Web Crawling & Machine Learning

Naver movie review analysis (positive, negative)

Song

Part.1 Web Crawling

Web crawling

task.

- Enter Naver movie review using Chrome Drive & Selenium
- Crawling the content of reviews using BeautifulSoup
- Automate using For loop
- Create Data Frame using Pandas

Enter Naver Movie Review

Chrome Drive & Selenium

Enter Naver Movie Review

Chrome Drive & Selenium

Import the necessary modules

```
1  from bs4 import BeautifulSoup
2  from selenium import webdriver
3  import time
4  import sys
5  import re
6  import math
7  import numpy
8  import pandas as pd
9  import xlwt
10 import random
11 import os
12 import urllib.request
13 import urllib
```

Enter Naver Movie Review

Chrome Drive & Selenium

```
15 name = input(      “Movie title”      )
16 reviewA = input(      “Number of reviews”      )
17
18 chrome_path = '/Users/inhyeoksong/Desktop/codding/wc/chromedriver'
19 driver = webdriver.Chrome(chrome_path)
20 url = "https://movie.naver.com/"
21 driver.maximize_window()
22 driver.get(url)
23 import time
24 time.sleep(2)
```

Enter the title of the movie and the number of reviews required.
(Movie title : 말할 수 없는 비밀,
Number of reviews : 105)

Enter Naver Movie Review

Chrome Drive & Selenium

```
27  SearchBar = driver.find_element_by_xpath(
28      '//*[@id="ipt_tx_srch"]')
29  SearchBar.click()
30  time.sleep(2)
31  SearchBar.send_keys('%s'%name)
32  time.sleep(2)
33  SearchR =driver.find_element_by_xpath(
34      '//*[@id="jAutoMV"]/ul/li[1]/a/div')
35
36  SearchR.click()
37  time.sleep(2)
38
39  time.sleep(2)
40
41  reviwePluse = driver.find_element_by_xpath(
42      '//*[@id="movieEndTabMenu"]/li[6]/a')
43  reviwePluse.click()
44  time.sleep(2)
45
46
```

Set the ‘xpath’ value to the ‘element’ of the page.
Click the search bar and specify Search term ‘%s send_key’

Enter Naver Movie Review

Chrome Drive & Selenium

Crawl the content of reviews

BeautifulSoup & user-defined function

Crawl the content of reviews

BeautifulSoup & user-defined function

“Search for tags containing the necessary information”

Crawl the content of reviews

BeautifulSoup & user-defined function

If there is a star rating

```
no = 0
def single_page (no):

    try:
        html = driver.page_source
        soup = BeautifulSoup(html,'html.parser')
        no+=1

        INreview = soup.find('div',class_='obj_section noline center_obj')


        star = INreview.find('div',class_="star_score").text.strip()

        print("별점",star)
        print("review no : ", no)

        title = INreview.find('strong',class_='h_lst_tx').text.strip()
        date = INreview.find('span',class_='wrt_date').text.strip()
        content = INreview.find('div',class_="user_tx_area").text.strip()
        reviewpoint = INreview.find('em',id='goodReviewCount').text.strip()

        print("제목 : ", title)
        print("리뷰추천 수 : ", reviewpoint)
        print("날짜 : ", date)
        print("내용 : \n", content)

        print("="*120)
        reno.append(no)
        star_li.append(star)
        title_li.append(title)
        date_li.append(date)
        content_li.append(content)
        reviewpoint_li.append(reviewpoint)

        return(no)

```

Crawl the content of reviews

BeautifulSoup & user-defined function

If there are no stars rating

```
37      except:
38          html = driver.page_source
39          soup = BeautifulSoup(html,'html.parser')
40          INreview = soup.find('div',class_='obj_section noline center_obj')
41
42          print("별점이 없습니다.")
43          print("review no : ", no)
44
45          title = INreview.find('strong',class_='h_lst_tx').text.strip()
46          date = INreview.find('span',class_='wrt_date').text.strip()
47          content = INreview.find('div',class_="user_tx_area").text.strip()
48          reviewpoint = INreview.find('em',id='goodReviewCount').text.strip()
49
50          print("제목 : ", title)
51          print("리뷰추천 수 : ", reviewpoint)
52          print("날짜 : ", date)
53          print("내용 : \n", content)
54
55          print("="*120)
56          reno.append(no)
57          star_li.append("별점이 없습니다.")
58          title_li.append(title)
59          date_li.append(date)
60          content_li.append(content)
61          reviewpoint_li.append(reviewpoint)
62
63          return(no)
```

