2020 인공지능 데이터셋 경진대회

인스타그램 해시태그 기반의 감성글 데이터셋

인스타갬성터짐 조
대표 안영진

CONTENTS

01 제안 배경
02 데이터셋 소개
03 크라우드 소싱 활용 방안
04 기대 효과

필요성 및 목적

자연어처리 영역에서 영어는 높은 완성도의 방대한 데이터셋들이 다수 존재

but, 한국어 데이터셋은 상대적으로 부족

Jigsaw

Google Q&A

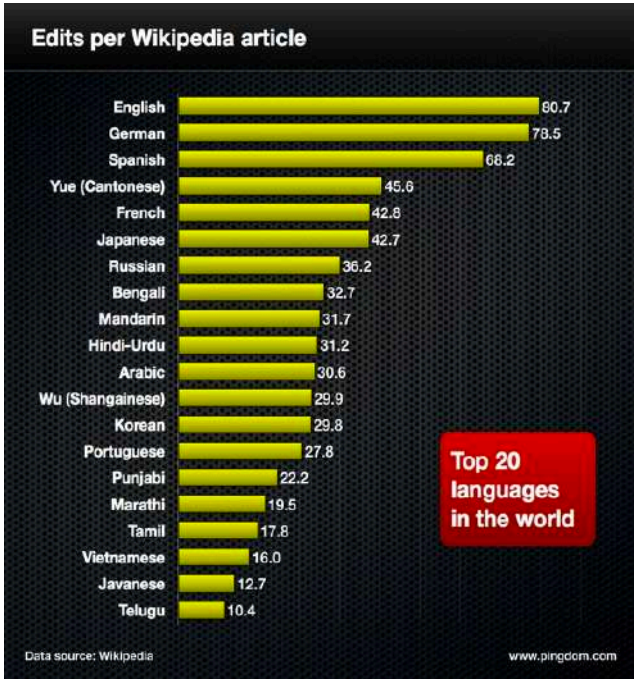
NLTK

SQuAD2.0

The Stanford QA Dataset

VQA

Visual Question Answering



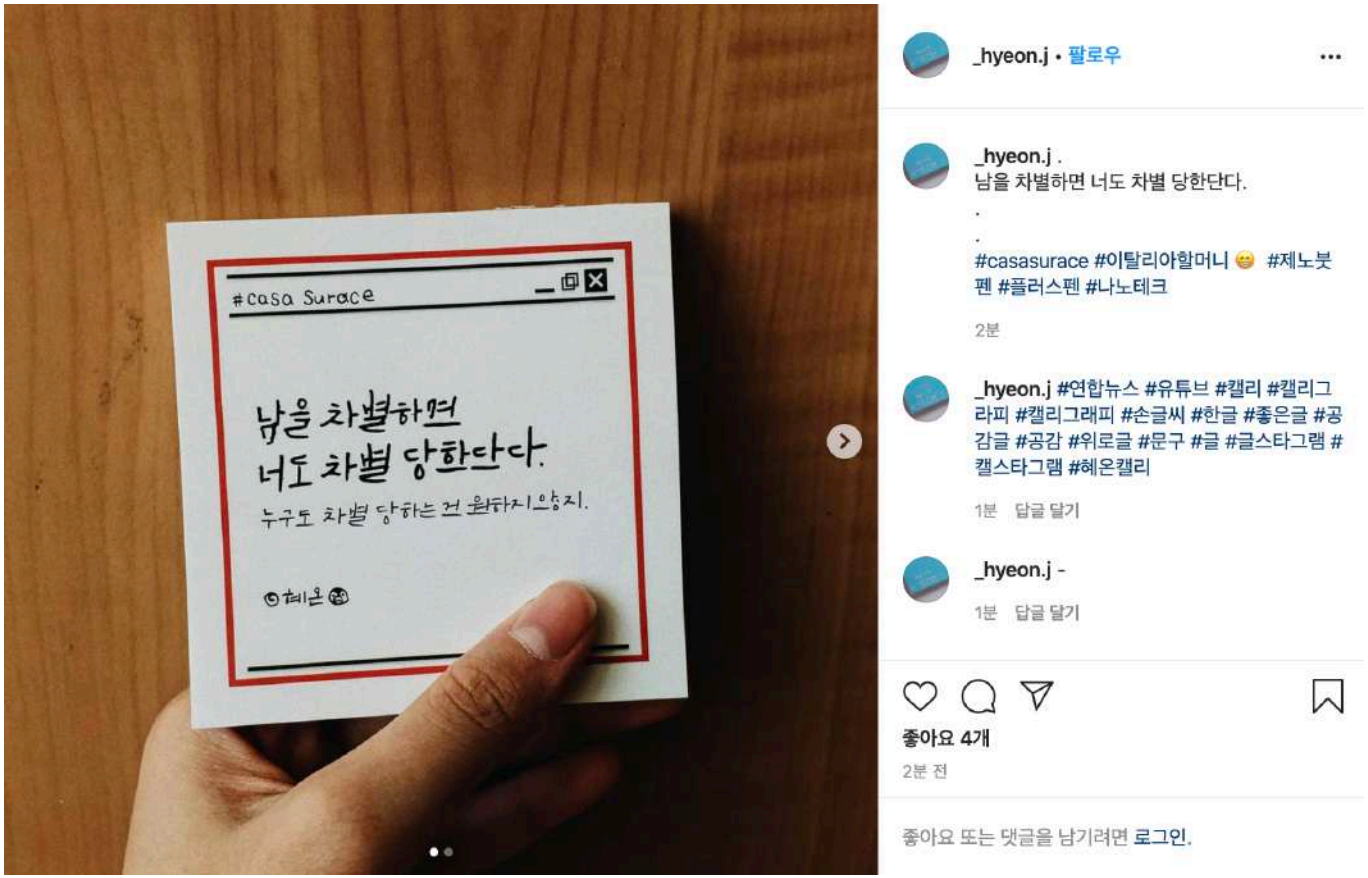
인스타그램 선정 이유

현재 국내에서 가장 활발히 이용되는 소셜 미디어
문학 작품을 표상하는 키워드를 해시태그로 사용

국내 소셜 미디어 앱 이용자 수 비교

