매력이 증가할 것이다. - 해당 제품의 실제 이용고객이 제품을 사용하고 실제로 좋은 제품이라면 회사의 신뢰성이 증가하여 차기 출시될 제품의 신규 구매력도 증가할 것이다. <p>[사용자 측면]</p> <ul style="list-style-type: none"> - 제품에 대한 리뷰를 알아보기 위해 번거롭게 여러 플랫폼을 일일이 찾아 검색하지 않아도 한번에 파악할 수 있다. - 리뷰에서 자주 언급된 단어를 통해 중요 키워드를 산출해내기 때문에 신뢰가 가는 제품인지 쉽게 파악할 수 있다. - 비슷한 조건의 다른 좋은 선택지가 없는지 한 눈에 알아볼 수 있다.
참고문헌	<p>https://kiwiprnad.blog.me/221458922512 [페이스북과 인스타그램 유저의 인구통계학적 특성]</p> <p>https://blog.naver.com/zzang9ha/221692276477 [Python] MySQL DB 연동(기본 crud 테스트)</p> <p>https://blog.naver.com/meibee55/221703098261 [Python] KoNLP 자연어 처리 (Kkma, Hannanum)</p> <p>파이썬을 활용한 클로러 개발과 스크레이핑 입문, 2019, 카토 카츠야, 요코야마 유우키, 위키북스</p> <p>파이썬 데이터 수집 자동화 한방에 끝내기 한입에 웹크롤링, 2018, 김경록, 서영덕, 비제이퍼블릭</p>

본인은 상기의 내용이 모두 사실임을 확인하며 『졸업작품 프로젝트』를 제안합니다.

2020 년 01 월 10 일

신청인 성명 한승주, 김성종 (한승주, 김성종)