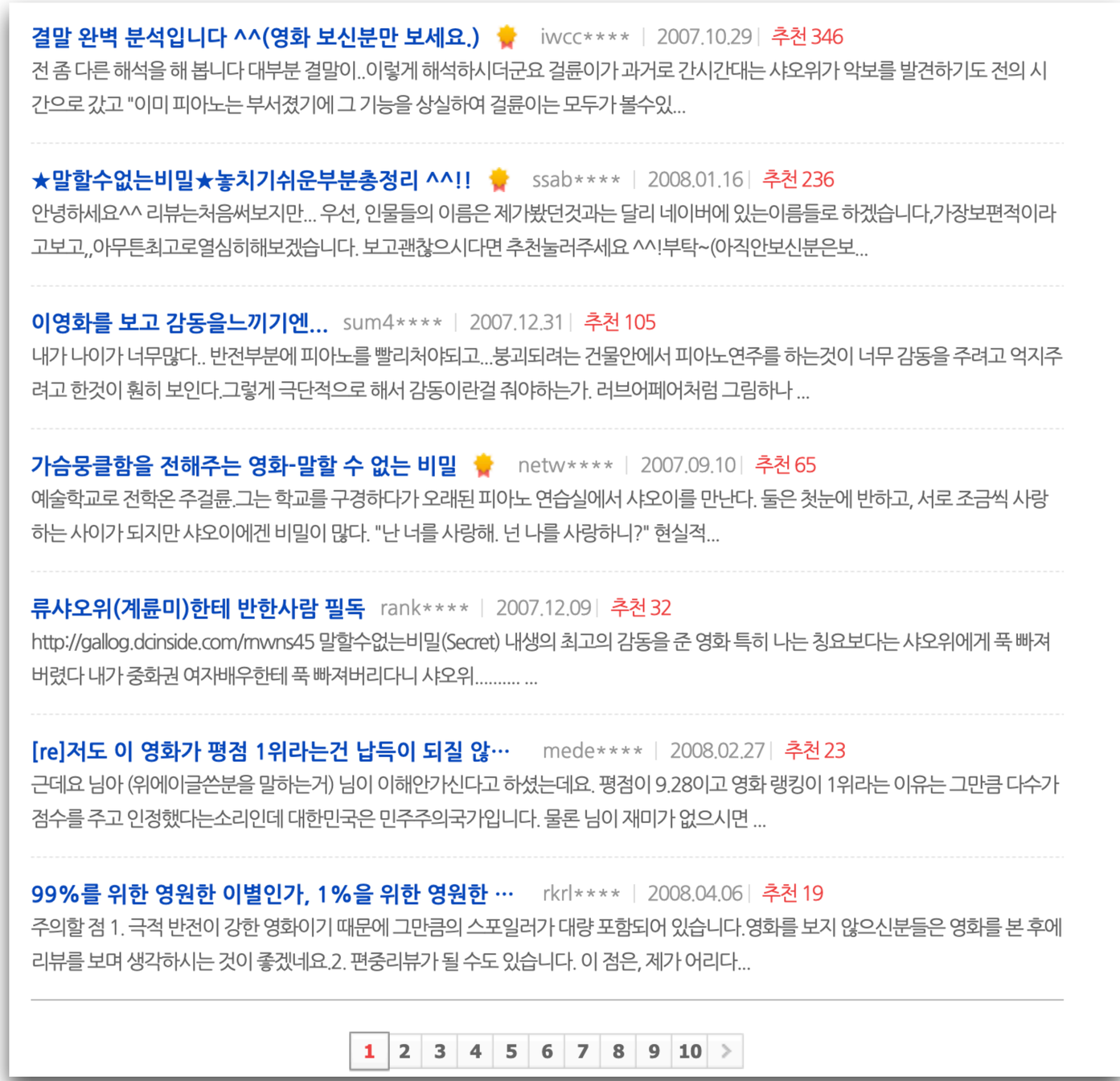
Automation

For loop

Automation

For loop

- Find the rules
- 10 posts per page
- Analysis of page buttons after 10



Automation

For loop

```
ten = 0

# 총 필요한 리뷰의 수 1의자리
end = int(reviewA)%10
# 페이지 넘김횟수
page = math.ceil(int(reviewA)/10)

# 마지막페이지 리뷰
endingArt = int(reviewA)%10
for x in range(1,int(page)+1):
    driver.find_element_by_xpath('//*[@id="pagerTagAnchor%s"]'%x).click()
    time.sleep(2)
    print(ten)

    #마지막 페이지 loop으로 빠지는 조건
    if ten/10 == int(page)-1:
        for e in range(1,end+1):
            driver.find_element_by_xpath('//*[@id="reviewTab"]/div/div/ul/li[%s]/a'%e).click()
            time.sleep(3)
            no = single_page(no)

            driver.back()

    else:
        for i in range(1,11):
            driver.find_element_by_xpath('//*[@id="reviewTab"]/div/div/ul/li[%s]/a'%i).click()
            time.sleep(3)
            no = single_page(no)

            driver.back()
            ten+=1
```

Data Frame Creation

Pandas

Data Frame Creation

Pandas

```
df = pd.DataFrame()

df["리뷰번호"] = pd.Series(reno)
df["리뷰 추천수"] = pd.Series(reviewpoint_li)
df["별점"] = pd.Series(star_li)
df["제목"] = pd.Series(title_li)
df["날짜"] = pd.Series(date_li)
df["내용"] = pd.Series(content_li)

f_name_xls = "/Users/inhyeoksong/Desktop/codding/wc/navermovies.csv"

df.to_csv(f_name_xls, encoding = 'utf-8-sig',index=False)
```

Part.2 Machine Learning

Machine Learning

task.

- Utilizing a model that learned positive and negative words in Korean
- Import CSV file & check data
- Data preprocessing
- Tokenization & Predictive Execution (for loop)
- Data Frame Creation
- visualization

Import CSV file and check data

Import CSV file and check data

```
!pwd
!ls
```

```
/content
best_model.h5  ratings_test.txt  ratings_train.txt  sample_data