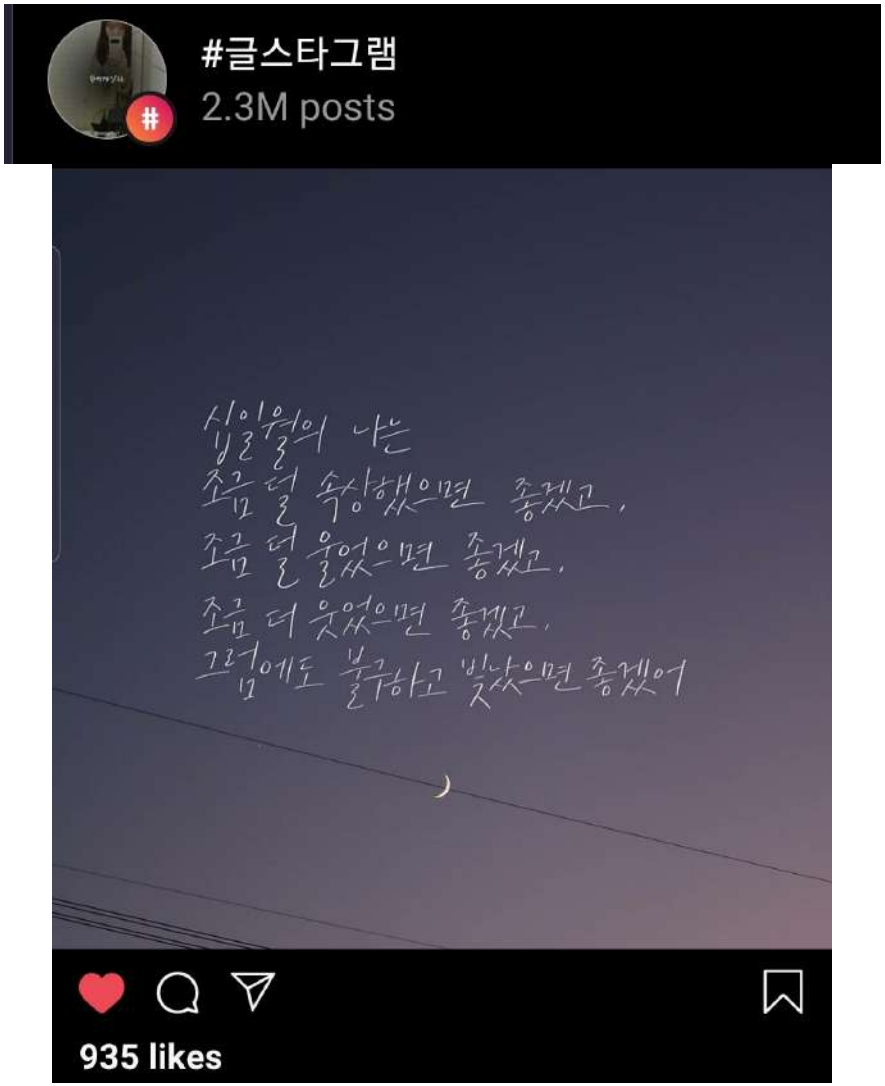
플랫폼 선정 이유

다수의 대중 문학 작품 확보 및 해시태그를 통해 작품의 핵심 주제 파악 가능



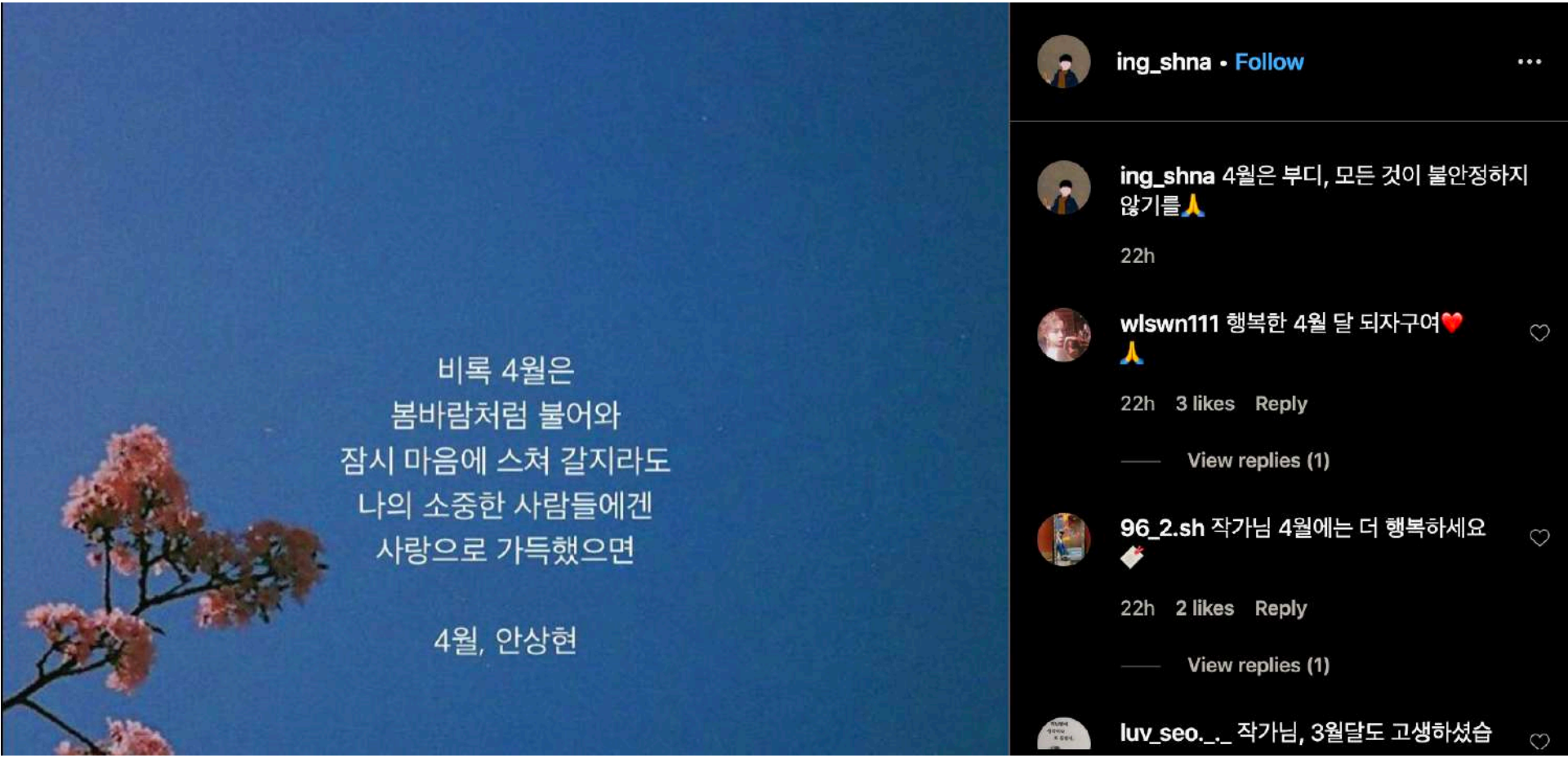
02
데이터셋 소개

용어 설명



글스타그램 : 짧은 글귀를 이미지(JPG, PNG) 형식으로 인스타그램에 업로드한 게시글

용어 설명



- 본문 텍스트 : 글스타그램과 함께 쓰인 글
- 댓글(코멘트) : 글스타그램에 달린 댓글 혹은 답글
- 인플루언서 : 글스타그램을 위주로 올리는 사람

데이터셋 소개

해시태그 기반으로 게시물 검색이 이루어지는 인스타그램의 특성을 이용

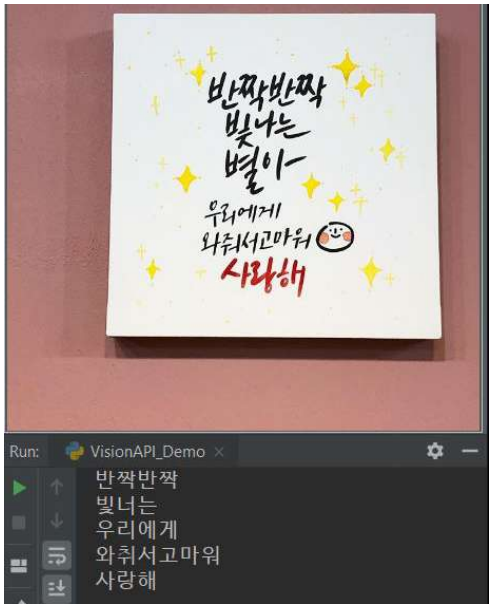
#글스타그램 데이터를 대량 수집 및 가공한 데이터셋

유저 ID				
유저 ID	게시글 ID	글 본문	이미지 OCR Text	해시태그
(string)	(URL Query)	(string 통째로)	(string)	(list of strings)
"jaulounge"	"B22i52NFTNc_1"	"비에 맞는 시를 한 번 써봤어요"	"비 온다니 장롱이 **뼈걱뼈걱** "	["#일상", "#사랑"]

데이터셋 소개

OCR

이미지를 글로 변환



#글스타그램

짧은 글귀를 이미지(JPG, PNG) 형식으로

인스타그램에 업로드한 게시글

본문 텍스트

#글스타그램 이미지에서

OCR로 추출한 텍스트

기존 데이터셋과의 차별성

- 우리 말의 서정적 정서가 담긴 글
- 여러 작가의 작품을 통해 규격화되지 않은 코퍼스 구축 가능
- 범용적이면서도 독창적인 시구 사용으로 다른 한국어 데이터셋과는 차별화된 글귀 수집
- 게시글 작성자가 작품과 함께 기재하는 키워드들을 통해 반자동적으로 글의 주제를 레이블링함으로써 어노테이션 효율화

