지역축제
방문객예측 모델

소프트웨어융합학과
2017103728
배이지

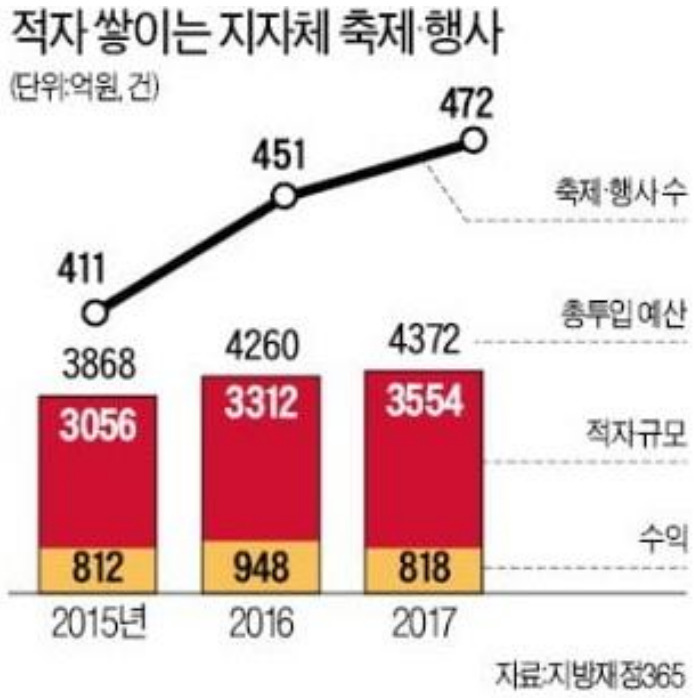
01 주제 선정 배경 및 문제점

지역축제의 규모가 커지는 만큼 **예산 분배를 더욱 효율적**으로 진행하고 **방문객 예측**을 통해 주차시설이나 휴게시설 증설 및 **지역축제 준비에 도움**을 줄 수 있을 것이다.

지방 > 지방일반

단양 소백산철쭉제 방문객 주차난·음식 불만 여전

등록 2019-07-16 13:21:10



02 관련논문

“축제 방문 수요 영향 요인 _ 권성수(2017)”

“ 방문객에 대한 방문동기로서 시장세분화에 관한 연구_손지군(2011, 배제대) ”

1) 존재하는 방문객 예측 논문은 대부분 **인구의 특성과** 관련된 연구가 많다.

2)방문객 예측 논문 외의 문화 및 미디어와 관련된 예측 연구는
SNS의 감정분석을 변수로 사용하는 연구가 많다.