```

```
[ ]
from google.colab import files
uploaded = files.upload()
```

파일 선택 선택된 파일 없음 Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving navermovies.csv to navermovies.csv
```

Import CSV file and check data

```
import pandas as pd
new_data = pd.read_csv('navermovies.csv', encoding='utf-8')
new_data.info()
new_data [5:]
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105 entries, 0 to 104
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   리뷰번호    105 non-null    int64 
 1   리뷰 추천수  105 non-null    int64 
 2   별점      105 non-null    object
 3   제목      105 non-null    object
 4   날짜      105 non-null    object
 5   내용      105 non-null    object
dtypes: int64(2), object(4)
memory usage: 5.0+ KB
```

	리뷰번호	리뷰 추천수	별점	제목	날짜	내용
5	6	105	3	이영화를 보고 감동을느끼기엔...	2007.12.31	내가 나이가 너무많다..\n반전부분에 피아노를 빨리쳐야되고...붕괴되려는 건물안에서...
6	7	65	9	가슴뭉클함을 전해주는 영화-말할 수 없는 비밀	2007.09.10	예술학교로 전학온 주걸륜.그는 학교를 구경하다가 오래된 피아노 연습실에서 샤오이를 ...
7	8	32	10	류샤오위(계륜미)한테 반한사람 필독	2007.12.09	http://gallog.dcinside.com/rnwns45\n \n말할수없는비밀...
8	9	23	10	[re]저도 이 영화가 평점 1위라는건 납득이 되질 않습니다.	2008.02.27	근데요 님아 (위에이글쓴분을 말하는거)\n \n님이 이해안가신다고 하셨는데요.\n ...
9	10	19	9	99%를 위한 영원한 이별인가, 1%을 위한 영원한 사랑인가.	2008.04.06	주의할 점 \n1. 극적 반전이 강한 영화이기 때문에 그만큼의 스포일러가 대량 포함...
...	...	...	...	...	...	...
100	101	2	7	관객수는 얼마 없는데 이렇게 많은 리뷰가 있다니..	2008.01.15	결론은 대부분 불법 다운로드 해서 봤다는 말..\n문제다 문제 정말.. 버젓이 현재...
101	102	2	7	짬뽕 짱깨 영화.....	2008.01.07	왠 홍콩 영화가 이렇게 평이 좋은지 의문으로 봤던 영화...\n \n왜 이렇게 별점...
102	103	2	6	볼만했던 하이틴 로맨스	2007.11.11	중후반까지는 너무 진부하고 내용없는 로맨스의 진행이며 중후반부터는 판타지틱한 재미가...
103	104	2	6	이 영화 평점은 7점 정도가 적당할듯 (수정)	2007.12.21	동감 지금 만나러 갑니다 시월애 등 \n시공간을 넘나드는 사랑을 소재로한 영화들을 ...
104	105	2	6	곱게 빚은 하이틴 로맨스.....	2010.04.08	곱게 빚은 하이틴 로맨스..... \n\n \n상륜(주걸륜)은 오랜 전통을 가진 ...

100 rows × 6 columns

Data preprocessing

Data preprocessing

