KSB 프레임 워크를 이용한 웹 크롤링 및 분석

팀장 김태환
팀원 김수민
팀원 남궁권

목차

Ⅰ. 연구내용

Ⅰ. 연구내용

1. 연구 배경 및 필요성

● 부정 이슈에 대한 신속한 대응의 필요성

◇대한항공의 '우왕좌왕 3일' vs 코오롱의 '발 빠른 9시간'

조현아 전 부사장이 기내 서비스를 매뉴얼대로 하지 않았다는 이유로 고함을 지르며 사무장을 내리게 한 사건은 이달 5일 오후 2시 50분(뉴욕 현지 시각으로는 0시 50분)에 벌어졌다. 대한항공이 이에 대한 입장을 처음 표명한 것은 그로부터 3일이 지난 8일 밤이었다.

부정이슈의 빠른 확산

⬇

신속한 대응 필요

- 코오롱 스포츠의 경우 부정 이슈에 대해서 빠르게 대응하여 이미지 실추를 최소화 시켰다.
- 그에 비해 대한항공과 포스코에너지는 늦은 대응으로 인해 이미지가 많이 실추되었다.

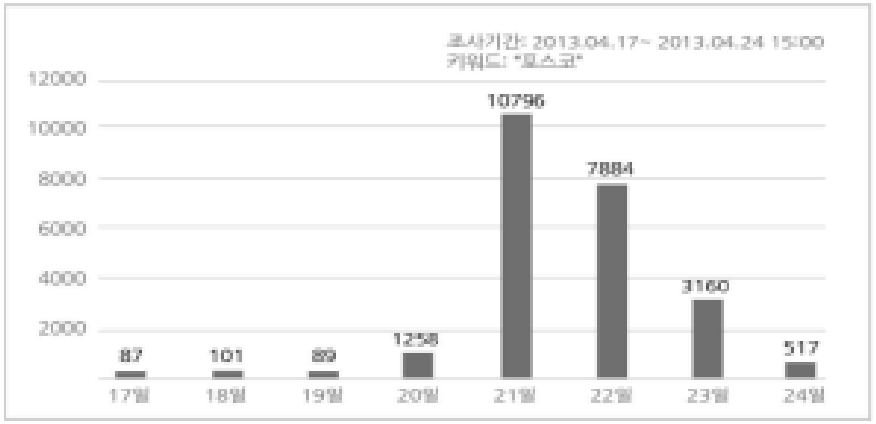