⇒ **축제의 특징과 SNS 감정분석 결과를**
바탕으로 지역방문객을 예측

02 관련논문_선행연구

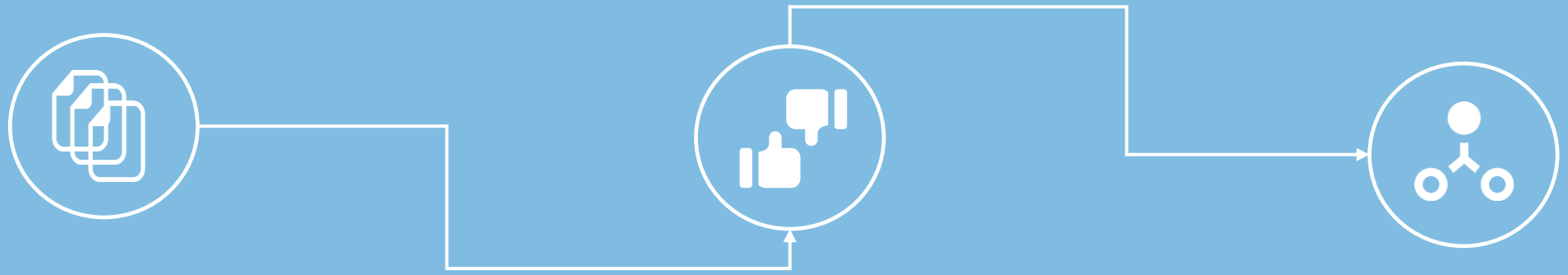
지역 축제에 영향을 미치는 요인

입장료 혹은 체험 프로그램이 비용이 존재하는 **유료 축제**에 완전 무료축제보다
평균적으로 많은 방문객들이 방문

방문축제 개최지에 **KTX의 정차역**이 존재하는 곳과 없는 곳 사이에도
평균적으로 많은 방문객 수가

축제기간에 따라 평균적으로 당일 개최되는 축제에 가장 적게 방문하는 것으로 나타났으나
기간이 장기화 될 수록 일일 평균 방문객 수가 증가하는 것은 아니었다.

03 연구 방법



데이터 수집 및 가공

1) 축제 관련 **데이터 수집**
2) 축제 데이터 **전 처리**
3) **SNS 데이터 수집**
4) **기상청** 데이터 연결
5) **KTX 노선** 유무 연결

감정분석

수집한
트위터 데이터를
감정사전을 이용해
분석한다.

알고리즘 작성

예측 알고리즘을
작성한다.
이후 **웹페이지** 작성

04 데이터 수집 및 가공

1) 축제 데이터 수집

- 축제정보_한국 지역진흥재단
- 지역축제_문화체육관광부
- 문화관광축제 주요 결과_관광지식정보시스템 통계게시판

➡ 장소, 기간, 지역, 등급 ,합계, 내국인 방문객 수 , 외국인 방문객 수, 경제효과

➡ 2007년부터 2017년까지의 500개의 데이터 사용

2) 데이터 전 처리

- 날짜 및 시기 찾기
- 축제 등급 채우기
- 세부 지역 채우기

04 데이터 수집 및 가공

3) SNS 데이터 수집

- **Instagram 데이터**를 selenium을 이용해 크롤링

➡ 인스타그램 내에서 크롤링을 제한

- **트위터 데이터** 크롤링

➡ **GetOldTweets3**를 이용해 데이터 수집

```
# 수집 기간 맞추기
start_date = days_range[0]
end_date = (datetime.datetime.strptime(days_range[-1], "%Y-%m-%d")
            + datetime.timedelta(days=1)).strftime("%Y-%m-%d") # setUntil이 끝을 포함하지 않으므로, day + 1

# 트윗 수집 기준 정의
tweetCriteria = got.manager.TweetCriteria().setQuerySearch(fes)\
                                           .setSince(start_date)\
                                           .setUntil(end_date)\
                                           .setMaxTweets(-1)

# 수집 with GetOldTweet3
print("Collecting data start.. from {} to {}".format(days_range[0], days_range[-1]))
start_time = time.time()

tweet = got.manager.TweetManager.getTweets(tweetCriteria)
```

04 데이터 수집 및 가공

4) 기상청 데이터 수집

- 기상청 데이터 url

"https://www.weather.go.kr/weather/climate/past_cal.jsp?stn={}&yy={}&mm={}&obs=1&x=18&y=6"

➡ **지역코드**를 필요로 하기에 기상청에서 직접 지역코드를 수집해 저장한다.

- 시작 달 종료 달이 같은 지 확인해 날짜만큼
  평균기온, 최고기온, 최저기온, 평균 운량, 강수량을 크롤링한다.

