REPORT

Social Media 데이터 분석

스타벅스와 이디야 신제품 반응 예측 모델 구축

유니윌 김선경 진주리

REPORT

SNS
- Data Analysis

Chameleon


EDIYA
COFFEE

REPORT

SNS

- Data Analysis

INTRODUCTION

프렌차이즈순위에서도 커피창업은 줄곧 상위를 지키고 있습니다. 프렌차이즈 창업의 커피(디저트)의 창업율은 40%에 달한다고 합니다. 통계청에 의하면 우리나라 사람 1인당 연간 커피 소비량은 484잔이라고 합니다. 그럼 매일 1인당 한 잔 이상씩 마시고 있는 것과 같습니다.

이처럼 빠른 성장세를 보이고 있는 커피시장에서 고객을 사로잡을 수 있는 확률이 높은 신제품을 제안하기 위해 이 프로젝트를 시작하게 되었습니다.

TABLE OF CONTENTS

Ⅰ. WHY NAVER & INSTAGRAM?

Ⅱ. WHY EDIYA & STARBUCKS?

REPORT

SNS

- Data Analysis

PART 1. 네이버블로그

Ⅰ.데이터 수집

Ⅱ.텍스트 정제

1. word2vec 알고리즘
2. 단어 통일
3. 게시글 정제
4. 기타 정제

Ⅲ.텍스트 정보화

1. 게시글 점수화
2. 점수 테이블 만들기
3. 점수 부여

Ⅳ. 텍스트 학습

1. 데이터셋 만들기
2. 데이터 병합
3. 점수 라벨링
4. 머신러닝

Ⅳ. 결과

네이버 블로그 크롤링이 안된다...?