그림 6) 포스코에너지 트위터 버즈량 추이

포스코에너지 라면 상무 사건

Ⅰ. 연구내용

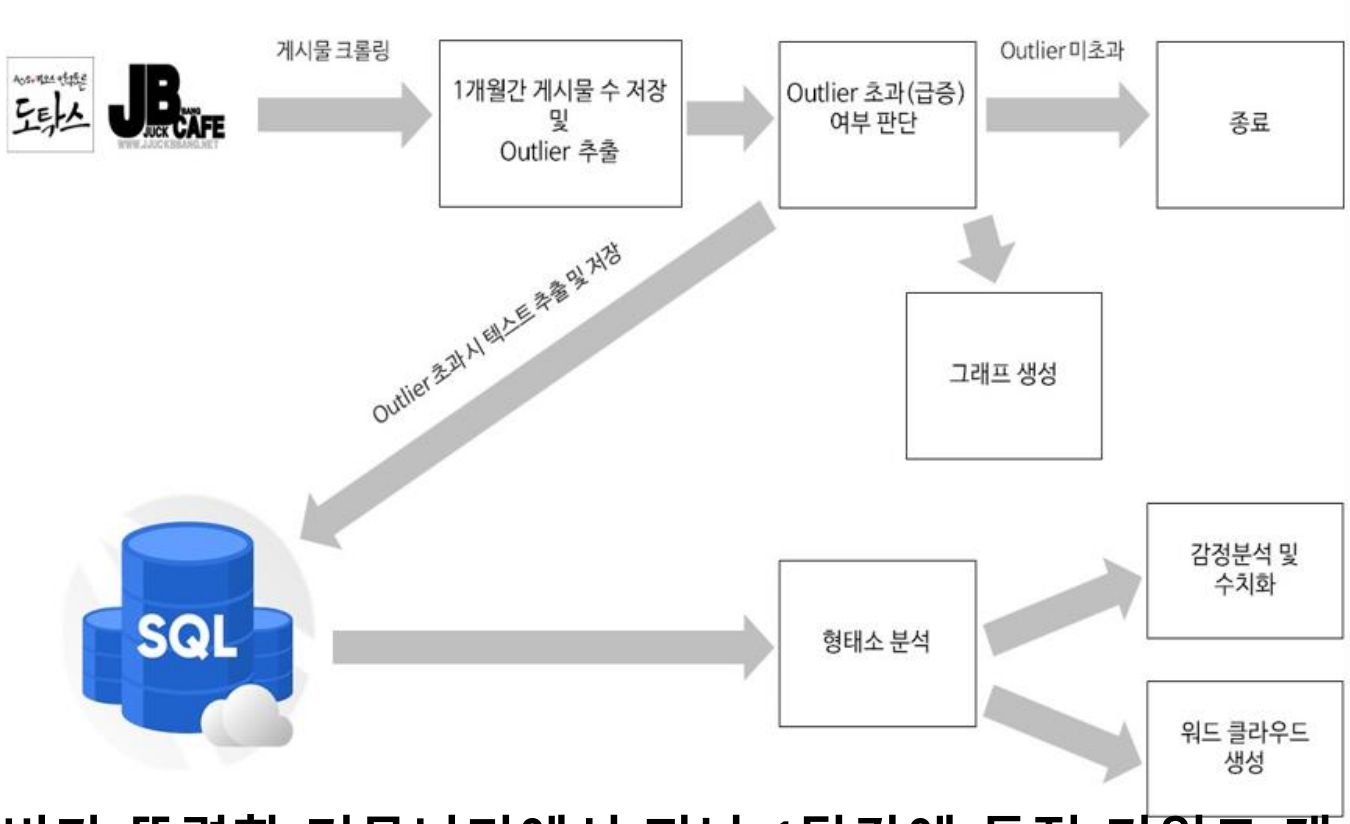
1. 연구 배경 및 필요성

따라서 기업들이 부정 이슈 확산에 대해

간편하게 확인할 수 있는 프로젝트 진행

Ⅰ. 연구내용

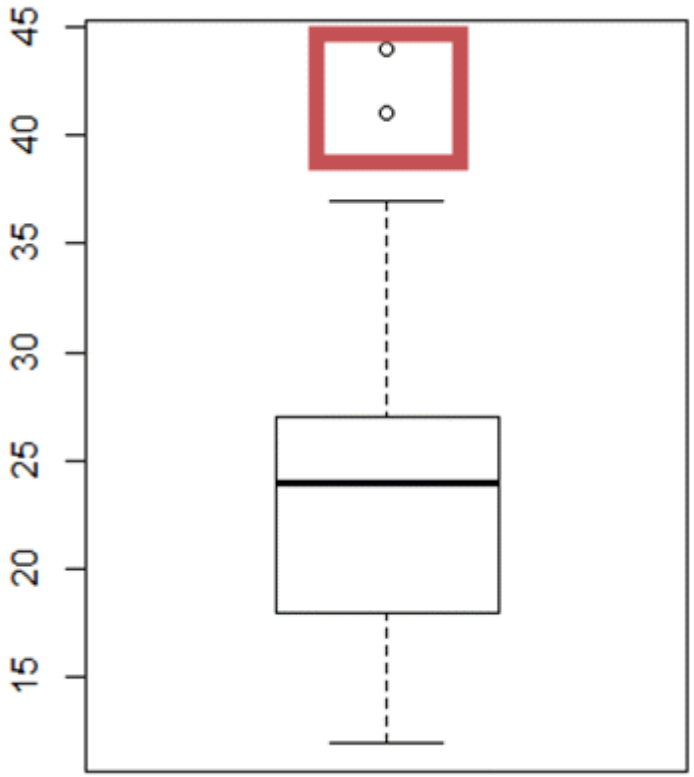
2. 연구 과정



1. 남녀 성비가 뚜렷한 커뮤니티에서 지난 1달간에 특정 키워드 게시물 언급량을 크롤링하여 일평균 얼마나 언급되는지 파악한다.
2. 일평균보다 많이 언급된다면 텍스트를 추출 및 저장을 한 뒤 형태소 분석 및 감정분석을 한다.
3. 그래프와 워드클라우드를 통해 시각화해준다.

Ⅰ. 연구내용

2. 연구 과정



Boxplot의 예시

Outlier는

Boxplot을 통해 정한다.