5) KTX 데이터 수집

- 코레일 홈페이지를 통해 직접 데이터 수집 및 연결

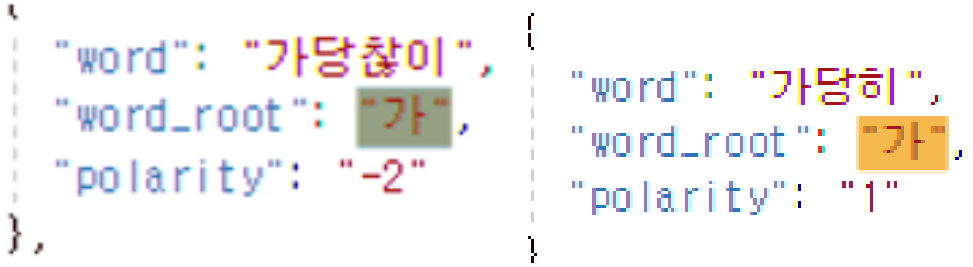
05 감정분석

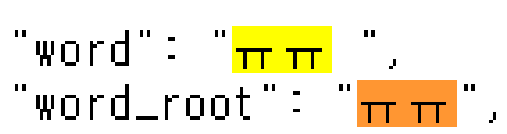
1) 감성분석 방법

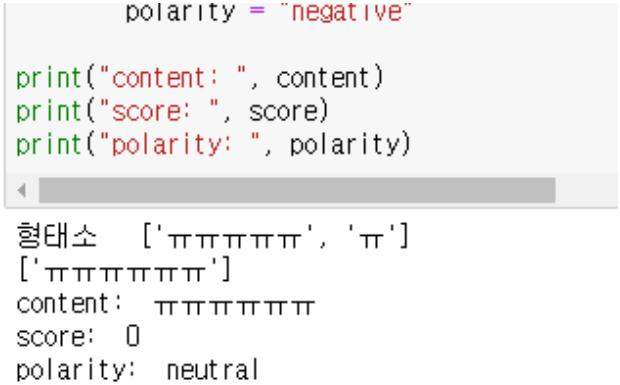
- 한글 감정분석은 연구 사례가 적고, labelling된 데이터가 없어 **KNU 감정사전**을 이용
- 감정사전 분석

  ➡ 사전이 형태소별로 구성되지 않는다. 이후 감정사전 형태소별로 재구축
- 이모티콘 정리

  ➡ 특정 개수가 넘어가면 인식이 불가능하다.







06 상관관계 분석

1) 상관관계 분석 (전체 축제 대상)

```
corr = total.corrwith(total['visitor'], method = 'pearson')
print("+++++","축제의 상관관계 분석+++++")
print(corr)
```

```
+++++ 축제의 상관관계 분석+++++
count         0.703857
visitor       1.000000
days          0.386138
avg_temp     -0.066663
high_temp    -0.010807
low_temp     -0.110431
cloud        -0.099353
rain         -0.096985
positive      0.033497
negative      0.056410
dtype: float64
```

- 부정비율 방문객에 미치는 영향보다 부정비율이 방문객의 양적 상관관계가 나타난다.

  ➡ 감정분석의 정확도 확인

06 상관관계 분석

2) 상관관계 분석 (개별 축제 대상)

```
+++++ 광주7080충장축제
visitor      1.000000
positive    -0.743096
negative    -0.777316
avg_temp     0.730638
high_temp    0.987279
low_temp     0.680117
cloud        0.030444
rain         0.950114
dtype: float64
```

```
+++++ 괴산고추축제 의 상관관계 분석+++++
visitor      1.000000
positive    -0.432466
negative    -0.815106
avg_temp    -0.181053
high_temp    0.302707
low_temp    -0.826671
cloud             NaN
rain        -0.522910
```

➡ 축제별로 진행해도 의미 없는 해당 데이터에 과적합된 상관관계 값이 나온다.

3) 데이터 재정의

- 감정분석의 정확도

➡ 더욱 세분화된 품사 태깅이 필요하다.

"달라진" => ('다르', 'P'), ('아', 'E'), ('지', 'P'), ('ㄴ', 'E')

```
wo:  ('지', 'P')
score:  -1
wo:  ('맞', 'P')
score:  -2
wo:  ('쓰겁', 'P')
score:  -2
wo:  ('맞', 'P')
score:  -3
negative
```

```
"word": "지다",
"word_root": "지",
"polarity": "-1",
```

➡ Hananum이라는 라이브러리에서 Komodo라는 라이브러리로 변경 후 감정사전 재구축

06 상관관계 분석

3) 데이터 재정의

- 날씨 데이터를 단순 날씨를 사용했다.

