한글 영화 리뷰 분석

학습 내용

- simple pre-processing: word tokenization, word count, data analysis
- stopwords 처리
- data labeling, train/test set 준비
- 성능 평가 기준 마련
- classification using machine learning algorithms

입력 data

45290 크리스토퍼 놀란 에게 우리는 놀란 다 10
45290 인셉션 정말 흥미진진하게 봤었고 크리스토퍼 놀란 감독님 신작 인터스텔라도 이번주 일요일에 보러갑니다 완전 기대중 10
45290 놀란이면 무조건 봐야 된다 왜냐하면 모든 작품을 다 히트 쳤으니깐 10
45290 나는 감탄할 준비가 되어있다 10
45290 얘들아 오늘나오는거지 밤에 ㅋㅋ 오늘 보러가야겟다 10
45290 이제 죽어도 여한이 없다 10
45290 기대감에 잠도 안온다 10
45290 나랑 같이 봐 줄까 ㅎ 10

영화 리뷰 분석 실습

- data load
- Counter 클래스를 이용한 단어의 빈도수 계산
- 상위 빈도 n개 단어 확인
- 빈도 n 이상인 단어만 선택

```
[('영화', 40000),
 ('정말', 18266),
 ('진짜', 14207),
 ('너무', 13397),
 ('이', 7893),
 ('영화를', 7006),
 ('그냥', 6844),
 ('더', 6560),
 ('최고의', 6394),
 ('보고', 5899),
 ('좀', 5725),
 ('수', 5671),
 ('영화가', 5588),
 ('최고', 5347),
 ('영화는', 5308),
 ('잘', 5039),
 ('꼭', 4944),
 ('ㅋㅋ', 4865),
 ('본', 4655),
 ('다', 4609)]

n = 1, the number of words : 400649
n = 2, the number of words : 100035
n = 3, the number of words : 61982
n = 4, the number of words : 46182
n = 5, the number of words : 37222
n = 6, the number of words : 31295
n = 7, the number of words : 27206
```

- train data 준비
  - 클래스
    - 1~3 점은 -1: 부정적 리뷰
    - 9,10점은 1: 긍정적 리뷰
    - 나머지는 사용하지 않거나 중립으로 처리
- 특정 단어에 해당하는 인덱스 알아보기
- LogisticRegression으로 학습
- 일부 리뷰를 골라 LogisticRegression으로 긍정/부정 리뷰예측

```
text: ['재미없다', '이상', '10자']
predicted class prob: (negative= 0.754, positive= 0.246
predicted class = negative
actual class = negative
-------------------------------------------------
text: ['정말정말', '대단합니다']
predicted class prob: (negative= 0.048, positive= 0.952
predicted class = positive
actual class = positive
-------------------------------------------------
text: ['관람객진짜', '인생영화다', '꼭', '봐야하는', '영화']
predicted class prob: (negative= 0.004, positive= 0.996
predicted class = positive
actual class = positive
-------------------------------------------------
text: ['빵점주고싶다']
predicted class prob: (negative= 0.660, positive= 0.340
predicted class = negative
actual class = negative
-------------------------------------------------
text: ['나도', '최악에', '한표', 'ㅋㅋ']

predicted class prob: (negative= 0.921, positive= 0.079
predicted class = negative
actual class = negative
```

- 긍정리뷰로 예측하는데 영향을 미친 단어들 상위 10 출력
- 부정리뷰로 예측하는데 영향을 미친 단어들 상위 10 출력
- N-fold cross-validation 으로 성능 측정 (accuracy, recall, precision, F1)
- Train, Test set을 나누어 성능 측정 (accuracy, recall, precision, F1)
- 2) 다른 머신러닝 알고리즘 (SVM, Decision Tree, Random Forest, Knn) 적용하여 성능 비교

알이즈웰 (4.316)
최고입니다 (3.503)
관람객재밌어요 (3.227)
기대됩니다 (3.199)
10점준다 (3.118)
좋았어요 (3.089)
웰 (2.987)
재미있었습니다 (2.978)
꿀잼 (2.934)
인생영화 (2.934)

0점은 (-4.261)
최악 (-4.248)
최악의 (-4.187)
0점이 (-3.892)
1점준다 (-3.749)
돈아깝다 (-3.736)
쓰레기영화 (-3.608)
1점대 (-3.566)
노잼 (-3.559)
최악의영화 (-3.474)

추가 테스트

1) konpy로 형태소 분석

2) stopwords 적용(너무 자주 등장하는 단어 등)

3) Tf-Idf 적용