Ⅰ. 데이터 수집

"selenium"을 통한 게시글 크롤링

— 웹 브라우저를 컨트롤하여 UI(User Interface)를 지정하는 도구

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
driver = webdriver.Chrome("c:/Windows/chromedriver")
for q in starbucks['click']:
    except_starbucks_naver = []
    driver.get("http://m.naver.com")
    driver.find_element_by_css_selector('#query').send_keys(q)
    driver.find_element_by_css_selector('#query').send_keys('\n')
……
    for z in range(1,11):
        while i < 16:
                num = (page-1)*15
                try:
                time.sleep(random.randrange(1,2))
                y = driver.find_element_by_css_selector( '#blog_' +str(num+i)+' > a > div.total.....
```

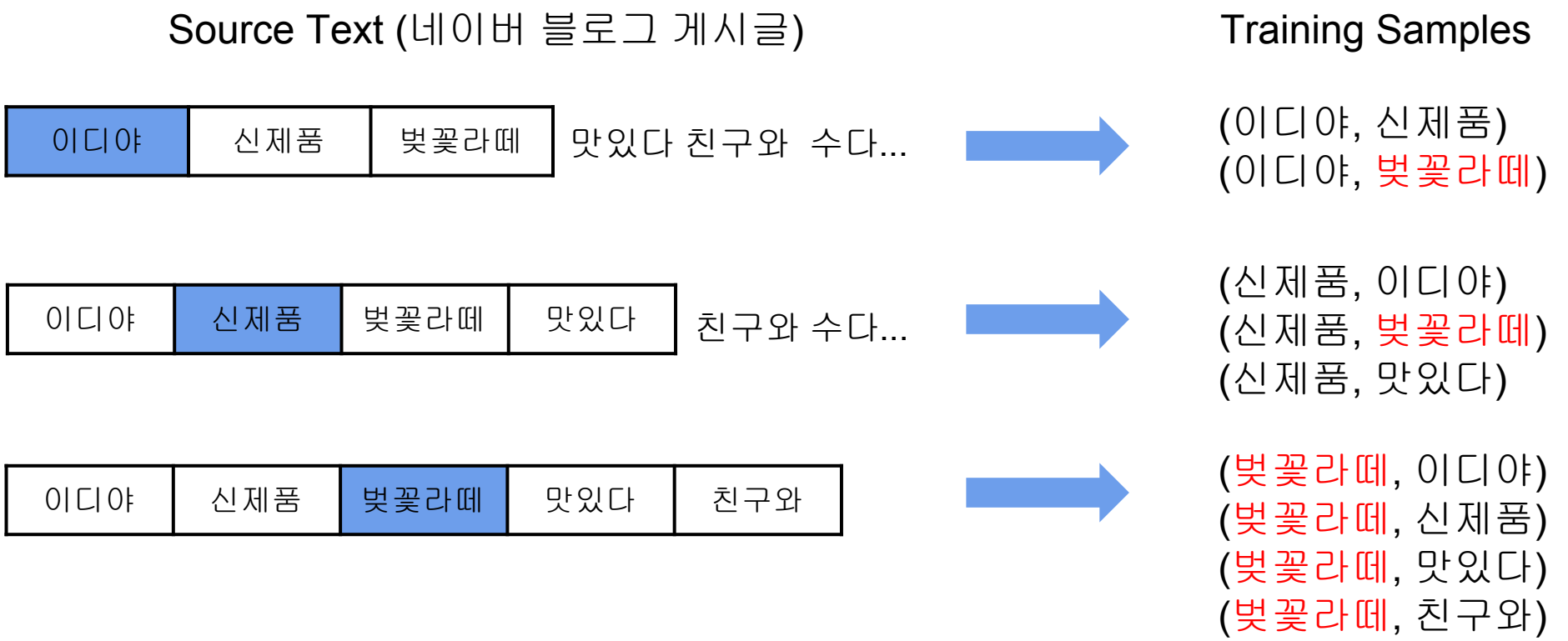
	스타벅스 블로그	:	12,424개
+	이디야 블로그	:	5,607개
	합계	:	18,031개

18,031개.. 블로그 어떻게 정제작업을 할까...?

Ⅱ - 1. Word2Vec 알고리즘

"word2vec"을 활용한 단어간 연관성분석

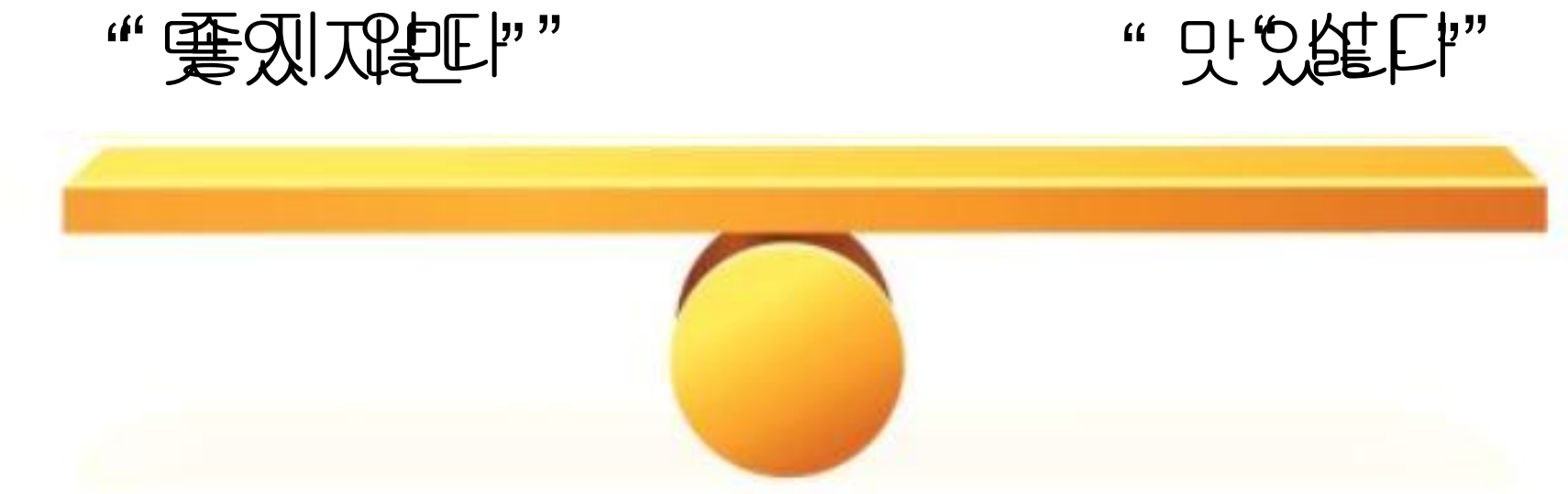
— 단어들을 벡터화시켜 단어간의 거리와 말뭉치의 빈도수를 계산해주는 알고리즘



한번에 학습할 수 있는 단어의 개수를 2개로 설정

“말이 아 다르고 어 다르다”

고객의 미세한 감정을 잘 활용해야 한다.



Ⅲ - 2. 단어 통일

선행부정어	→	**안**
후행부정어	→	않, **없**, 아니, 모르
반대 내용을 이끄는 접속부사	→	하나, **지만**, 한데, 하지만, 그러나
기타 처리	→	입맛에 딱, ~와 똑같다

Ⅲ - 2. 단어 통일

안맛있다

맛있지않다

맛있지만

맛있다

“ 사전등록 ”

(벚꽃라떼,안맛있다)
(벚꽃라떼, 맛있지않다)
(벚꽃라떼, 맛있지만)
(벚꽃라떼, 맛있다)

Ⅲ - 3. ckonlpy

Hyunjoong Kim
lovit

Ph.D student / Data science / Nat
Language Processing / Machine
Learning / soy.lovit@gmail.com

Block or report user

https://lovit.github.io

Overview | Repositories 43 | Stars 15 | Followers 100 | Following 4

Pinned repositories

soynlp

한국어 자연어처리를 위한 파이썬 라이브러리입니다. 단어 추출/ 토크나이저 / 품사판별/ 전처리의 기능을 제공합니다.

customized_konlpy

zed KoNLPy - Korean Natural Language Processing
t KoNLPy wrapping code

9

입니다. CRF 와 같은 머신러닝
법으로 띄어쓰기를 교정합니

textmining-tutorial

clustering4docs

customized_konlpy

Customized KoNLPy - Korean Natural Language Processing Toolkit KoNLPy wrapping code

Jupyter Notebook ★ 26 9

https://github.com/lovit

Ⅲ - 3. 게시글 정제 작업

"..... 모카프라푸치노에서 좀더 **진한 맛이긴한데**, 또 가 ....."

나는 이 음료에 시시한 추억을 가지
2009년 겨울, 내 생에 첫 텀블러 쿠폰으로 주문한 음료가 ㅂ ㅡr 프라푸치노인 것이다.
그 당시 텀블러 쿠폰은 그야말로 무ㅎ, ㅡ 쿠폰이었다.
음료 종류와 사이즈 불문, 엑스트라 횟수 제한도 없어 뭐든지 가능했다.
물론 스타벅스의 엑스트라 서비스를 잘 아는 사람도 없을 때였다.
어쨌든 몹시 다정한 파트너가 조심스레 포장한 프라푸치노를 들고 총총걸음으로 기숙사에 돌아가던 길은 설레고 즐거웠다.
춥고 매서운 바람이 불어도 마냥 신났다.

Ⅲ - 3. 게시글 정제 작업