➡ 해당 월의 평균 날씨 – 해당 행사기간 동안의 평균날씨로 변경
(평균 기온/최고기온/최저기온/평균운량)

06 상관관계 분석

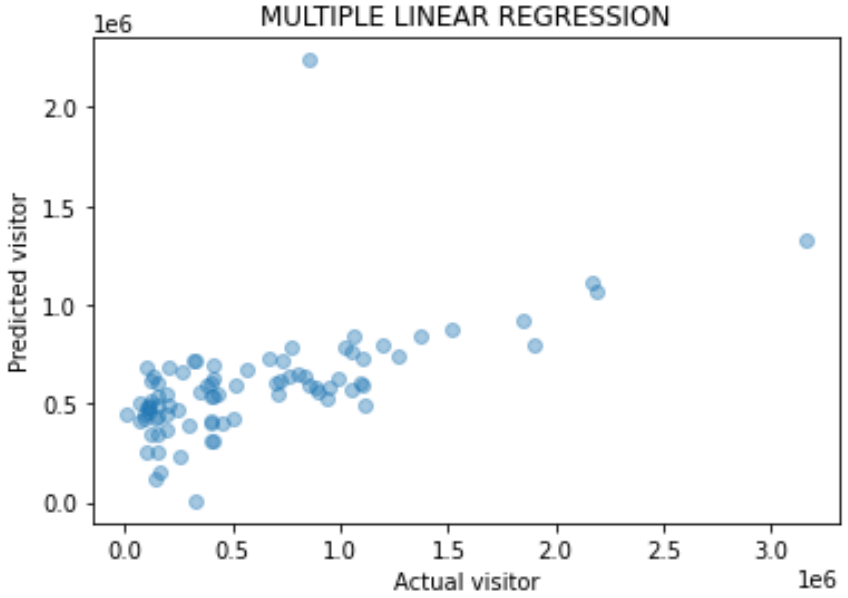
4) 최종 상관관계 분석 결과

1) 기간은 축제 인원과 뚜렷한 상관관계를 보인다.
2) KTX역의 유무도 뚜렷한 상관관계를 보인다.
3) 경제효과는 축제인원과 강한 상관관계를 보인다.
   => 추후의 인원 예측 후 경제효과를 예측
4) 평균 운량은 축제 인원과 약한 상관관계를 보인다.
5) 평균 강수량은 상관관계가 거의 없다.
6) 평균기온 절댓값과 최저기온 절대값도 약한 상관관계를 보인다.
7) Content 개수는 강한 상관관계를 보인다.
8) Positive content는 약한 상관관계를 보인다.
9) Negative content는 상관관계가 거의 없다.

07 알고리즘 구축

1) regression

- KTX 역 유무, 기간, 평균 운량, 평균기온 절댓값, 최저기온 절대값, tweet 개수, Positive content, Neutral content를 이용해 regression 분석을 시작




```
print("훈련세트점수: {}".format(mlr.score(x_train, y_train)))
print("테스트세트점수: {}".format(mlr.score(x_test, y_test)))
```

```
훈련세트점수: 0.3126467522337151
테스트세트점수: 0.3575721496357528
```