```python
import numpy as np

#데이터 전처리 수행
new_data.drop_duplicates(subset = ['내용'], inplace=True) # document 열에서 중복인 내용이 있다면 중복 제거
new_data['내용'] = new_data['내용'].str.replace("[^ㄱ-ㅎㅏ-ㅣ가-힣 ]","") # 정규 표현식 수행
new_data['내용'] = new_data['내용'].str.replace('^ +', "") # 공백은 empty 값으로 변경
new_data['내용'].replace('', np.nan, inplace=True) # 공백은 Null 값으로 변경
new_data = new_data.dropna(how='any') # Null 값 제거
print('전처리 후 테스트용 샘플의 개수 :',len(new_data))
new_data2 = new_data['내용']
new_data2
```

```
전처리 후 테스트용 샘플의 개수 : 105
0       이 영화 정말 스토리나 연출 캐스팅연기남주는 감독각본까지 했으므로 좀 봐줘서ㅋ음악영...
1       말할 수 없는 비밀우리나라에서 개봉 됐을때 부터 숱하게 올라온 리뷰와 칭찬글들그 유...
2       점이 적당하다평점이 좋길래봤는데평점에 속은 느낌이다나름 중국영화 특유의 과장된 몸짓...
3       전 좀 다른 해석을 해 봅니다대부분 결말이이렇게 해석하시더군요걸륜이가 과거로 간시간...
4       안녕하세요 리뷰는처음써보지만 우선 인물들의 이름은 제가봤던것과는 달리 네이버에 있는...
                              ...
100     결론은대부분 불법 다운로드 해서 봤다는 말문제다 문제 정말 버젓이 현재 개봉하고 있...
101     왠 홍콩 영화가 이렇게 평이 좋은지 의문으로 봤던 영화왜 이렇게 별점이 높은지원일본...
102     중후반까지는 너무 진부하고 내용없는 로맨스의 진행이며 중후반부터는 판타지틱한 재미가...
103     동감 지금 만나러 갑니다 시월애 등 시공간을 넘나드는 사랑을 소재로한 영화들을 많이...
104     곱게 빚은 하이틴 로맨스 상륜주걸륜은 오랜 전통을 가진 예술학교에 갓 들어온 전학생...
Name: 내용, Length: 105, dtype: object
```

Tokenization & Predictive Execution
(for loop)

Tokenization & Predictive Execution(for loop)

Probably a positive review.
Probably a positive review.

