

# 인공지능 과제 2

## 영화 리뷰 긍정/부정 분류하기

### 1. 코드 설명

```
4 from konlpy.tag import Mecab
5 import sys
6 import os
7 import string
8 import math
9
10 pos_cnt = 0
11 neg_cnt = 0
12 pos_word_cnt = 0
13 neg_word_cnt = 0
14 pos_words = {}
15 neg_words = {}
```

주어진 한글 데이터의 형태소 분석을 위해 Konlpy의 Mecab 클래스를 사용하였습니다. `mecab.morphs()` 함수를 이용하여 텍스트를 형태소 단위로 잘라주었으며, 각각의 형태소를 출현 빈도와 함께 딕셔너리에 저장하였습니다. 코드 상단의 전역변수는 다음을 의미합니다.

`pos_cnt` : 긍정 문장의 개수.

`neg_cnt` : 부정 문장의 개수.

`pos_word_cnt` : 긍정 문장에 속한 단어의 전체 개수.

`neg_word_cnt` : 부정 문장에 속한 단어의 전체 개수.

`pos_words` : 긍정문장에 속한 단어들을 key로, 각각 단어의 출현 빈도를 value로 저장한 딕셔너리.

`neg_words` : 부정문장에 속한 단어들을 key로, 각각 단어의 출현 빈도를 value로 저장한 딕셔너리.

```
268 def main():
269
270     # read_train_file("./ratings_data/ratings_train.txt")
271     # save_train_result("./ratings_data/trained_data_save.txt")
272
273     load_train_result("./ratings_data/trained_data_save.txt")
274
275     #test_valid_file("./ratings_data/ratings_valid.txt")
276     classify("./ratings_data/ratings_test.txt", "./ratings_data/ratings_result.txt")
```

다음은 메인함수입니다. 학습 과정과 학습결과를 이용한 classification과정을 분리 하였습니다. read\_train\_file(학습 데이터 경로) 과 save\_train\_result(학습 결과를 저장할 텍스트파일 경로)를 먼저 호출하여야 하며, 이후 save\_train\_result()에서 저장된 텍스트 파일을 load\_train\_result() 에서 읽어 메모리에 올린 다음 classify(test파일 경로) 혹은 test\_valid\_file() 을 호출하여 분류작업을 하게 됩니다. 학습결과를 저장한 텍스트 파일의 포맷은 다음과 같습니다(trained\_data\_save.txt).

```
1 trained_data_save.txt Buffers
1 89966 90034
2 1615426 1688731
3 친구 531 들 12621 끼리 72 의 24403 우정 152 을 19726 생각 3916 하 24137 게 1
4 갈 5262 은 19304 동족 4 이 42435 라는 1551 걸 1380 떠나 188 서 2212 인간 609
```

첫번째 줄에 pos\_cnt, neg\_cnt, 그 다음 줄에 pos\_word\_cnt, neg\_word\_cnt, 3번째 줄에 pos\_word 딕셔너리, 4번째 줄에 neg\_word 딕셔너리를 각각 저장합니다. 딕셔너리는 item 사이는 '\t', key 와 value 사이는 공백문자로 구분하였습니다.