훈련 세트와 테스트 세트 둘 다 R-square 값이 작다.
➡ 과소 적합

2) 과소적합 해결을 위한 라쏘

- 모델에서 제외되는 특성을 생기게 해 적합문제 해결
- alpha 값 개선으로 모델의 복잡도를 증가시켜 성능 향상

Alpha = 1일 때

훈련세트점수: 0.354930098743094
테스트세트점수: 0.11784949158791802
사용한 특성의 수: 9

Alpha = 0.1일 때

훈련세트점수: 0.3549301007160057
테스트세트점수: 0.11784147050134797
사용한 특성의 수: 9

Alpha = 0.0001일 때

훈련세트점수: 0.3549301007359515 5
테스트세트점수: 0.1178405794349544
사용한 특성의 수: 9

➡ **특성 수가 적다.**

3) BASS 모형

- 특성 수의 부족으로 인한 문제를 해결하기 위해 수요량을 추정하는 확산모델을 사용
- 수요량의 패턴을 곡선으로 추정하는 모형이다.

$$Y_{ij} = m\frac{(p+q)^2}{p}\frac{e^{-(p+q)j}}{(1+\frac{q}{p}e^{-(p+q)j})^2} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim iid\,(0,\sigma^2)$$

- 각각의 M(최대누적 값), p(혁신 계수), q(모방계수)를 계산하기 위해 이전 구매 및 수요량을 기반으로 NLS를 이용해 값을 계산한다.
- BASS 모형은 축제별로 이전 수요량을 입력해 값을 추정한다.

3)Hybrid 모형

- 이전의 regression 값과 bass 추정치를 weigh를 이용해 최종 결과값을 예측한다.
- Hybrid 모형을 이용해 값을 추정할 때 RMSE 값이 가장 weight를 선택한다.

➡ 그 결과 bass가 0.6, regression이 0.4일 때 가장 작은 RMSE 값을 가진다.

$$\hat{y} = w\hat{y}_{LSS} + (1-w)\hat{y}_{Bass}$$

```
from sklearn.metrics import explained_variance_score, mean_squared_error, mean_absol

print('r2_score: {}'.format(r2_score(y_test, final_predict)))
```

```
r2_score: 0.711571449581172
```

08 Web page 구현

1) select page

- 파이썬 장고를 활용해 축제와 날씨를 선택할 수 있는 웹페이지를 만든다.
- 입력된 날씨 정보에 따라 날씨 데이터를 크롤링을 이용해 받아온다.

지역축제 방문객 예측

지역축제이름 무주반딧불축제

시작날짜 2020-06-01

종료날짜 2020-06-06

저장

08 Web page 구현

2) result page

- 기본정보

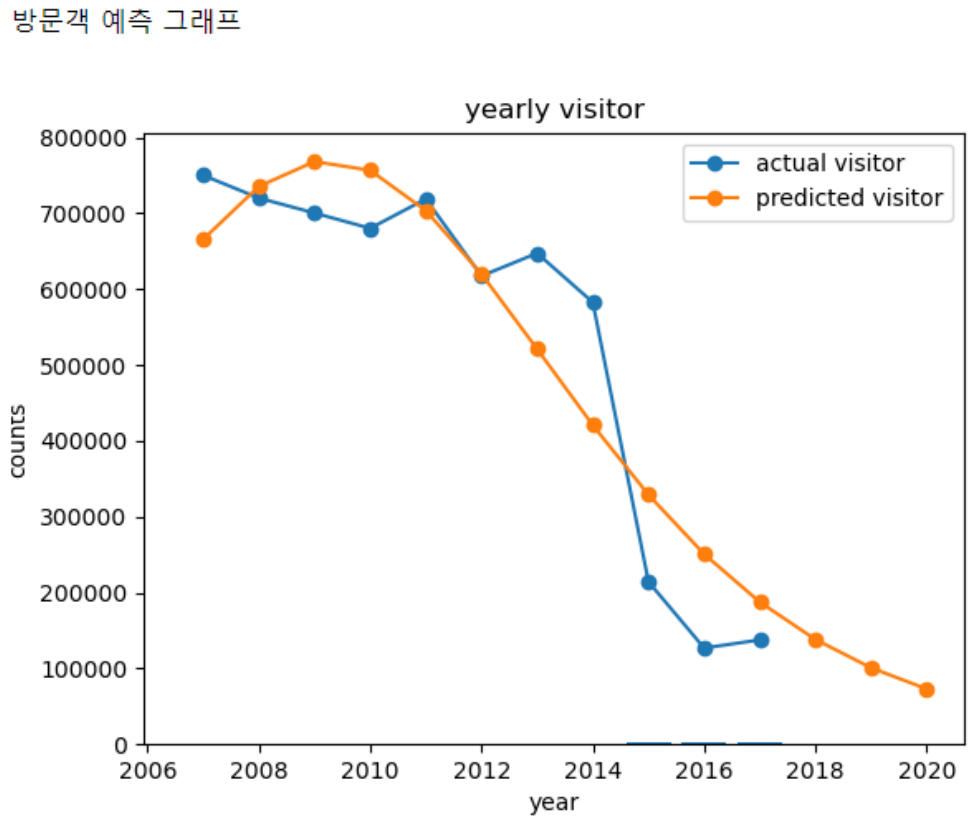
축제명: 무주반딧불축제
시작: 2020년 6월 1일
종료: 2020년 6월 6일
예측 축제인원: 77714.35228889462