```
re_adjective('좋','좋다','좋지않다','좋지만',docs)
re_adjective('색다','색다르다','색다르지않다','색다르지만',docs)
re_adjective('다르','다르다','다르지않다','다르지만',docs)
re_adjective('똑같','똑같다','똑같지않다','똑같지만',docs)
re_adjective('귀여','귀엽다','귀엽지않다','귀엽지만',docs)
re_adjective('아까','아깝다','아깝지않다','아깝지만',docs)
re_adjective('괜찮','괜찮다','괜찮지않다','괜찮지만',docs)
re_adjective('따뜻','따뜻하다','따뜻하지않다','따뜻하지만',docs)
re_adjective('나쁘','나쁘다','나쁘지않다','나쁘지만',docs)
re_adjective('아쉬','아쉽다','아쉽지않다','아쉽지만',docs)
re_adjective('아쉽','아쉽다','아쉽지않다','아쉽지만',docs)
re_noun('밍밍','밍밍하다','밍밍하지않다','밍밍하지만',docs)
re_noun('달달','달달하다','달달하지않다','달달하지만',docs)
re_noun('달','달달하다','달달하지않다','달달하지만',docs)
re_noun('뿜뿜','뿜뿜하다','뿜뿜하지않다','뿜뿜하지만',docs)
re_noun('실패','실패하다','실패하지않다','실패하지만',docs)
re_noun('추천','추천하다','추천하지않다','추천하지만',docs)
re_noun('취저','취저하다','취저하지않다','취저하지만',docs)
re_noun('심쿵','심쿵하다','심쿵하지않다','심쿵하지만',docs)
re_noun('합격','합격하다','합격하지않다','합격하지만',docs)
re_noun('풍미','풍미롭다','풍미롭지않다','풍미롭지만',docs)
re_noun('고급','고급지다','고급지않다','고급지지만',docs)
```

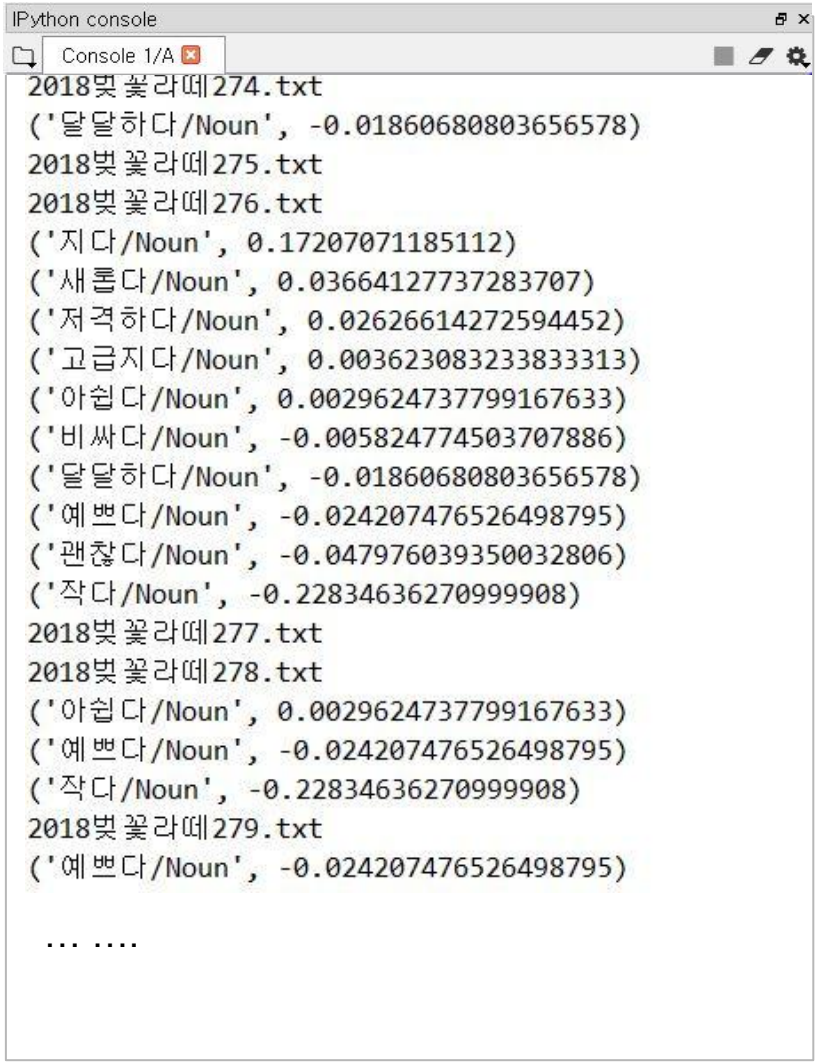
Ⅲ - 4. 기타 처리

'입맛에 딱이였음!!!!!! 아',
'입맛은 아니라며 복숭아',
'입맛 무엇..?..) 카페에서 쇼케이스를',
'입맛인지. 그래도 이제',
'입맛에는 아니예요. ㅋ',
'입맛돋고 좋드라구요^^ 쇠고기',
'입맛에 너무 달다ㅠ',
'입맛인 나에게 정말',
'입맛에는 괜찮았어여!! 그런데
'입맛에딱!!맛는ㅎㅎ 살짝'

'그 헤이즐럿 아메리카노와 똑같',
'엑스트라로 카라멜드리즐을 추가했다. 똑같',
'에이 이거 딸기우유랑 똑같',
'항아리모양 딸기우유랑 맛 똑같',
'오 블로그후기를에서 봤던거랑 똑같']

Ⅳ - 1. 게시글 점수화