Classification 과정을 수행하기 위해서는 위의 텍스트 파일을 로드한 후 , classify() 함수를 호출하여야 합니다. Classify 함수에서는 각각의 line text 에 대해서 형태소 분석을 한 후, word 들의 긍정 출현 빈도, 부정 출현 빈도 를 이용하여 해당 comment 가 긍정일 확률 과 부정일 확률을 각각 계산하게 됩니다.

```
#각각의 word에 대해 긍정과 부정의 확률을 계산합니다.
log_pos_prob = math.log(pos_cnt / (pos_cnt + neg_cnt) )
log_neg_prob = math.log(neg_cnt / (pos_cnt + neg_cnt) )
for word in analyzedLine:
    log_pos_prob += caculate_prob(1, word)
    log_neg_prob += caculate_prob(0, word)

# print("POS:", log_pos_prob)
# print("NEG:", log_neg_prob)

# 결과 파일에 태그를 달아 기록합니다.
if log_pos_prob >= log_neg_prob:
    fw.write(lineSplited[0]+' \t'+lineSplited[1]+' \t'+str(1)+'\n')
else:
    fw.write(lineSplited[0]+' \t'+lineSplited[1]+' \t'+str(0)+'\n')
```

```

18 # 각 단어의 긍정빈도 혹은 부정빈도(확률)를 로그를 취한 값으로 리턴합니다.
19 def caculate_prob(bool, word):
20     global pos_word_cnt, neg_word_cnt, pos_words, neg_words
21
22     if bool == 1:
23         total = pos_word_cnt
24         dic = pos_words
25
26     else:
27         total = neg_word_cnt
28         dic = neg_words
29
30     # test case 에서 처음 등장한 단어의 확률을 계산할때 0을 곱하게 되는 것을 방지하기 위해(혹은 로그0 계산)
31     # 적당히 작은 상수 k를 분자와 분모에 더해줍니다.
32     k = 0.5
33
34     if dic.get(word) == None:
35         v = 0
36     else:
37         v = dic[word]
38
39
40     return ( math.log(k + float(v)) - math.log(2.0 * k + float(total)) )
41
42
43 def dic_input(dic, word):
44     if dic.get(word) == None:
45         dic[word] = 1
46
47     else:
48         dic[word] += 1

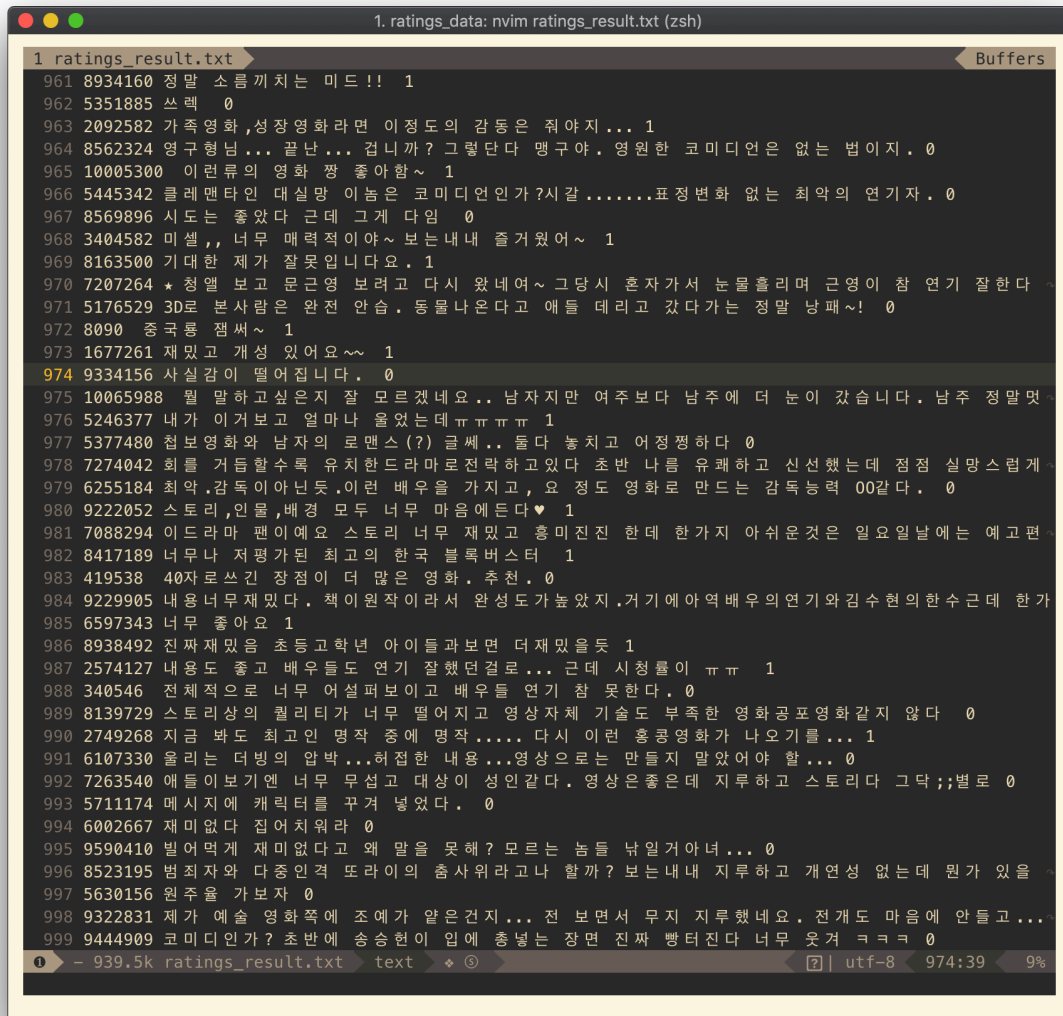
```