날씨 정보
평균기온: 22.26
최저기온: 18.08
최고기온: 28.0

08 Web page 구현

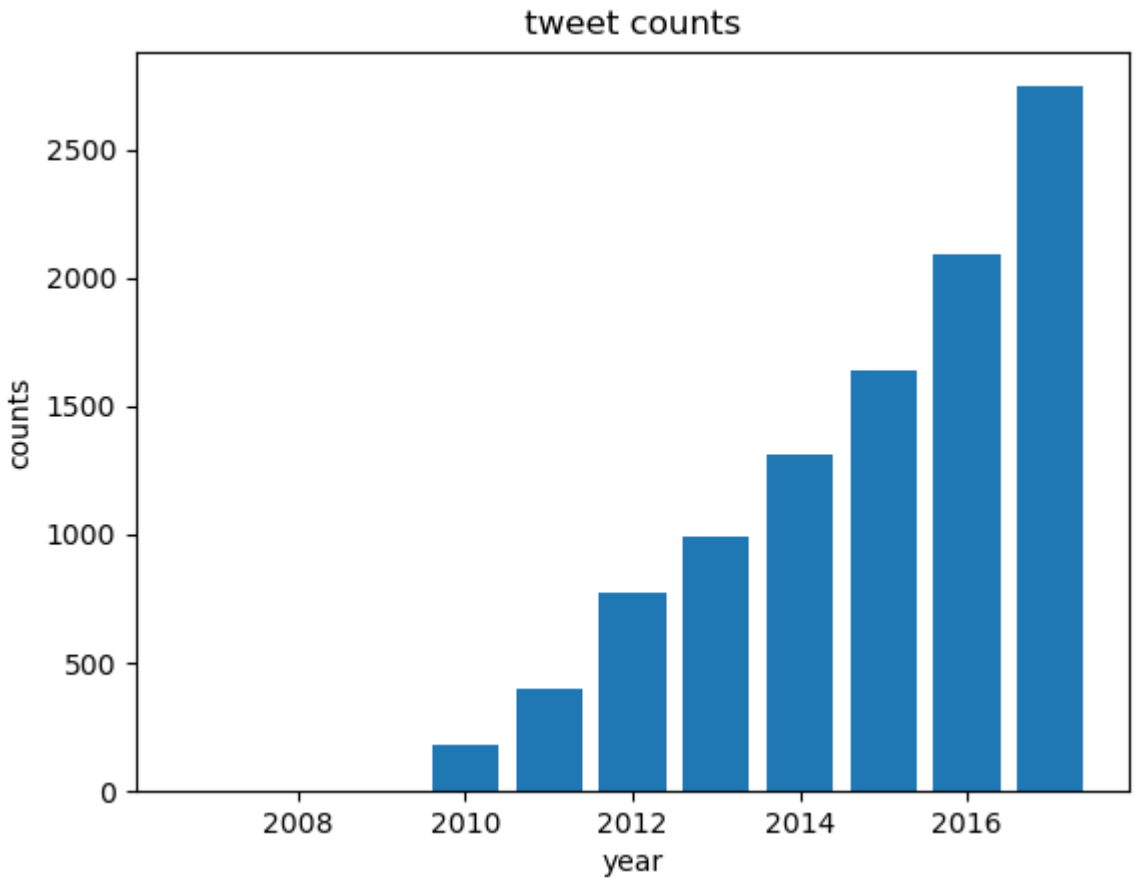
2) result page

- 방문객 예측 그래프



2) result page

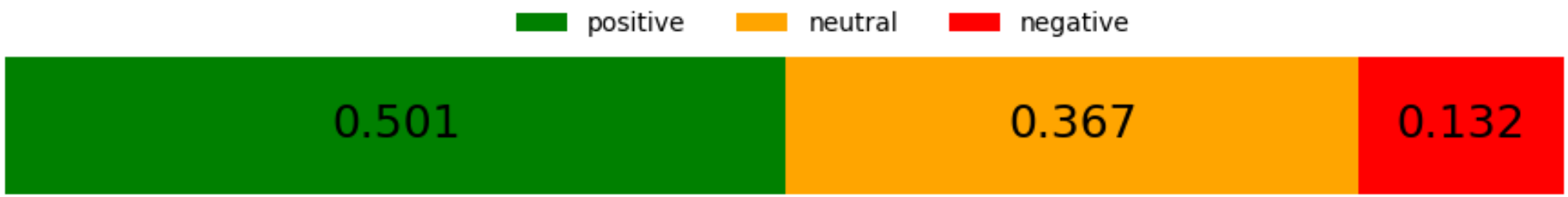
- 년도 별 트윗 개수 그래프



08 Web page 구현

2) result page

- 감정분석 비율 그래프



09 분석 및 한계점_데이터 수집 및 가공

1) 데이터의 개수
- 관광진흥처에서 제공해주는 축제 데이터의 경우, 년도별로 **결과가 좋았던 축제 데이터**만을 제공해 **년도별**로 일정한 **데이터가 부족**했다.

2) SNS 데이터
- 축제에 대한 온전한 후기 및 평가만 있는 것이 아니라 **홍보성 글**이 많았다.
- 트위터의 특성상 **리트윗**되는 글이 많다.
- **관련 없는 기사**에 해쉬태그를 해당 축제를 추가한 경우가 많다.

09 분석 및 한계점 _감정분석

1) 감정사전