```
from ckonlpy.tag import Twitter
t = Twitter()
final_beverage = []
postprocessor = Postprocessor(t, replace = dic_replace)
file_name_kobill = kobill.fileids()#kobill안에 있는 파일이름
docs_ko = [kobill.open(i).read() for i in kobill.fileids()]
from gensim.models import word2vec
for i in range(len(docs_ko)):
    print(file_name_kobill[i])
    if len(re.findall('\w{0,}',docs_ko[i])) != len(docs_ko[i])+1:
        docs_ko[i] = docs_ko[i]*20
        pos = lambda d: ['/'.join(p) for p in postprocessor.tag(d)]
        texts_ko = pos(docs_ko[i])
        for Z in range(len(texts_ko)):
            texts_ko[Z] = [texts_ko[Z]]
        wv_model_ko = word2vec.Word2Vec(texts_ko)
        wv_model_ko.init_sims(replace=True)
        wv_model_ko.save('ko_word2vec_e.model')
        try:
            file_name = file_name_kobill[i]
            a = score_beverage()
            final_beverage.append(a)
        except:
            continue
final = []
final.append(final_beverage)
```



Ⅳ - 2. 점수 테이블 생성

NEGATIVE ← → POSITIVE

-20점	-15점	-10점	-5점	+5점	+10점	+15점	+20점
좋지않다	싫지만	편하지않다	불편하지만	편하지만	편하다	좋지만	좋다
싫다	아깝다	불편하다	예쁘지만	괜찮지만	불편하지않다	싫지않다	굉장하다
맛없다	아쉽다	예쁘지않다	가깝지만	따뜻하지만	예쁘다	아깝지않다	추천하다
굉장하지않다	멀다	아깝지만	아름답지만	새롭지만	괜찮다	아쉽지않다	애정하다
나쁘다	맛있지않다	아쉽지만	뜨겁지만	적당하지만	따뜻하다	멀지않다	뿜뿜하다
추천하지않다	맛없지만	괜찮지않다	똑같지만	즐겁지만	새롭다	맛있다	저격하다
애정하지않다	부드럽지않다	따뜻하지않다		멋지지만	가깝다	부드럽다	심쿵하다
욕하다	시원하지않다	새롭지않다		많지만	맛있지만	시원하다	취저하다
뿜뿜하지않다	깔끔하지않다	가깝지않다		뜨겁지않다	맛없지않다	깔끔하다	추천하다
저격하지않다	고소하지않다	멀지만		귀엽지만	부드럽지만	고소하다	
심쿵하지않다	달콤하지않다	아름답지않다		재밌지만	아름답다	달콤하다	
취저하지않다	진하지않다	싱겁지만		설레지만	시원하지만	진하다	
추천하지않다	싱겁다	적당하지않다		만끽했지만	깔끔하지만	싱겁지않다	
	저렴하지않다	저렴하지만		반했지만	고소하지만	굉장하지만	
	비싸다	비싸지만			달콤하지만	저렴하다	
	비슷하다	비슷하지만			진하지만	비싸지않다	

Ⅳ - 3. 점수 부여

def score_beverage():

```
for i in wv_model_ko.most_similar(pos(j),topn=2000):
    if (i[0][i[0].find('/')+1:] == 'Noun') and (i[0][:i[0].find('/')][-1:] == '다'):
        print(i)
        lst.append(i)
for i1 in wv_model_ko.most_similar(pos(j),topn=2000):
    if (i1[0][i1[0].find('/')+1:] == 'Noun') and (i1[0][:i1[0].find('/')][-1:] == '지만'):
        print(i1)
        lst.append(i1)
for i2 in wv_model_ko.most_similar(pos(j),topn=2000):
    if (i2[0][i2[0].find('/')+1:] == 'Noun') and (i2[0][:i2[0].find('/')][-1:] in noun):
        print(i2)
        lst.append(i)
for z in range(len(lst)):
    if lst[z][0][:lst[z][0].find('/')] in positive_5:
        score = lst[z][1]
        score = score * 5
        score_final += score

… ….

    if lst[z][0][:lst[z][0].find('/')] in negative_20:
        score = lst[z][1]
        score = score * (-20)
        score_final += score
score_dic[j1] = score_final
return score_dic
```

print score_dic():