하나의 comment 가 긍정일 확률은 (긍정 comment의 개수/ 전체 comment의 개수) \* (각 word가 긍정 comment 에서 나온 빈도/긍정 comment에 속한 word 전체 개수)\*(...) 와 같이 계산 합니다. 부정일 확률도 유사합니다. 작은 확률을 계속 곱하다 보면 너무 작은 숫자가 되어 underflow 를 일으킬 수 있는데 이를 방지하기 위해 log를 취해서 계산 하도록 합니다. 함수 caculate\_prob 는 각 단어의 긍정 빈도 혹은 부정 빈도(확률)을 로그를 취한 값으로 리턴합니다. 또한, test case에서 처음 등장한 단어일 경우, 0의 확률을 곱하는 것(혹은 0에 로그를 취하는 것)을 방지하기 위해 적당히 작은 상수 k 를 분모와 분자에 더하여 계산합니다.

위와 같이 각각의 comment가 긍정일 확률과 부정일 확률을 계산한 후, 둘을 비교하여 긍정이 더 클 경우 1, 부정이 더 클 경우 0을 result file 에 기록하였습니다.

## 2. 실행 결과

ratings\_result.txt 파일이 생성되었으며, 그 일부는 다음과 같습니다.



```
1 ratings_result.txt Buffers
961 8934160 정말 소름끼치는 미드!! 1
962 5351885 쓰레 0
963 2092582 가족영화,성장영화라면 이 정도의 감동은 줘야지... 1
964 8562324 영구형님... 끝난... 겁니까? 그렇단다 맹구야. 영원한 코미디언은 없는 법이지. 0
965 10005300 이런류의 영화 짱 좋아함~ 1
966 5445342 클레멘타인 대실망 이놈은 코미디언인가?시갈.....표정변화 없는 최악의 연기자. 0
967 8569896 시도는 좋았다 근데 그게 다임 0
968 3404582 미셀,, 너무 매력적이야~ 보는내내 즐거웠어~ 1
969 8163500 기대한 제가 잘못입니다요. 1
970 7207264 ★ 청өл 보고 문근영 보려고 다시 왔네여~ 그당시 혼자가서 눈물흘리며 근영이 참 연기 잘한다
971 5176529 3D로 본사람은 완전 안습. 동물나온다고 애들 데리고 갔다가는 정말 낭패~! 0
972 8090 중국풍 짬썩~ 1
973 1677261 재밌고 개성 있어요~~ 1
974 9334156 사실감이 떨어집니다. 0
975 10065988 뭘 말하고싶은지 잘 모르겠네요.. 남자지만 여주보다 남주에 더 눈이 갔습니다. 남주 정말멋
976 5246377 내가 이거보고 얼마나 올랐는데 π π π π 1
977 5377480 첩보영화와 남자의 로맨스(?) 클레.. 둘다 놓치고 어정쩡하다 0
978 7274042 회를 거듭할수록 유치한드라마로전락하고있다 초반 나를 유쾌하고 신선했는데 점점 실망스럽게
979 6255184 최악,감독이아닌듯.이런 배우를 가지고, 요 정도 영화로 만드는 감독능력 00갈다. 0
980 9222052 스토리,인물,배경 모두 너무 마음에든다♥ 1
981 7088294 이드라마 팬이에요 스토리 너무 재밌고 흥미진진 한데 한가지 아쉬운것은 일요일날에는 예고편
982 8417189 너무나 저평가된 최고의 한국 블록버스터 1
983 419538 40자로쓰긴 장점이 더 많은 영화. 추천. 0
984 9229905 내용너무재밌다. 책임원작이라서 완성도가높았지,거기에아역배우의연기와김수현의한수근데 한가
985 6597343 너무 좋아요 1
986 8938492 진짜재밌음 초등고학년 아이들과보면 더재밌을듯 1
987 2574127 내용도 좋고 배우들도 연기 잘했던걸로... 근데 시청률이 π π 1
988 340546 전체적으로 너무 어설피보이고 배우들 연기 참 못한다. 0
989 8139729 스토리상의 퀄리티가 너무 떨어지고 영상자체 기술도 부족한 영화공포영화같지 않다 0
990 2749268 지금 봐도 최고인 명작 중에 명작..... 다시 이런 홍콩영화가 나오기를... 1
991 6107330 올리는 더빙의 압박...허접한 내용...영상으로는 만들지 말았어야 할... 0
992 7263540 애들이보기엔 너무 무섭고 대상이 성인같다. 영상은좋은데 지루하고 스토리다 그닥;;별로 0
993 5711174 메시지에 캐릭터를 꾸겨 넣었다. 0
994 6002667 재미없다 집어치워라 0
995 9590410 빌어먹게 재미없다고 왜 말을 못해? 모르는 놈들 닦일거야녀... 0
996 8523195 범죄자와 다중인격 또라이의 춤사위라고나 할까? 보는내내 지루하고 개연성 없는데 뭔가 있을
997 5630156 원주율 가보자 0
998 9322831 제가 예술 영화쪽에 조예가 알은건지... 전 보면서 무지 지루했네요. 전개도 마음에 안들고...
999 9444909 코미디인가? 초반에 송승헌이 입에 총넣는 장면 진짜 뻔뻔하다 너무 웃겨 ㅋㅋ 0
- 939.5k ratings_result.txt text 974:39 9%
```

