**수집한 해시태그 분석 현황 및 실험 진행표 (20.12.04)**

1. **데이터 수집**
2. GCPMuseumData

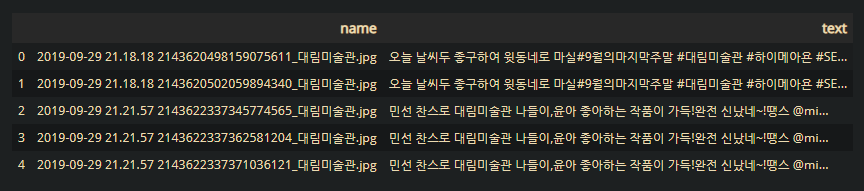
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Image** | **Image w.text** | **Distinct text** | **Cleansing text** | **date** |
| #britishmuseum | 11187 | 11166 | 3988 | 3953 | 16.05.18~19.10.14 |
| #centrepompidou | 14552 | 14498 | 6857 | 6823 | 19.08.12~19.09.27 |
| #metropolitanmuseumofart | 15242 | 15194 | 5418 | 5401 | 19.08.09~19.09.24 |
| #moma | 14314 | 14185 | 7666 | 7605 | 19.09.19~19.10.08 |
| #museedulouvre | 15434 | 15303 | 7093 | 7003 | 11.08.07~16.11.27 |
| #tatemodern | 13280 | 13225 | 5484 | 5410 | 19.09.10~19.09.24 |

1. GCPKoreanMuseumData

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Image** | **Image w.text** | **Distinct text** | **Cleansing text** | **date** |
| #daelim | 10402 | 10345 | 2145 | 2078 | 19.09.29~19.11.02 |
| #leeum | 10977 | 10854 | 3950 | 1585 | 18.04.16~19.11.02 |
| #mmcaseoul | 10339 | 10230 | 2497 | 2425 | 19.08.04~19.11.02 |
| #museumkorea | 10841 | 10805 | 2347 | 2325 | 19.10.05~19.11.02 |
| #nfmkorea | 10369 | 10341 | 2131 | 2112 | 19.03.17~19.10.31 |

* 각 column 설명
  + Name : 박물관 이름
  + Image : 수집한 이미지 개수
  + Image w.text : 수집한 이미지 중 텍스트 포함된 개수
  + Distinct text : 수집된 서로 다른 텍스트 개수
  + Cleansing text : 전처리 완료 된 텍스트 개수
  + Date : 수집된 데이터의 날짜

1. **데이터 전처리**
2. 데이터 csv 파일 예시 및 사용 프로그램



* name : 수집된 인스타그램 이미지의 이름
* text : 해당 이미지에 포함된 텍스트
* python 라이브러리인 pandas 를 사용하여 데이터 전처리 진행

1. 텍스트 중복 제거

* 인스타그램은 게시글(텍스트) 하나 당 여러 이미지를 올릴 수 있으므로 다른 name을 가진 중복된 text 가 존재
* 중복된 text 들 중, 가장 첫 번째 text만 남겨놓고 제거

1. 텍스트 클렌징

* ‘GCPKoreanMuseumData’는 국내 전시관을 수집하였으므로 주로 한글 text 포함
* ‘GCPMuseumData’는 해외 전시관을 수집하였으므로 주로 영어 text 포함. 다만, 아랍어 등 다양한 언어가 존재하나 그 개수가 영어의 수보다 작음
* 인스타그램 특성 상, 텍스트에 다양한 이모지(emoji) 를 포함하는 경우가 많음
* 따라서 #, emoji, 특수문자 등을 제거 후, 이중공백 처리 등을 통해 데이터 전처리 진행

1. 불용어(stopword) 제거

* Text 데이터에서 의미있는 정보들만 뽑아내기 위하여, 불용어(stopword)를 제거함
* 국립국어원에서 제공하는 ‘세종 코퍼스’ 에서 고 빈도어 상위 100개를 추출하여 불용어 리스트로 제작
* 불용어 리스트는 [형태 / 품사태깅 / 평균 빈도수]로 구성되어있음

1. **토픽 모델링(예시/진행중)**

수집된 데이터의 개략적인 정보를 파악하고자, 간단한 토픽 모델링 진행. 이후 수집된 데이터 특성에 맞게 방법을 수정 및 보완 예정

1. 품사 추출 및 단어장 생성

* 전처리 완료 된 텍스트 중, ‘명사(Noun), 동사(Verb), 형용사(Adjective)’ 3가지 품사로 된 단어들을 추출하여 단어장 생성

1. Nonnegative Matrix Factorization (NMF)

* Nonnegative Matrix Factorization (NMF) 을 사용하여 각 전시관 텍스트 데이터로부터 토픽 추출
* 예시

|  |
| --- |
| 텍스트이(가) 표시된 사진  자동 생성된 설명 |
| 텍스트이(가) 표시된 사진  자동 생성된 설명 |

* 각 전시관의 성격이 다름에 따라 추출된 토픽도 다른 양상을 띄고 있음

1. WordRank

* graph ranking 알고리즘을 이용하여 단어를 추출하는 WordRank 사용

1. **시각화(예시/진행중)**

추출된 토픽을 워드클라우드로 시각화 시험 진행. 이후 추가적인 분석및 시각화 방식을 도입 예정

* 각 전시관의 로고를 이용한 워드 클라우드 예시

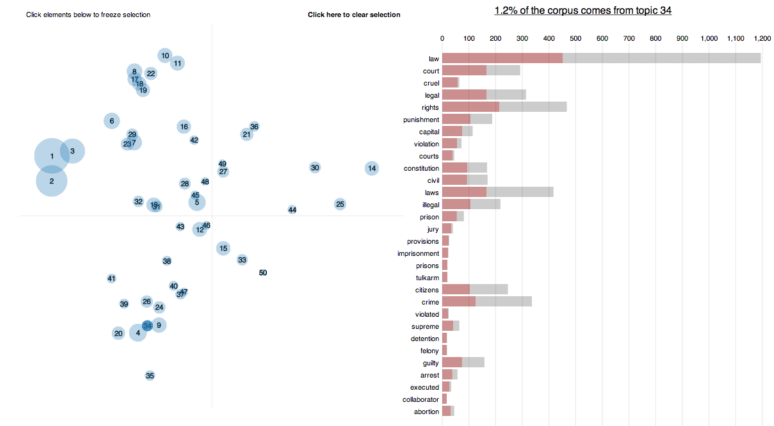
|  |  |
| --- | --- |
| 전시관 로고 | 워드클라우드 |
|  | 텍스트이(가) 표시된 사진  자동 생성된 설명 |
|  |  |
| 화살이(가) 표시된 사진  자동 생성된 설명 | 텍스트이(가) 표시된 사진  자동 생성된 설명 |
| 그리기이(가) 표시된 사진  자동 생성된 설명 | 텍스트이(가) 표시된 사진  자동 생성된 설명 |
|  | 텍스트이(가) 표시된 사진  자동 생성된 설명 |

- 워드클라우드 이외의 시각화 예시

1) 각 토픽 별 단어 비율

- 각 전시관별 나타나는 주제와 그 주제들에 포함하는 단어들 정렬

- 혹은 각 주제별로 포함되는 전시관 나타내기



2) 토픽-키워드 네트워크

- 비슷한 토픽이더라도, 전시관 별로 토픽-키워드 네트워크가 다를 수 있음

- 전시관 별로 비교분석