```
{'카라멜리스트레토비안코71.txt': 0.1971839740872383},
{'카라멜리스트레토비안코74.txt': -1.0110549442470074},
{'카라멜리스트레토비안코77.txt': -0.16882652416825294},
{'카라멜리스트레토비안코79.txt': -0.25920677930116653},
{'카라멜리스트레토비안코8.txt': -0.4964792914688587},
{'카라멜리스트레토비안코82.txt': 1.9731426797807217},
{'카라멜리스트레토비안코83.txt': 0.7857213355600834},
{'카라멜리스트레토비안코84.txt': 0.06514834240078926},
{'카라멜리스트레토비안코85.txt': 2.914459779858589},
{'카라멜리스트레토비안코87.txt': 0.9093738161027431},
{'카라멜리스트레토비안코90.txt': 1.02241612970829},
{'카라멜리스트레토비안코91.txt': 6.798028033226728},
{'카라멜리스트레토비안코92.txt': -0.5029981397092342},
{'카라멜리스트레토비안코94.txt': 1.2897718138992786},
{'카라멜리스트레토비안코95.txt': -1.5630182810127735},
{'카라멜리스트레토비안코96.txt': -0.4891696572303772},
{'카라멜리스트레토비안코97.txt': 1.6434356570243835},
{'벚꽃라떼1.txt': -2.7710622549057007},
{'벚꽃라떼10.txt': -0.19313422963023186},
```

Ⅳ. RECIPE 데이터셋 만들기

카라멜 소스

카라멜 드리즐

카라멜 시럽

카라멜

에스프레소 샷

비니스트마일드

비니스트오리지널

과테말라 에스프레소샷

샷

Ⅳ - 1. RECIPE 데이터셋 만들기

menu	샷	우유	...	카라멜시럽	...	카라멜드리즐	...	카라멜소스	...	벚꽃파우더	...
카라멜리스트레토비안코	1	0	...	0	...	1	...	0	...	0	...
...	...	...	...	...	...	...	...	...	...	...	...
벚꽃라떼	0	1	...	0	...	0	...	0	...	1	...
...	...	...	...	...	...	...	...	...	...	...	...

약 300개에 가까운 column

menu	샷	우유	...	카라멜	...	블라썸	...
카라멜리스트레토비안코	1	0	...	1	...	0	...
...	...	...	...	...	...	...	...
벚꽃라떼	0	1	...	0	...	1	...
...	...	...	...	...	...	...	...

119개의 column

Ⅳ - 2. MERGE

RECIPE TABLE

menu	샷	우유	...	점수
카라멜리스트레토비안코	1	0	…	
…	…	…	…	
벚꽃라떼	0	1	…	
…	…	…	…	

SCORE_DIC()

{‘카라멜리스트레토비안코1.txt’: -0.5459930375218391},

…

{‘카라멜리스트레토비안코11.txt’: 2.0612977258861065},

…

{‘벚꽃라떼114.txt’: -2.256549783051014}

…

{‘벚꽃라떼282.txt’: 1.6226249374449253}

Ⅳ - 2. MERGE

RESULT_TABLE

index	menu	샷	우유	...	점수
0	공주보늬밤라떼	**1**	**1**	**...**	
....	**...**	**...**	**...**	**...**	**...**
1700	벚꽃라떼	0	1	...	0
1701	벚꽃라떼	0	1	...	-2.5356519..
....	**...**	...	...	...	**....**
13121	카라멜리스트레토비안코	1	1	...	5.6823108..
13122	카라멜리스트레토비안코	1	1	...	-1.2097426..
...	**...**	...	...	...	...
18030	화이트코코리스트레토	1	1	...	

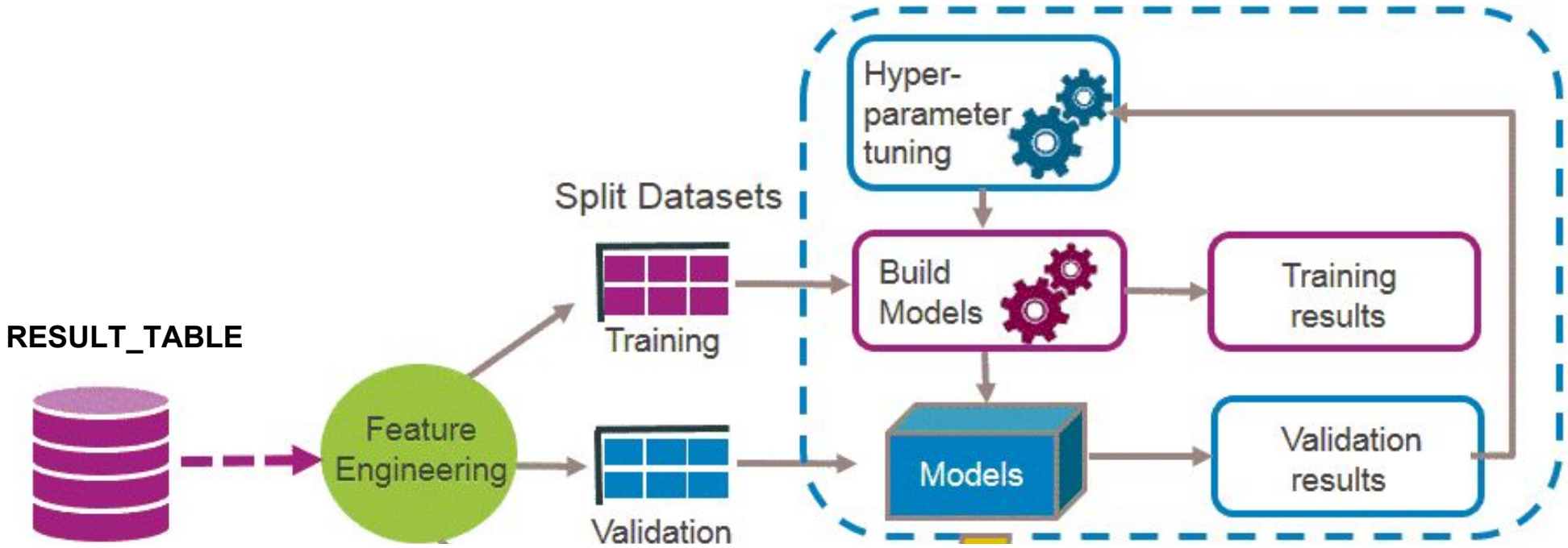
Ⅳ - 3. 점수 라벨링

RESULT_TABLE

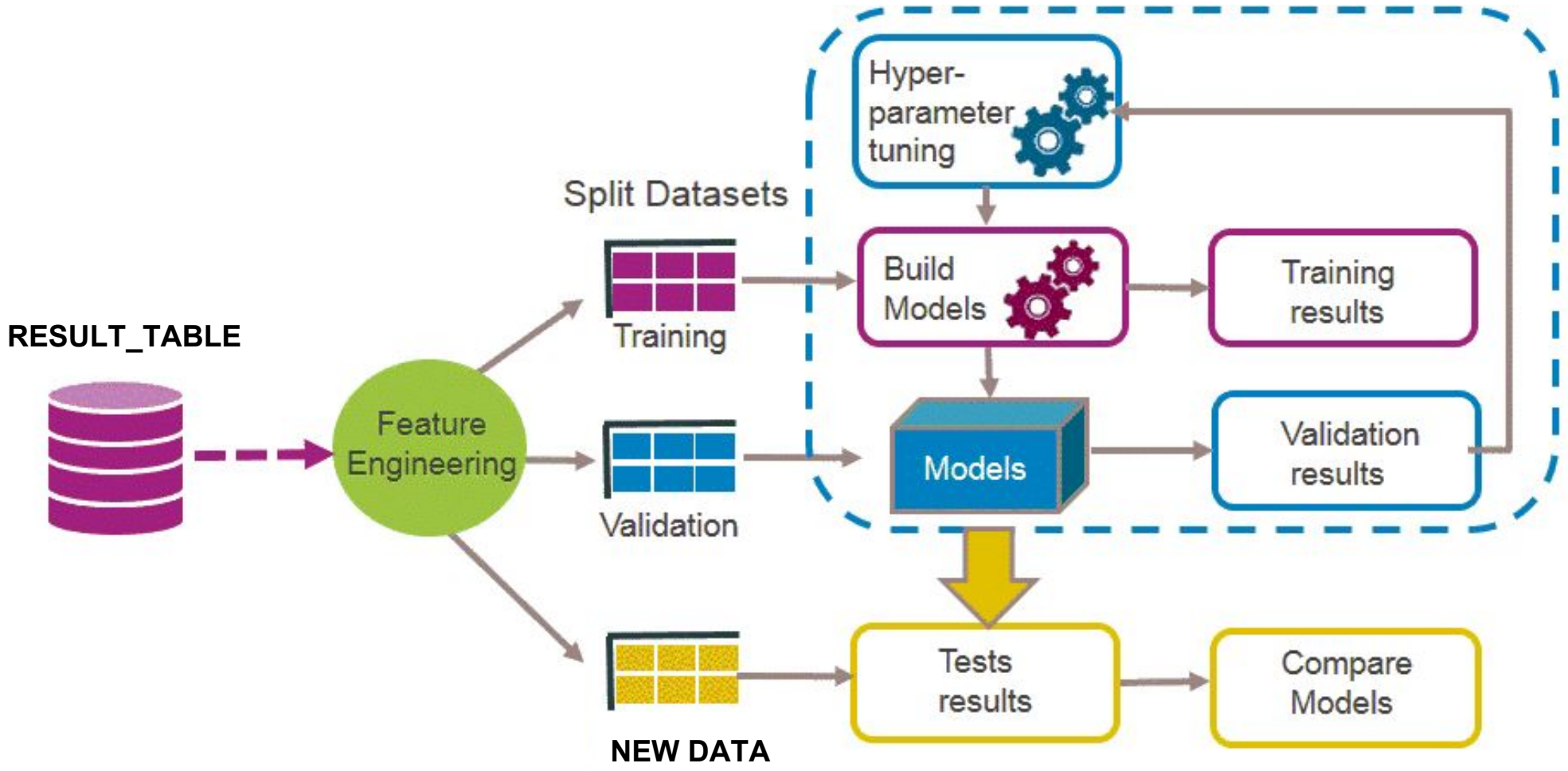
index	...	LABEL
0	...	
....	...	...
1700	...	2
1701	...	1
....	...	....
13121	...	4
13122	...	1
...	...	...

기준	LABEL
-10 이상 ~ - 5 미만	0
-5 이상 ~ 0 미만	1
0점	2
0점 초과 ~ 5점 이하	3
5점 초과 ~ 10점 이하	4

Ⅳ - 4. Machine Learning



Ⅴ. FINAL RESULT



