지역축제
방문객예측모델

소프트웨어융합학과
2017103728
배이지
2020-04-17

축제 데이터 결측 값 확인 및 채우기

1) 결측 값: 날짜, 시기, 등급
   => 인터넷 검색 후 직접 채우기

2) 축제 데이터 통일: python 파일

3) KTX 노선 연결: python 파일

4) 기상청 데이터 연결: 파이썬 **크롤링**을 이용
   - 기상청 데이터에 맞게 지역코드 생성 => 파이썬 이용
   - 축제가 같은 월인지 확인해 그 간격만큼 기상청 데이터 수집
   - 그 기간만큼의 평균기온과 일 강수량의 평균을 계산한다.

SNS 데이터 수집

1) 인스타그램 데이터 크롤링

- 파이썬 selenium 라이브러리를 활용해 content, hashtags, year 수집

=> 데이터 값이 너무 많고 인스타그램 내에서 **크롤링을 제한**

2) 트위터 데이터 크롤링

- **“GetOldTweets3”** 이용해 데이터를 수집

SNS 감정분석

1) content를 **형태소**로 나눈다.
   - 파이썬 konlpy 라이브러리를 활용해 형태소 나누기
   - 형태소: 의미를 가지는 최소단위

2) SNS 데이터 수집
   - 인터넷에 있는 불용어파일을 찾고 통합본을 만든다.

3) 형태소별 극성 계산
   - 군산대학교 감정사전에는 어근 및 단어별로 극성이 나와있다.
   - 어근 비교 및 단어비교를 같이 진행해 극성을 계산한다.

To do

1) 축제 데이터 결측 값 확인 및 채우기
   - 기상청 데이터 평균 **기온 및 평균 일 강수량 계산함수** 작성

2) SNS 데이터 수집
   - 데이터 수집 코드 작성 완료 축제 데이터 수집하는게 시간이 오래 걸림

3) SNS 감정분석
   - 감정분석 test 코드 적성 완료
   - **csv 파일 읽고 감정분석 후 데이터 추가하는 코드 추가**

대분류	중분류	소분류	설명	닐짜
표본선정	사용할 데이터 정리			O
데이터 수집 및 가공	축제데이터 결측 값 확인 및 채우기	날씨 및 시기, 등급, 세부지역 및 축제 이름 통일	Python	O
	기상청 데이터 연결하기			△
	KTX 노선 연결하기			O
	트위터 감정분석	트위터 content, hashtag, 언급 횟수 크롤링	Pyhon	△
		Content 감정분석	감성사전	△
변수 정의	독립변수 및 종속 변수 확정하기	트위터 언급횟수 ,긍정리뷰수 , 부정리뷰수		4/19
모델 구축 및 평가	데이터 분할			5/18
	변수 조합 탐색		R stepwise	5/18
	예측 모델 훈련	데이터 적합도 그래프		5/18
		다중선형회귀, 인공신경망, 서포트벡터머신	R	5/18
	예측모델평가	R-square		6/15
웹페이지구축	선택페이지	축제명, 날씨, 기간, 날짜 등을 선택	장고	6/15
	결과 페이지	예측 방문객 수, 내외국인 비율, 경제효과, 축제등급		6/15
		이전 년도 방문객 수 및 정보		6/15
		인스타그램 분석결과		6/15