빅데이터분석및실험 기말시험 문제

<2021. 12. 15. 14:00 ~ 15:50>
<open books/internet, close mind!!!>
<**주의: 교수 PC에서 여러분들의 화면을 monitoring 하고 있습니다. 메신저/메일 프로그램이 작동되면 즉시 해당 PC의 전원을 끄겠습니다**>
<각 문제는 제공하는 파일을 사용하세요>

제출물
- ULMS 의 16주차 <기말시험>에 등록
- 제출물: python(ipython notebook) 프로그램, 결과물(생성파일 등)
- 압축파일 하나로 제출: 압축파일이름 학번_이름.zip

1. pandas 및 visualization 사용 (20)
- 서울시주민등록인구_2020.xls, cctv_in_seoul_2020.csv, crime_in_seoul_2020.xls
- 서울시 구별 주민등록인구, CCTV 설치 현황 및 관서별 범죄(발생, 검거) 건수의 csv 파일을 이용
- 구별 전체인구와 CCTV 설치 비율을 구하고 이를 수평막대 그래프(비율내림차순)으로 표현하라. (5)
- crime_in_seoul_2020 데이터를 minmax 정규화를 이용하여 변경하라(5)
- 2020년 구별 전체 범죄율 및 검거율과 인구수 및 CCTV수와의 관계를 pairplot으로 나타내라 (10)

2. 네이버 검색 API를 이용하여 "빅데이터"의 검색 결과 분석 (30)
- 파일: naver_search_빅데이터.csv

(1) 각 description의 내용을 한국어 형태소 분석(Konlpy의 Okt 형태소분석기 사용) 하여 전체 description에 출현한 명사에서 상위 100개 단어(명사)를 빈도 순으로 출력하라 (10)

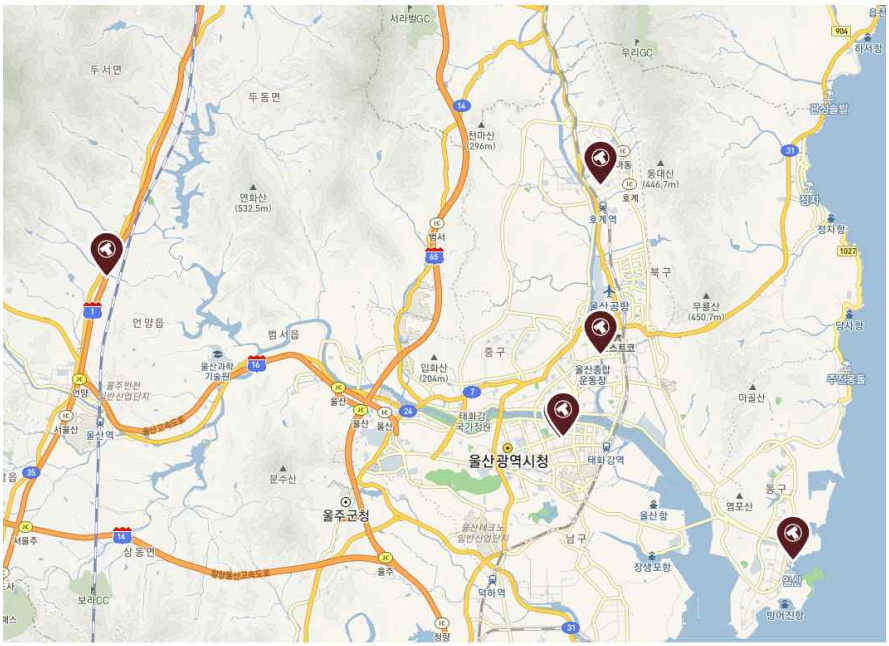
(2) (1)의 결과에서 상위 10개의 명사를 stopword로 하여 제거한 후 Word Cloud를 그려라. (10)

(3) (1)의 결과에서 “빅데이터”와 같이 사용된 상위 30개의 feature를 가진 TfidfVectorizer를 구하고 이를 기반으로 Word Network을 그려라. (10)

3. 지리정보 시각화 (20)

탐앤탐스 국내매장찾기(https://tomntoms.com/store/domestic_store_search.html)

- 파일: TomNToms_울산.csv
- 네이버 지도 API를 이용하여 "울산 남구" 지역의 매장을 지도 위에 나타내라 (20)
- NAVER 지도 API client ID가 없는 경우 config_NaverAPI_okcy.py 사용



4. 물품 추천시스템(20)

아래 표는 각 사용자(행, 15명)이 구입한 제품(열, 10개)의 별점의 행렬이다.
(별점이 0인 것은 구매하지 않은 제품)

- 파일: 추천시스템.csv

	i01	i02	i03	i04	i05	i06	i07	i08	i09	i10
user01	2	0	0	4	4	0	0	0	0	0
user02	0	0	0	0	0	0	0	0	0	5
user03	0	0	0	0	0	0	1	0	4	0
user04	3	3	5	0	3	0	3	2	2	0
user05	5	5	4	0	0	0	0	0	1	0
user06	0	0	0	0	0	5	0	0	5	0
user07	4	0	4	0	0	0	0	0	0	5
user08	0	0	0	0	0	4	0	0	0	4
user09	0	0	0	0	0	5	0	0	5	0
user10	0	0	0	3	0	0	0	4	5	0
user11	1	1	2	1	1	2	1	0	4	0
user12	3	0	0	3	5	0	0	0	1	3
user13	0	0	4	0	0	1	2	0	2	2

user14	2	2	3	0	2	1	0	0	0	1
user15	4	0	3	0	0	2	0	5	1	0

- i01과 가장 유사한 제품은? (Euclidian 유사도로 측정) (5)
- 위 사용자-물품 행렬에 대해 SVD을 구해서 전체에 대해 80%에 해당되는 특이값의 개수는? (5)
- user1이 구입하지 않은 제품에 대해 svd 평가 점수로 상위 3개의 제품을 추천하라 (10)

5. 강의 내용 중 어떤 부분이 부족했는가? 추가 보완되어야 할 내용? (10)

평가

- 출석 10% (출석은 2-7, 9-14주) webex 보고서 기반으로 5분 지각 (-0.5), 강의시간 1/2 불참의 경우 (-1.0), 결석 (-1.0)으로 하여 (20 + 지각결석)/2로 계산
- 중간시험 25%
- 기말시험 25%
- 과제 20% (과제 6개의 합계를 20%로 계산)
- term project 20%: (최종 발표순서 공지사항 참조)
  주제(관련도, 20), 완성도(계획서기준, 20), 복잡도(연관데이터, 20), 발표(20), 팀인원(20)
  위 항목으로 교수, 조교가 평가한 것의 평균