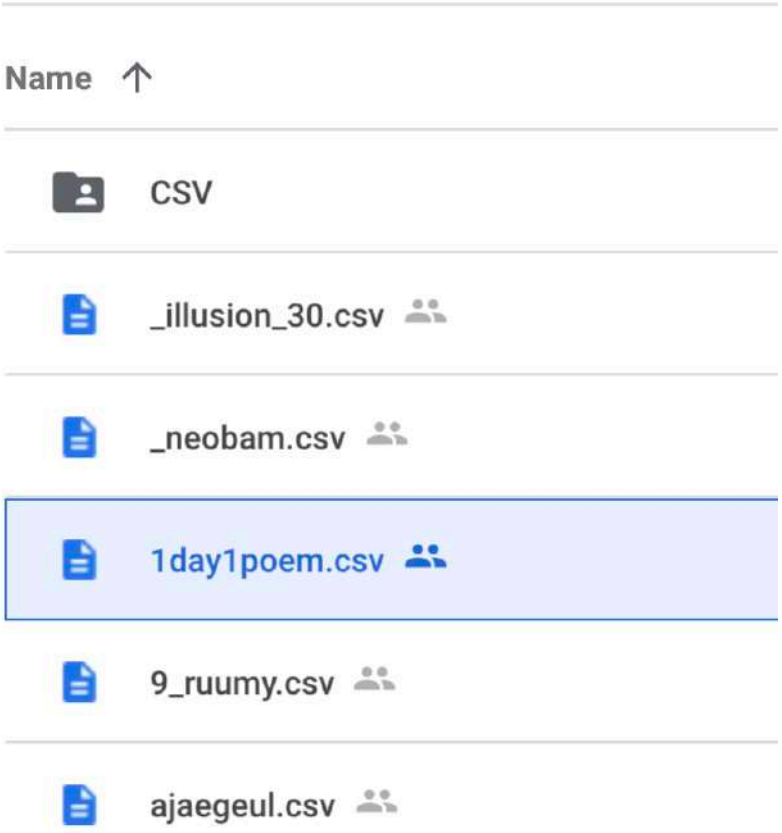
데이터 구축

인스타그램 크롤러 Instaloader를 이용하여
"글스타그램" 해시태그를 가진 게시글 자동 수집

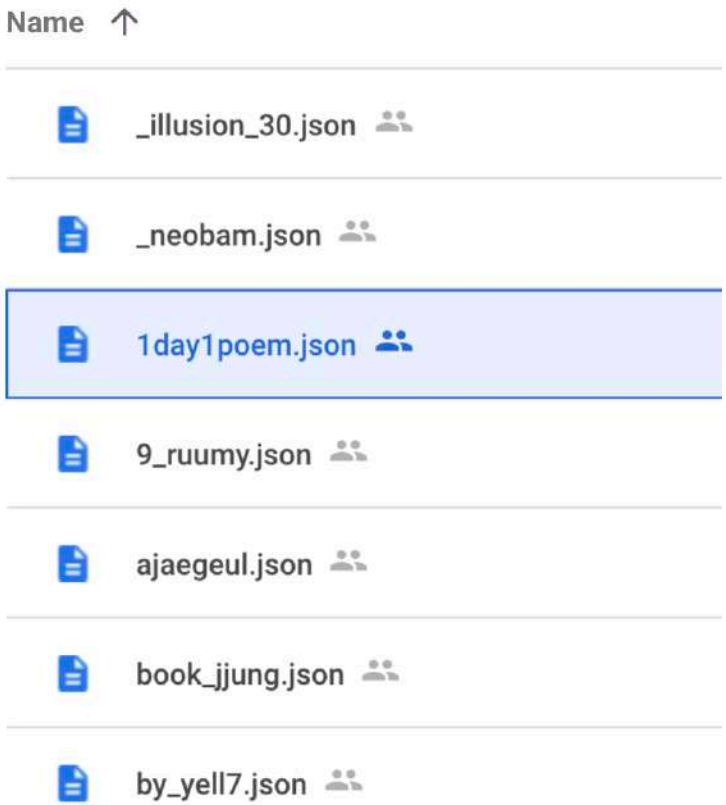
```
import instaloader

def getContentsWithTags(searchKeywordTag, numbersOfPosts):
    #get contents only
    L = instaloader.Instaloader(download_pictures=False,
                                download_video_thumbnails=False,
                                download_videos=False,
                                download_geotags=True,
                                download_comments=False,
                                save_metadata=False)

    while numbersOfPosts > 0:
        for post in L.get_hashtag_posts(searchKeywordTag):
            L.download_post(post, target='#'+searchKeywordTag)
            numbersOfPosts-=1
            # print(numbersOfLoop)
            if numbersOfPosts ==0:
                break
```