IF, 딸기와 돌체시럽을 넣은 신제품이 나오면 어떤 점수가 나올까?

REPORT
SNS
- Data Analysis

PART 2. 인스타그램

Ⅰ. 데이터 수집

1. 문제점
2. 데이터 수집
3. 데이터 수집 코드
4. 데이터

Ⅱ. 데이터 정제

1. 인스타그램 약어 정제
2. 불필요한 단어 정제
3. ~그램 정제
4. 한글화
5. 기타 정제

Ⅲ. WORD CLOUD

1. 전체 데이터
2. menu 제외
3. 결과

Ⅳ. 이벤트 제안

Ⅰ. 데이터 수집


Instagram
검색
jinaest • 팔로우
연세세브란스빌딩
jinaest 역시 시즌음료는 별쿠폰으로✨
.
.
#스타벅스 #커피 #모닝커피 #starbucks #연세세브란스빌딩 #시즌음료 #뉴그린티프라푸치노 #먹지마세요 #맛이없음
gu_seul_ah, gfbjcsp0rs1, ukisgod, hee.jjung2, ggovely_2016, ilikai.official, hogyster, by_velyeun님이 좋아합니다
6분 전
댓글 달기...

I - 1. 문제점

Instagram

죄송합니다. 페이지를 사용할 수 없습니다.

클릭하신 링크가 잘못되었거나 페이지가 삭제되었습니다. Instagram으로 돌아가기.

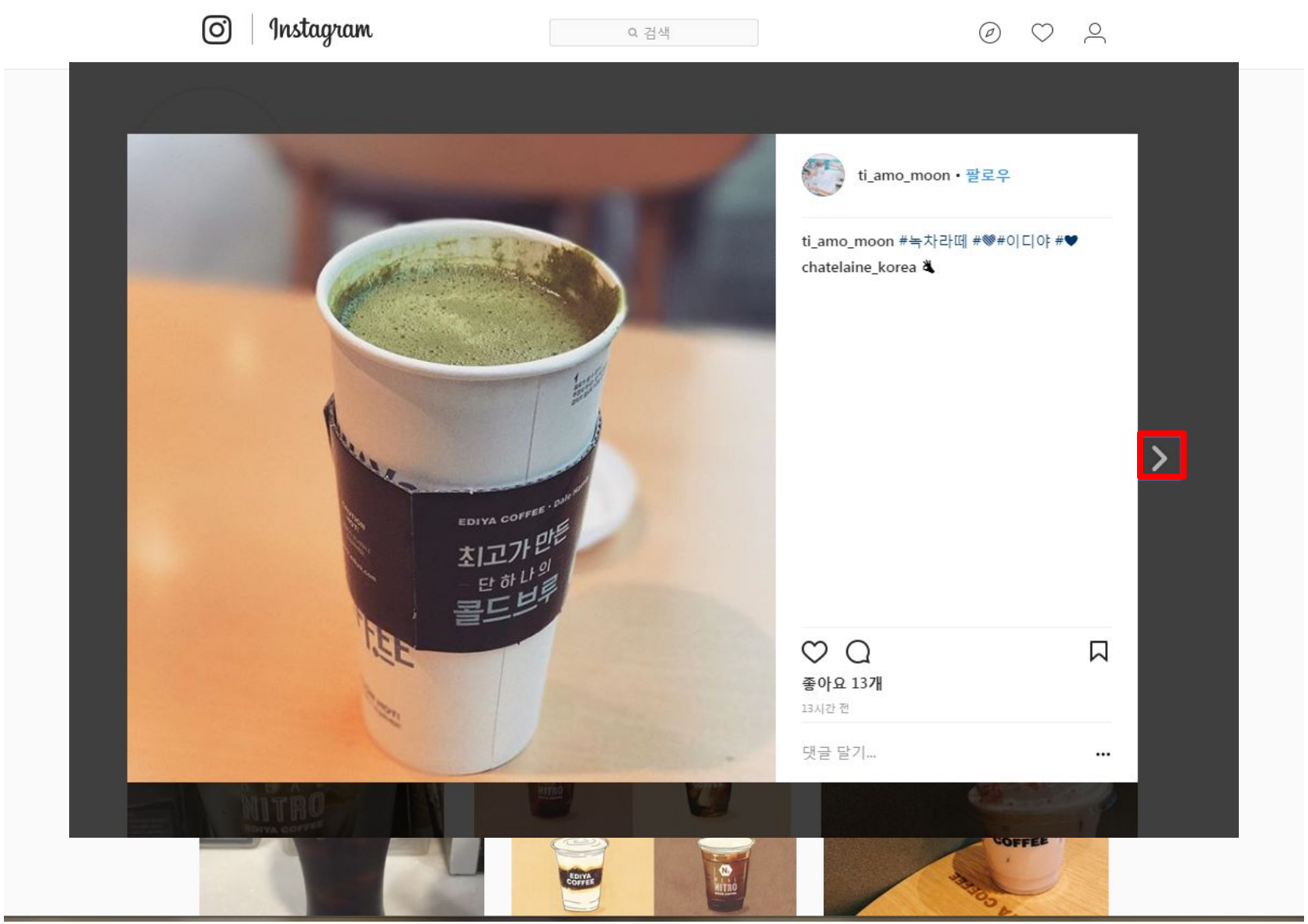
사라진 게시글

多

I - 2. 데이터 수집


Instagram
검색
9009da • 팔로우
9009da
이건 꼭 사야되!!!
아침부터 #스벅 다녀왔어요!
여름시즌 한정 #피크닉매트
주머니 전용 파우치안에 쏙 들어가니 좋구만!
넉넉한 145 *145cm 빅사이즈라서 진짜 대박ㅁㅁ
일러스트가 진짜 넘넘 사랑스러워용
완전 내스타일임! 완전 #소장각
특특한 재질의 소재라서 써본사람만이 알수있답니당!
한정판이니 빨리 고고고!
퀄리티가 진짜 넘 좋고 고급져서
난 집 벽면에 걸어둬야지
나나나 피크닉매트하나로 #스벅러버녀 됨ㅁㅁ
좋아요 3,605개
2일 전
댓글 달기...

I - 2. 데이터 수집


Instagram
검색
ti_amo_moon • 팔로우
ti_amo_moon #녹차라떼 #♥#이디야 #♥
chatelaine_korea
좋아요 13개
13시간 전
댓글 달기...
EDIYA COFFEE
최고가 만든
단 하 나 의
콜드브루

Ⅰ - 3. 데이터 수집 코드

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
driver = webdriver.Chrome('C:\selenium\chromedriver.exe')
driver.get('http://www.instagram.com')
driver.find_element_by_css_selector('#email').send_keys('*****@*****.***')
driver.find_element_by_css_selector('#pass').send_keys('*******')
driver.find_element_by_css_selector('#loginbutton').click()

for i in ['이디야']:
        driver.find_element_by_css_selector('....').send_keys(i)
        time.sleep(random.randrange(1,3))
        driver.find_element_by_css_selector('....').send_keys('\n\n')
        ...
        text1 = []
        if count != 0:
                driver.find_element_by_css_selector('....').click()
                for j in range(1, int(count)):
                        try:
                                text1.append(....').text)
                        except Exception as error:
                                time.sleep(random.randrange(1,3))
                                driver.find_element_by_css_selector('...').click()
                                j += 1
                with open('C:/…..txt','w',encoding='UTF-8') as file:
                        for j in text1:
                                file.write(j+'\n')
```