- 감정사전이 단어별로 구성 되어있지 않고 “가당치 않다”라는 식으로 **주어부와 술어부가 함께 구성되어** 매칭해 감정분석 값을 지정하기 어렵다.

09 분석 및 한계점 _알고리즘 작성 및 웹페이지

1) 과소적합 문제 해결
- 특성이 많이 부족해 라쏘를 이용해도 많이 해결하지 못했다.

2) 다양한 분석 알고리즘 사용
- 단순히 regression과 bass 모형을 사용했지만 더 다양한 알고리즘을 사용해 분석하지 못했다.

3) 웹페이지 작성
- UI를 개선을 하지 못했다.
- 날씨 API를 사용하지 못해 result page를 로드하는데 시간이 오래 걸린다.

10 의의

1) 지역축제에 대한 SNS 감정분석과 관련된 선행연구가 많이 없었지만
지역축제 관련 감정분석을 진행해 결과를 얻어보았다.

2) 이전까지의 지역축제 예측의 특성으로는 감정분석 결과가 사용되지 않았지만
감정분석을 특성으로 사용해 예측을 진행해보았다.

3) 지역축제 방문객 예측을 **웹페이지**로 구현해 몇몇의 입력 값만 설정하면
예측 축제인원 값을 얻을 수 있고 그 외의 정보를 받을 수 있다.

감사합니다.