데이터 구축

인스타그램 크롤러 Instaloader를 이용하여
"글스타그램" 해시태그를 가진 게시글 자동 수집



Text Data



Meta Data

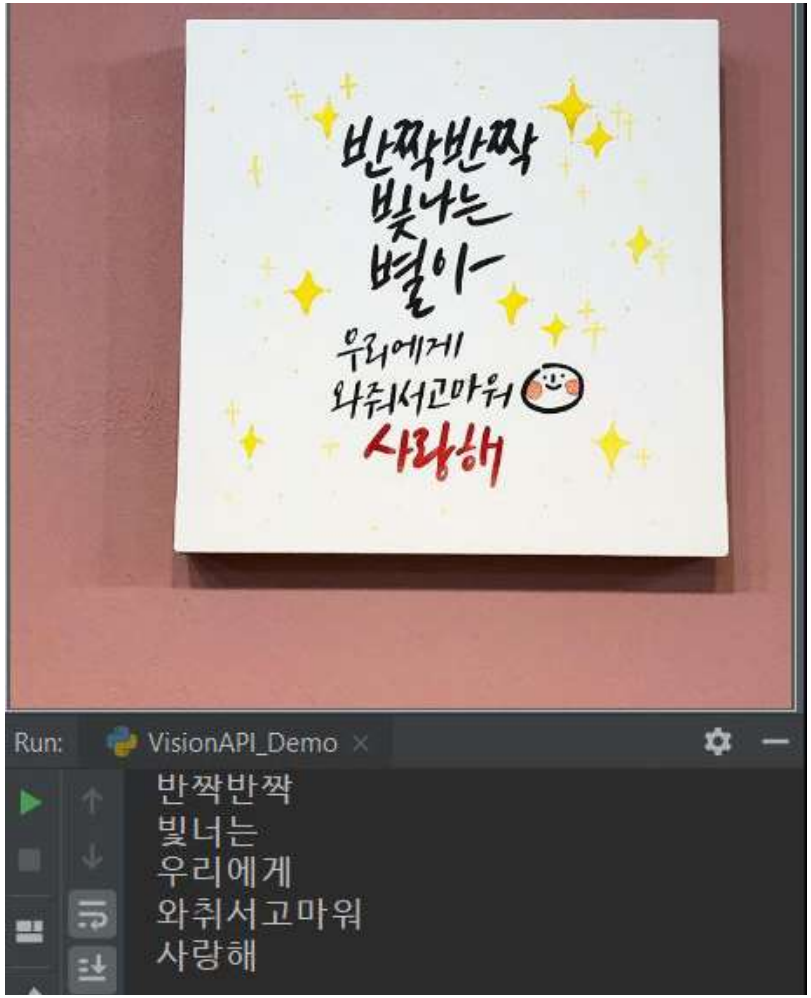
데이터 구축

Meta Data 기반으로 글스타그램 감성 테마에 맞는 데이터셋 정제


```
"__typename": "GraphSidecar",
"accessibility_caption": "Image may contain: text",
"caption_is_edited": false,
"commenting_disabled_for_viewer": false,
"comments_disabled": false,
"dimensions": {
    "height": 1080,
    "width": 1080
```

데이터 구축

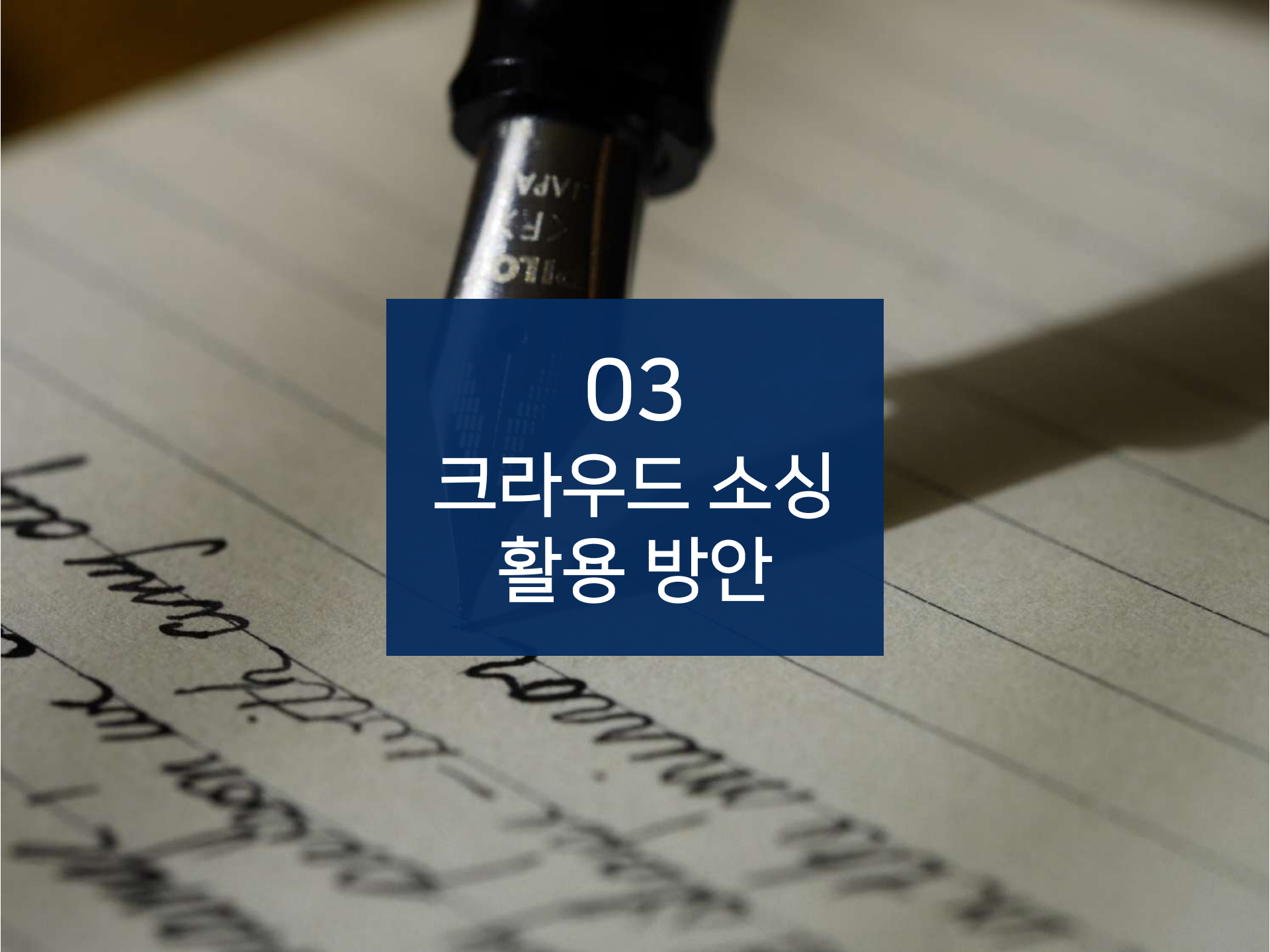
Google Cloud Vision API를 활용하여 게시글 사진 내 텍스트 추출