4月 3日 ~ 4月 5日

I - 4. 데이터

nam_yeeeeh갤럭시s9+죠흐당♡♥
.
.

#컵홀 #커플폰 #기념일 #3주년 #갤럭시s9플러스 #아이폰언제써보냐 #듀얼카메라 #화질고와 #기념셀카 #귀요미 #스노우카메라 #봄 #라이더자켓 #영화 #곤지암 #제발좀가지말라고 #말좀들어라 #존나무서워 #가장맛있는족발 #육회공작소 #마이쩡 #먹스타그램 #데일리 #일상 #데이트 #좋아요 #이디야 #당보충 #바닐라라떼
umcertain장조림과 수육사이 차슈. 문동이 일본가서 사온 라멘. 반숙을 원했지만 완숙이 된 계란. 드디어 사용한 스티치 접시. 이디야 치즈스틱케익이 없어서 초코 구입. 고양이 젓가락받침대와 많이 세탁해서 바랜 밤색 린넨. 집에는
버스커버스커 재생중.주말오전
the_court_clerk8#20180323
아침부터 혼자 바쁨돋았음 ㅎㅎ
동전지갑 득템하고요(이어폰 케이스로 찰떡)

#컵홀 #커플폰 #기념일 #3주년 #갤럭시s9플러스 #아이폰언제써보냐 #듀얼카메라 #화질고와 #기념셀카 #귀요미 #스노우카메라 #봄 #라이더자켓 #영화 #곤지암 #제발좀가지말라고 #말좀들어라 #존나무서워 #가장맛있는족발 #육회공작소 #마이쩡 #먹스타그램 #데일리 #일상 #데이트 #좋아요 #이디야 #당보충 #바닐라라떼

Ⅱ. 데이터 정제

#인스타그램 #데일리 #일상 #daily #dailylook #ootd #instadaily #selfie #selca #like4like #셀스타그램 #얼스타그램 #셀카 #셀피 #카페 #cafe #카페스타그램 #이디야

#친구 #일상 #거울샷 #폭풍업뎃 #셀스타그램 #이디야 #카페 #시험기간 #고딩 #좋아요반사 #좋아요 #맞팔 #소통 #selfie #cafe #mirrorselfie #l4l #f4f #like4like

#이디야카페 #이디야커피 #카페스타그램 #이디야 #손그림 #그림 #art #doodle #penart #drawing #ediyacoffee #ediya #followme #like4like #좋아요반사 #followforfollow #선팔하면맞팔 #언팔❌ #낙서

"제발 통일 좀 해줘..."