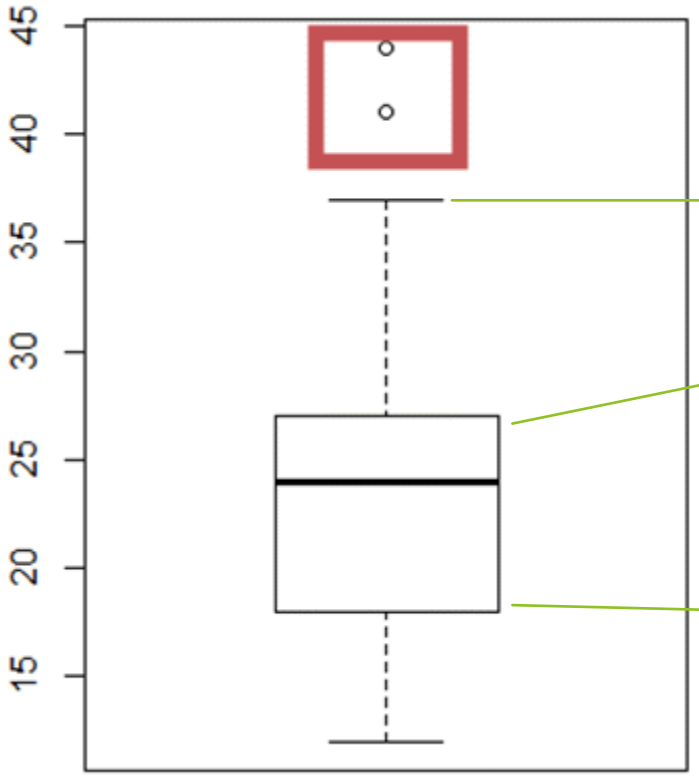
상자 그림	값	설명
상자 아래 세로 점선	아랫수염	하위 0~25%
상자 밑면	1사분위수(Q1)	하위 25% 위치 값
상자 내 굵은 선	2사분위수(Q2)	하위 50% 위치 값(중앙값)
상자 윗면	3사분위수(Q3)	하위 75% 위치 값
상자 위 세로 점선	윗수염	하위 75~100%
상자 밖 가로선	극단치 경계	Q1, Q3 밖 1.5 IQR 내 최대값
상자 밖 점 표식	극단치	Q1, Q3 밖 1.5 IQR을 벗어난 값

Boxplot의 사분위수

Ⅰ. 연구내용

2. 연구 과정

극단치 경계보다 많은 양의 게시물이 언급되면
급증한다고 판단하고 게시물의 텍스트들을 크롤링한다.



Outlier(극단치 경계) = Q3 + 1.5*IQR

Q3

IQR : Q3 - Q1의 값

Q1

Ⅰ. 연구내용

2. 연구 과정

● 추출한 텍스트를 통해 감정 분석(Doc2Vec)을 한다.

```
#doc2vec parameters
cores = multiprocessing.cpu_count()

vector_size = 300
window_size = 15
word_min_count = 2
sampling_threshold = 1e-5
negative_size = 5
train_epoch = 100
dm = 1
worker_count = cores
```

1.Doc2vec train data parameters 설정한다.

적중률이 약 70% 되는 모델을 생성할 수 있다.

```
# 사전 구축
doc_vectorizer = doc2vec.Doc2Vec(size=300, alpha=0.025, min_alpha=0.025, seed=1234)
doc_vectorizer.build_vocab(tagged_train_docs)

# Train document vectors!
for epoch in range(10):
    doc_vectorizer.train(tagged_train_docs, total_examples=doc_vectorizer.corpus_count, epochs=doc_vectorizer.iter)
    doc_vectorizer.alpha -= 0.002  # decrease the learning rate
    doc_vectorizer.min_alpha = doc_vectorizer.alpha  # fix the learning rate, no decay

#To save
doc_vectorizer.save('model/doc2vec.model') # doc2vec.model(사전)을 생성한다.
```

2. Training data로 사전을 구축한다.

```
# train data로 만든 사전을 load한다.
doc_vectorizer = Doc2Vec.load('model/doc2vec.model')

# 분류를 위한 피쳐 생성
train_x = [doc_vectorizer.infer_vector(doc.words) for doc in tagged_train_docs]
train_y = [doc.tags[0] for doc in tagged_train_docs]
test_x = [doc_vectorizer.infer_vector(doc.words) for doc in tagged_test_docs]
test_y = [doc.tags[0] for doc in tagged_test_docs]

classifier = LogisticRegression(random_state=1234)
classifier.fit(train_x, train_y)

# 테스트 socre 확인 적중률
print( classifier.score(test_x, test_y) )
# 0.66712

# save the model to disk
filename = 'model/finalized_model.sav'
pickle.dump(classifier, open(filename, 'wb'))
```