데이터 구축

Torch 기반 Chatspace로 맞춤법 교정

원래 문장	띄어쓰기를 수정한 문장	Mecab 결과	이전 결과
저녁머글준비즁	저녁 머글준비즁	저녁 머글 준비 즁	저 녁머글준비즁
왱 항댸마신다니까	왱 항댸 마신다니까	왱 항 댸 마신 다니까	왱 항 댸마신다니까
공부개열심히하네	공부개 열심히 하네	공부 개 열심히 하 네	공부 개열 심히 하 네
헤헤 비안올거예요	헤헤 비 안올거에요	헤헤 비 안 올 거 에요	헤헤 비안 올 거 예요
굿모닝 난가는중	굿모닝 난 가는 중	굿모닝 난 가 는 중	굿모닝 난가 는 중



03
크라우드 소싱 활용 방안

크라우스 소싱을 통한 수집 및 가공

대상 데이터의 확정: 반자동 수집된 글스타그램 텍스트 및 해시태그

원본 문장 :

너를 두고 망설인 그 찰나가 내 평생 가장 큰 사치였다.
늘 그렇듯 분에 넘친 소모가 남기는 것은 깊은 **후회**뿐이다.
값비싼 찰나 비싼 대가.

데이터 가공 방식

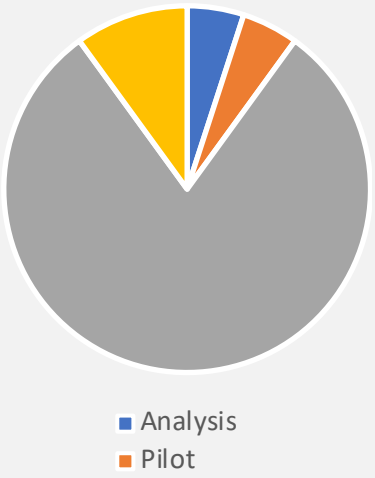
1. 이미지를 참고하여, 자동화로 정제되지 않은 컨텐츠 교정
   - 작품 당 1인

2. 게시글 기반의 해시태그 확장 및 확정
   - 작품 당 최소 2인
     ex) 사치, 소모, **후회**, 대가

크라우스 소싱을 통한 수집 및 가공

데이터 분배

1. 가이드라인 제작을 위한 **데이터 분석** (5%)
2. 주석자 선정을 위한 **파일럿 스터디** (5%)
    - 파일럿에서 좋은 성적을 거둔 주석자를 기용
3. 크라우드소싱에 활용하는 **최종 데이터셋** (90%)
    - **Train 및 Validation**에 80% 활용
    - **Test**에 10% 활용