#99 #00 #20 #19 #고딩 #여고생 #수원역 #인친 #데일리 #일상 #소통 #셀스타그램 #selfie #like4like #likeforlike #followme #follow #고3 #셀기꾼 #팔로우 #팔로미 #모모스테이크 #이디야

#일상 #일상그램 #데일리 #데일리그램 #데일리룩 #데일리라이프 #라이프 #라이프그램 #daily #dailygram #dailypic #dailylook #daiylife #dailylifegram #starbucks #스타벅스 #스벅md #스벅텀블러 #스벅 #스타벅스투톤핑크사이렌보온병 #그러데이션사이렌핸들글라스 #그라데이션사이렌핸들글라스

Ⅱ-1. 인스타그램 약어 정제

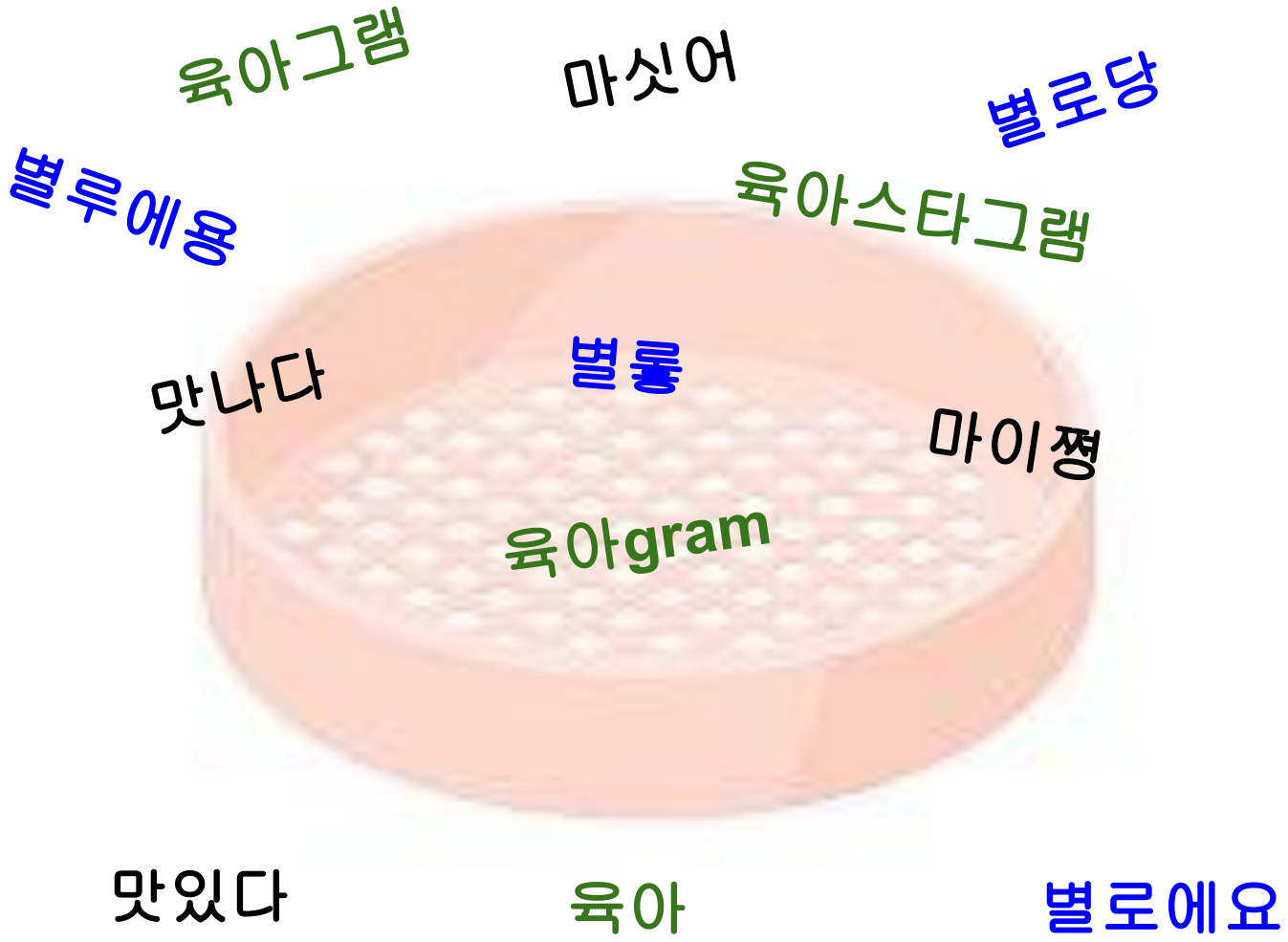
Ⅱ-2. 불필요한 단어 정제


dailylook
fff
ootd
맞팔선팔
followme
l4l
like4like
lfl
follow4follow
f4f
out fit of the day


육아그램
아기스타그램
lovestagram
럽스타그램
육아스타그램
babystagram
연애스타그램
애스타그램
육아gram
아기
육아
연애중


カフェモカ
ソウル
caffemocha
さくら
櫻花
Seoul
beotkkot
汉城
Caffè mocha
서울
카페모카
벚꽃

Ⅱ-5. 기타 정제



Ⅲ. WORD CLOUD 1

STARBUCKS

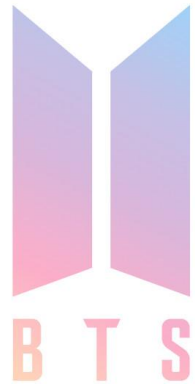
EDIYA

Ⅲ. WORD CLOUD 2

STARBUCKS

EDIYA

공통적으로 사진, 셀피 多



스타벅스

친구
방탄소년단

이디야

육아
연애중

STARBUCKS & EDIYA

STARBUCKS


친구랑 함께하면
다양한 혜택들이
쏟아진다!


BTS Score Top 10 Debut on Billboard 200 With 'Love Yourself: Her' Album
9/24/2017 by Jeff Benjamin , Keith Caulfield
Courtesy of Big Hit Ent
BTS
BTS
BTS_official
@bts_bighit
[공지] 방탄소년단 세 번째 정규앨범 LOVE YOURSELF 轉 'Tear' 발매 안내

EDIYA

이구동성
퀴즈 이벤트

THANK YOU