정확도를 확인하기 위해 주어진 ratings\_valid.txt 를 test\_valid\_file() 함수를 이용하여 classify 해본 후, 기존의tag와 비교하여 보았습니다. 해당 함수에서는 맞고 틀린 comment의 개수를 true positive, true negative, false positive, false negative 로 출력하도록 하였습니다.

```
1. ~/Desktop/AIAssignment/assignment2 (cat)

어디서 볼 수 있나요? 상영작이라고 뜨는데 볼 수 있는 영화관이 없네요. 1
분석할 필요 없다. 그냥 누구에게나 어떻게든 볼 영화 1
저 사람들은 저렇게 살아 하고 연인의 일상을 보여준다 0
안보면 진심 후회함...!! 1
과연 이런 사랑이 정말 현실적인걸까. 퓨처룸에서 관계를 맺던 신디의 표정을 보고서 영화를
꺼버렸다. 이런 사랑은 하고 싶지 않아서 0
옥소리 프로필 사진에 1점 남기고 갑니다 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 완전 대박이다 진짜 아
우 짜증나! 0
'난징, 난징' 만큼의 영화는 아니다. 0
유승호랑 결혼하고 싶엉~ 1
솔직한 평점 하지만 심감독님 언젠간 대박 터트릴 듯 0
이연걸의 소림사 시리즈 이후 오랜만의 작품이었다. 0
그냥 만화책이나 한번 더 보는게 정신건강에 이로움 ㅎㅎ 0
내가 오늘 과학의 날 행사가 있어서 이 영화를 봤는데 재미가 좀 그렇다 0
카메라웍만으로 무협스토리를 멜로로 만든 그게 진짜반전이다.. 왕가위의 힘!! 1
언더더스톰급일거 같다.. 1
고단한 인생... 0
마지막 장면 빌딩이 국회쪽으로 넘어가는거 같던데 감독의 의도가 있었겠죠? 1
모녀끼리 손 잡고 가서 손수건 좀 적실영화 0
이문식 존나 웃기네 ㅋㅋ 1
스틸러 내가 제일 좋아하는데 뭐라 표현을 할수가없네 1
이렇게 리얼하다니 전혀 연출스럽지 않다 1
알ㅋㅋㅋㅋㅋㅋ 0
누가 이거 이상하다고 하나 긴장감최곤데 1
자꾸 정이 간다.. 정이 1
다 떠나서 스토리 자체가 이상함. 마틴아저씨는 정말 멋지지만,.. 0
재밌기는 한데, 이벤트호라이즌과 비교하니 좀 우울하네요 .. 0
아 조금만 더 스틸 있었다면 스토리는 좋은데;; 0
이런걸 영화라고본 내눈을 뽀아버리고 싶을정도의 영화 한마디로 대박 초딩영화! 0
정말내가 본 영화 중에 가장 이상한 영화 0
말이필요없다.... 1
세이지는 세상을 너무 잘 이해하는것이 아니라 잘못 이해하고 있는것이다 0
귀신의 공포보다는 입시의 공포...그리고 공포를 뛰어넘지 못한자들..... 1
같은시나리오에 다른 배우들이 연기했음 어땠을까? 1
그들은 즐겁게 춤을 추는데 난 왜 흥겹지 않을까? 0
true positive: 4186
true negative: 4196
false positive: 782
false negative: 836

~/Desktop/AIAssignment/assignment2 master*
> □
```

10000개의 데이터 가운데, 8382 개가 정답이었으므로, 80 퍼센트 이상의 정확도를 보이는 것을 확인할 수 있었습니다.