크라우스 소싱을 통한 수집 및 가공

파일럿 연구

1. 검토 과정: **맞춤법, 교정의 섬세함** 등 측면에서 리뷰어들에게 높은 점수를 얻은 주석자들을 선정하여 최종적으로 기용

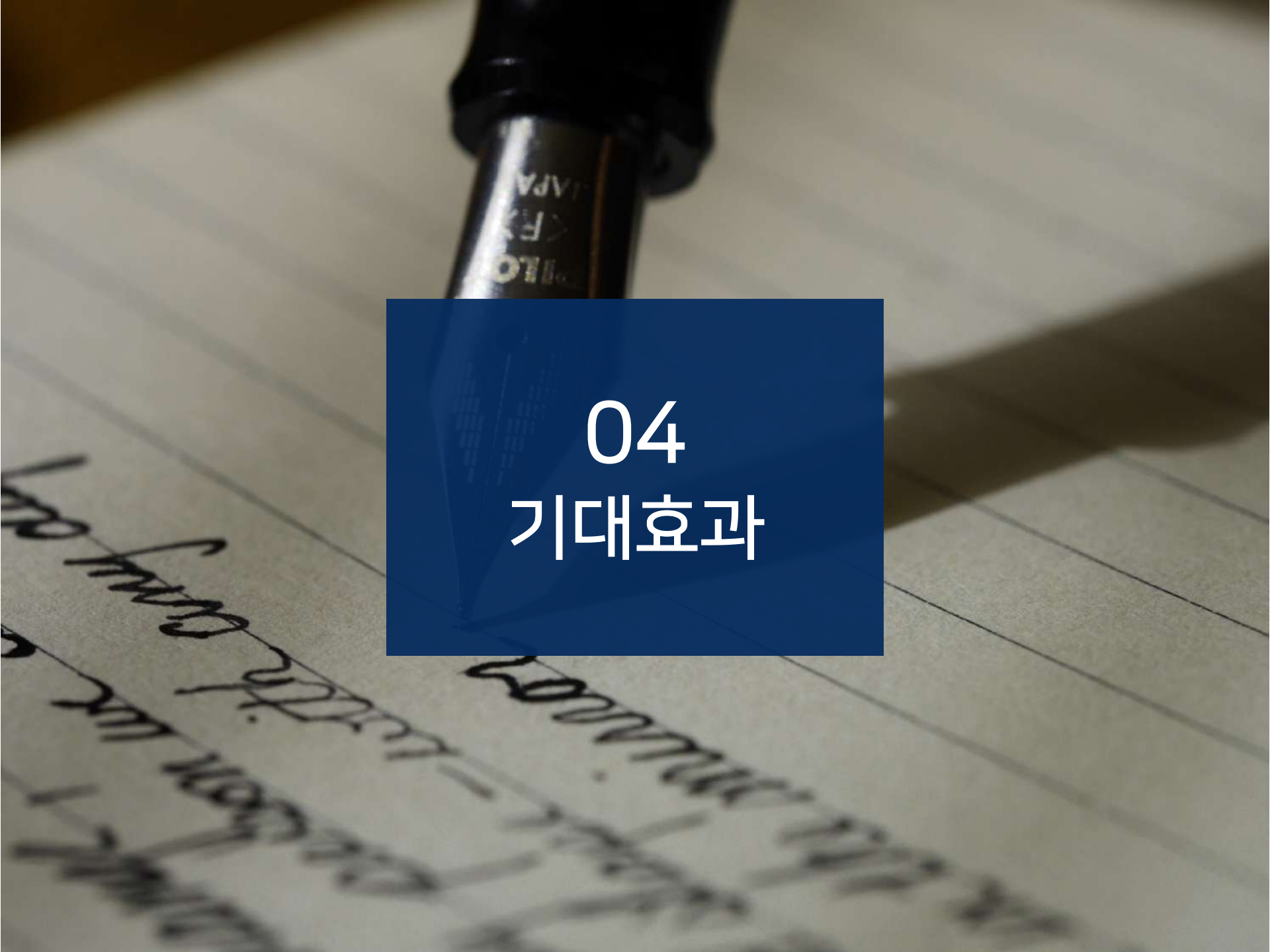
2. 소분류 태깅: 기존 해시태그들과 **의미나 형태적으로 비슷한 태그들**을 많이 추출 및 고안하는 주석자들을 기용

최종 데이터 결정 및 배포 포맷

1. 검토 과정: 샘플링 검수 이외 별도의 최종 결정 없음
   - 정제된 글스타그램 텍스트 및 해시태그 목록 취득 (우선 시행)

2. 소분류 태깅:
   1. 최소 두 명의 주석자를 통해 태그 목록 확보
   2. 반자동수집 및 정제된 태그와 유사한 것을 우선적으로 메타데이터로 남김
   3. 텍스트 당 최대 10개까지의 태그 기록

   - 글스타그램 텍스트와 그에 따른 해시태그 각 최대 10개 배포



04
기대효과

기대 효과

트렌드에 민감한 인스타그램의 특성을 고려하여
사회 현상의 분석 자료로 활용 가능

Twitter Reveals: Using Twitter Analytics to Predict Public Protests

Mohsen Bahrami [1,2 *], Yasin Findik [3], Burcin Bozkaya [2], Selim Balcisoy [3]

