Project-01

Sentiment Analysis on Movie Reviews(with Kaggle)

*개인 Project / 데이터 수집 및 모델링 / 2016. 11 ~2016. 12

Abstract

- 영화 리뷰 감성 분석
- kaggle Sentiment Analysis on Movie Reviews
- project (모델 성능 최적화)
- kaggle 점수 확인

프로젝트 개요

Why ? (프로젝트를 하게 된 계기)
- 텍스트 분석을 통해서 개인의 성향을 판단이 가능할 수 있다는 가정하에 프로젝트 진행

How? (데이터 수집 및 분석 방법론)
- 데이터 수집
  kaggle 영화 리뷰 감성분석 데이터 활용.
  (The Rotten Tomatoes movie review dataset is a corpus of movie reviews used for sentiment analysis, originally collected by Pang and Lee[1])

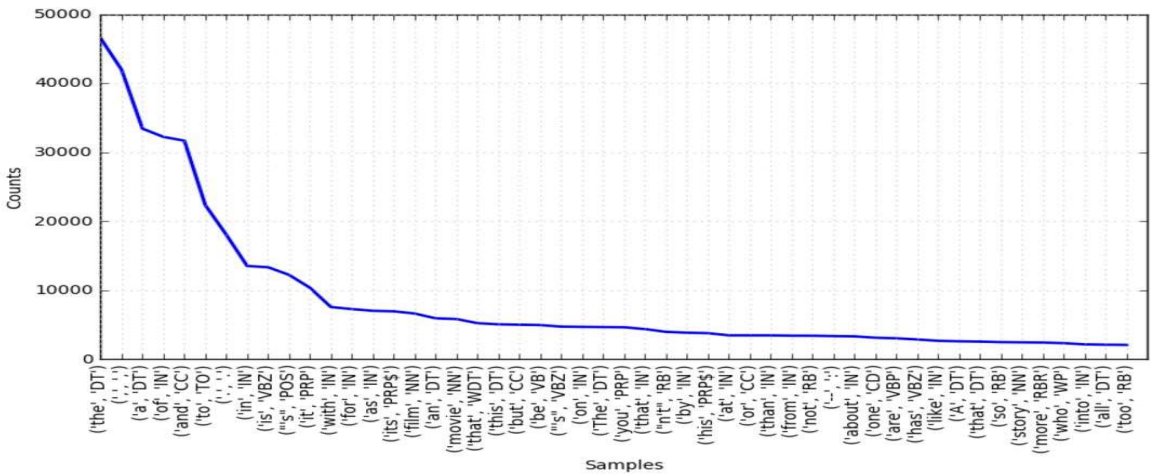
- 평가방법 (Y) 예측
  0 - negative
  1 - somewhat negative
  2 - neutral
  3 - somewhat positive
  4 - positive

방법론?

1. 데이터 전처리

```
read_file = pd.read_csv('./resource/train.tsv', sep='₩t')
test_file = pd.read_csv('./resource/test.tsv', sep='₩t')
```

<그림1. 데이터 불러오기.>



Project-01

Sentiment Analysis on Movie Reviews(with Kaggle)

*개인 Project / 데이터 수집 및 모델링 / 2016. 11 ~2016. 12

프로젝트

2. Model 선택

```
clf_2 = Pipeline([
        ('tv' , TfidfVectorizer(analyzer = 'word', tokenizer=tokenizer)),
        (('clf', SVC(kernel='linear')))
    ])
```

```
clf_2.fit(x_train, y_train)
```

```
Pipeline(steps=[('tv', TfidfVectorizer(analyzer='word', binary=False, decode_error=u'strict',
        dtype=<type 'numpy.int64'>, encoding=u'utf-8', input=u'content',
        lowercase=True, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), norm=u'l2', preprocessor=None, smooth_idf=True,
  ...,
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False))])
```

```
print(classification_report(y_test, clf_2.predict(x_test)))
print('*' * 100)
print(accuracy_score(y_test, clf_2.predict(x_test)))
```

```
             precision    recall  f1-score   support

          0       0.53      0.15      0.23      1339
          1       0.50      0.37      0.43      5236
          2       0.66      0.86      0.75     15273
          3       0.55      0.44      0.49      6367
          4       0.59      0.18      0.28      1785

avg / total       0.60      0.62      0.59     30000

****************************************************************************************************
0.615266666667
```

<그림2. 서포터 벡터 모델 선택.>

- kaggle 결과

416	↓35	Ryan Zhao	0.60701	2	Tue, 09 Dec 2014 02:26:28	
-		**Lee Sung Guk**	**0.60680**	-	**Mon, 19 Dec 2016 14:35:47**	Post-Deadline
Post-Deadline Entry If you would have submitted this entry during the competition, you would have been around here on the leaderboard.						
417	↓35	charalson	0.60674	1	Sat, 31 Jan 2015 20:22:42	

<그림3. 예측 데이터 평가.>

Project-01

Sentiment Analysis on Movie Reviews(with Kaggle)

*개인 Project / 데이터 수집 및 모델링 / 2016. 11 ~2016. 12

프로젝트 결론

분석?

1. 정확도 Test : 0.62, 실제 데이터 예측: 0.60
2. 형태소 분석 시 원형 어원 사용.
3. 서포트 벡터 머신, 랜덤 포레스트, 나이브 베이즈 모형 사용

향후 보완점?

1. 형태소 분석 시 어근 추출, 품사 태깅 방법 사용
2. 텍스트 분류 및 감정 사전 이용방법 숙지
3. 확률론적 언어 모형 및 gensim, word2vec 사용 모델 최적화
4. 영어가 아닌 한글 모델 구현 필요

인사이트?

1. 텍스트 분석 시 욕설을 stop word(해당하는 단어 필터링)을 구성 후 필터링 기능 구현
2. 키워드 분석
3. 댓글, 소셜 글을 통해서 행동 분류