3. Training data로 만든 model을 test data로 test한다.

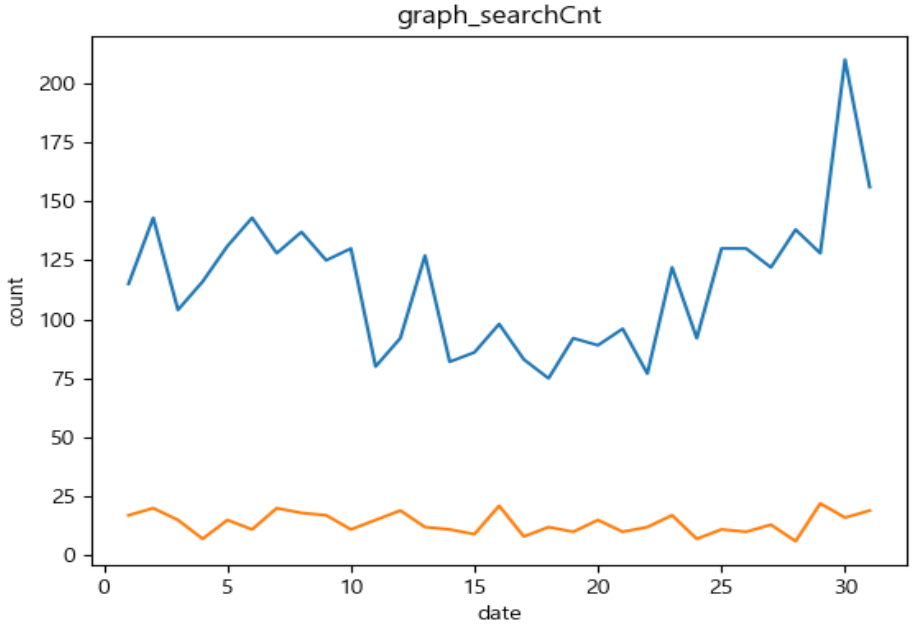
Ⅰ. 연구내용

2. 연구 과정

● 앞서 만든 모델을 이용하여 추출한 텍스트를 감정분석한다.
Ex) 키워드 : 아이돌

1. 워드클라우드

2. 그래프 (파랑 : 女, 주황 : 男)
일별 각 커뮤니티에서의 언급량



Ⅰ. 연구내용

3. 연구 성과

● 연구 과정을 크게 세가지 모듈로 나눌 수 있다.

1. 데이터 확보, 추출

2. 감정 분석을 이용한 데이터 분석

3. 감정 분석이 된 결과물 추출
   - 그래프, 워드클라우드

1 2 3번의 Python 모듈을 Dockerize 시켜서 사용한다.

Dockerize 시키면 더 많은 커뮤니티에서 데이터를 추출할 수 있으며 더 정교한 감정 분석이 가능해진다.

Ⅱ. 기대성과 및 활용계획

Ⅱ. 기대성과 및 활용계획

1. 기대성과

1. 기업의 부정 이슈 확산 방지	2. 엔터테인먼트 이미지 관리	3. 커뮤니티별 성향이나 선호도 파악
		
커뮤니티, SNS 모니터링을 통해 부정 이슈가 확산 되는 것을 초기에 막을 수 있고 앞서 말 했듯이 초기 대응을 제대로 하지 못해 기업 이미지를 실추시키는 일을 최소화할 수 있다.	잘못된 정보가 확산되는 것을 막을 수 있기 때문에 연예인의 이미지 관리에도 도움이 될 수 있다.	남녀 성비가 뚜렷한 커뮤니티를 대상으로 크롤링하므로 성비에 맞춘 마켓팅 전략을 내세울 수 있다.

Ⅱ. 기대성과 및 활용계획

2. 활용계획

스타트업 및 중소기업	일반인	소상공인
브랜드 및 신규제품 고객 반응 모니터링 프로그램으로 사용	관심분야 키워드를 모니터링 키워드로 등록하여 분야 트렌드 분석 프로그램으로 사용	지역 커뮤니티 카페에 매장 언급량을 모니터링하여 부정 이슈 대응