[1] *MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA*

Predicting consumer sentiments from online text

Xue Bai ✉

⊞ Show more

https://doi.org/10.1016/j.dss.2010.08.024 Get rights and content

기대 효과

Hate speech detector 개발에 활용 가능

Instagram Comments Dataset for Hate Speech Detection

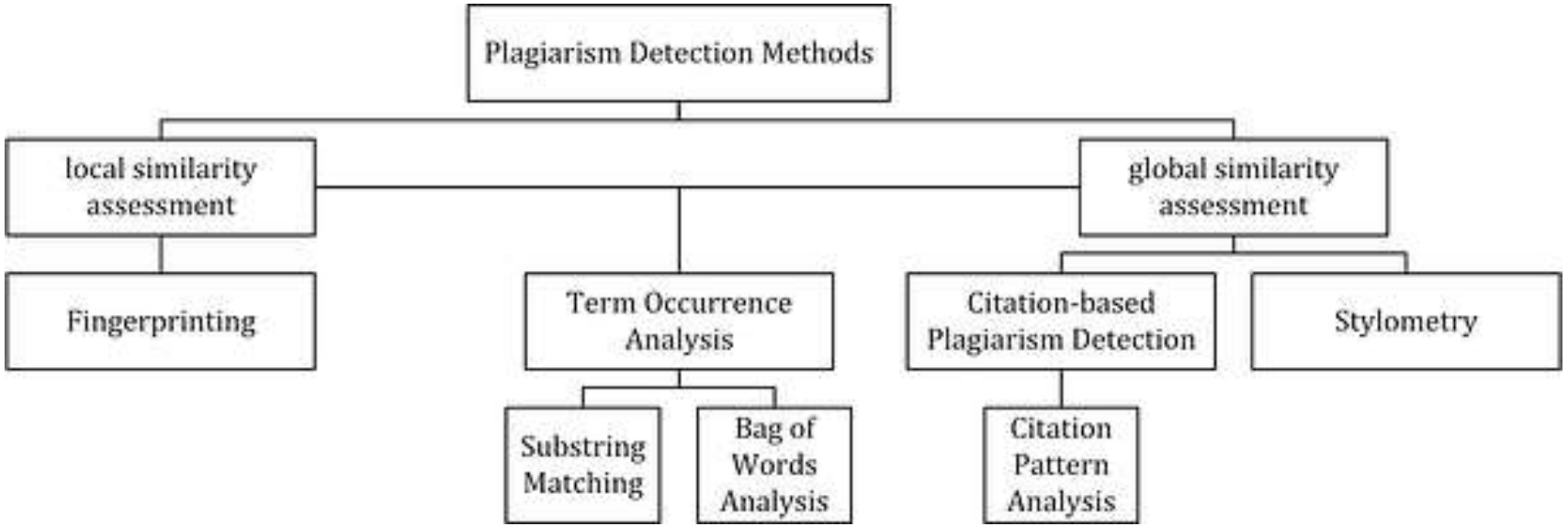
This dataset of Instagram comments is used as source of study "Hate Speech Detection on Indonesian Instagram Comments using FastText Approach".

Hate Speech Detection on Twitter

Implementation of our paper titled - "Deep Learning for Hate Speech Detection" (to appear in WWW'17 proceedings).

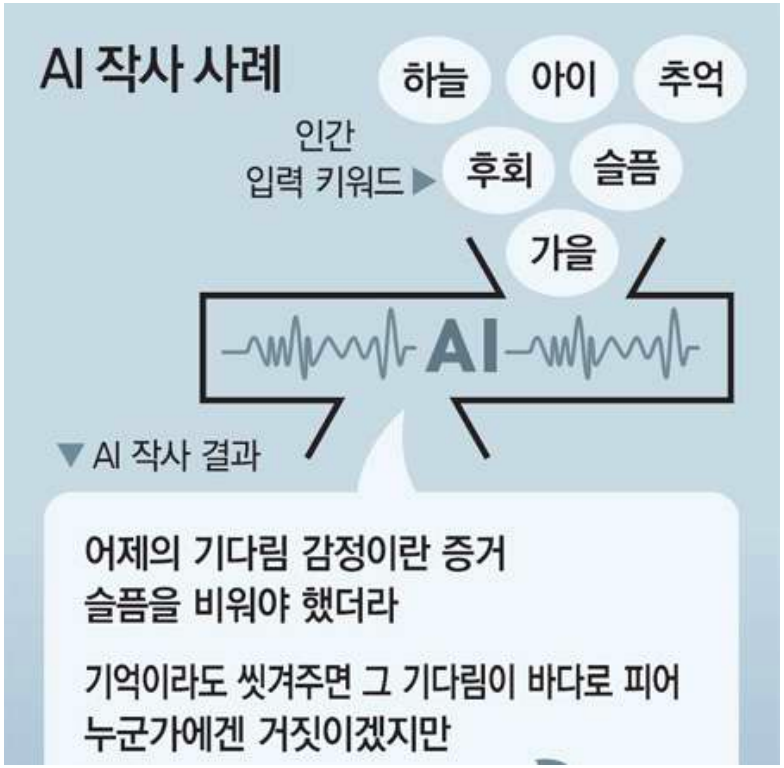
기대 효과

문학 작품 표절 및 가이드라인 검사에 활용 가능



기대 효과

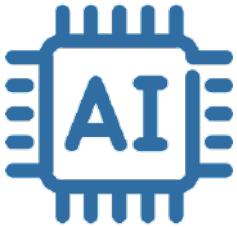
새로운 장르의 문학 신설 및 문학 영역 확장 및 활성화




```python
def on_epoch_end(epoch, _):
    print('\n----- Generating text after Epoch: %d' % epoch)

    start_index = random.randint(0, len(text) - maxlen - 1)
#     for diversity in [0.2, 0.5, 1.0, 1.2]:
#         print('----- diversity:', diversity)

    generated = ''
    sentence = text[start_index: start_index + maxlen]
    generated += sentence
    print('----- Generating with seed: "' + sentence + '"')
    sys.stdout.write(generated)
```

Keras 단편 소설 제작 모델

기대 효과



- 구어체/문어체 구분 인공지능 학습 데이터로 활용
- 감성적 표현 위주 문학 최적화 시스템을 위한 인공지능 연구 활성

문어체

구어체

나는 서울 대학교 학생이다	저는 서울 대학교 학생입니다
나는 영국 노래를 좋아한다.	영국 노래 좋아하지.

기대 효과

문체, 어구 및 성향 등에 대한 파악이 가능해
최신 문학 관련 연구 데이터로 활용 가능

http://dx.doi.org/10.9717/kmms.2015.18.11.1391

인스타그램 해시태그를 이용한 사용자 감정 분류 방법

남민지[†], 이은지[††], 신주현[†††]

A Method for User Sentiment Classification using Instagram Hashtags

Minji Nam[†], EunJi Lee[††], Juhyun Shin[†††]

End of Document