```
[ ]  new_data3 = new_data2.sample(n=105)
     len(new_data3)

     105
```

```
[ ]  #새로운 문장 예측해주는 함수
     def sentiment_predict(new_sentence):
       new_sentence = re.sub(r'[^ㄱ-ㅎㅏ-ㅣ가-힣 ]','', new_sentence)
       new_sentence = okt.morphs(new_sentence, stem=True) # 토큰화
       new_sentence = [word for word in new_sentence if not word in stopwords] # 불용어 제거
       encoded = tokenizer.texts_to_sequences([new_sentence]) # 정수 인코딩
       pad_new = pad_sequences(encoded, maxlen = max_len) # 패딩
       score = float(loaded_model.predict(pad_new)) # 예측
       if(score > 0.5):
         print("{:.2f}% 확률로 '긍정' 리뷰입니다.\n".format(score * 100))
         score_list.append(score * 100)
         pos_neg_list.append(1)
       else:
         print("{:.2f}% 확률로 '부정' 리뷰입니다.\n".format((1 - score) * 100))
         score_list.append((1-score) * 100)
         pos_neg_list.append(0)
```

Tokenization & Predictive Execution(for loop)

```
no = 1
score_list = []
pos_neg_list = []
for i in new_data3:
    print(no)
    sentiment_predict(i)
    no += 1
```

```
18
97.97% 확률로 '긍정' 리뷰입니다.

19
76.29% 확률로 '긍정' 리뷰입니다.

20
99.12% 확률로 '긍정' 리뷰입니다.

21
68.55% 확률로 '긍정' 리뷰입니다.

22
90.49% 확률로 '부정' 리뷰입니다.

23
97.94% 확률로 '긍정' 리뷰입니다.

24
99.63% 확률로 '긍정' 리뷰입니다.

25
82.66% 확률로 '부정' 리뷰입니다.

26
82.34% 확률로 '부정' 리뷰입니다.
```

```
import pandas as pd
new_data = pd.read_csv('navermovies.csv', encoding='utf-8')
new_data.info()
new_data [18:27]
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105 entries, 0 to 104
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   리뷰번호    105 non-null    int64
 1   리뷰 추천수  105 non-null    int64
 2   별점      105 non-null    object
 3   제목      105 non-null    object
 4   날짜      105 non-null    object
 5   내용      105 non-null    object
dtypes: int64(2), object(4)
memory usage: 5.0+ KB
```

	리뷰번호	리뷰 추천수	별점	제목	날짜	내용
18	19	9	10	모든 궁금증을 말끔히...	2007.11.26	안녕하세요?\n \n이 영화에서 생각하고 계시는 오류가 개별시간으로 보십니다 . 또...
19	20	8	9	말할 수 없는 비밀, 그 비현실적인 사랑 이야기	2007.12.09	말할 수 없는 비밀\n \n소감문 쓰기를 시작하기 전, 직접 다운받아 한 번 더 ...
20	21	8	10	제발그만좀하자..딱두가지만.	2008.02.05	1.알바가지고 ㅈㄹ하는것들\n \n내가이말만몇번되풀이하는지모르겠다. 이영화, 개봉관...
21	22	8	5	겉치레는 화려하나...	2008.04.22	첫사랑에 대한 추억처럼 아련한 느낌의 영상미는 헐리웃 유명 멜로 영화 못지 않은 세...
22	23	8	별점이 없습니다.	비밀스러운 음악이 아름다운영화 "말할수없는 비밀 (不能說的秘密, Secret, ...	2020.07.20	오늘 다시볼 영화는 "말할수없는 비밀"입니다.일본영화는 아닌 대만 영화인데요. 일본...
23	24	7	10	말할 수 없는 비밀	2008.01.10	이 글은 스포일성 글이므로 영화를 보신분에 한해서만 보시길 바랍니다. \n영화 안보...
24	25	7	10	아니, 진짜 웃긴 사람들 많네.	2008.02.07	사람들이 각자 개인의 관점으로 영화를 보고 난 뒤에\n \n자기 자신이 느낀, 남의...
25	26	7	7	평점이란 대중성에 근거한 것일 뿐...	2007.10.31	나는 미술학도다. 그것도 순수미술학도... 내가 배우는 내용과 그리는 것들은 공감대...
26	27	6	10	우리나라 사람들 옛날영화가 평점 1위여야 한다는 이상한 생각에 빠져있다	2007.10.16	꼭 옛날 영화 가 1위여야 한다는\n생각을 버려야 한다 \n언제까지 쇼생크탈출,레옹...

Compare with actual content

Data Frame Creation

Data Frame Creation

```
new_dataframe = pd.DataFrame()
new_dataframe['내용'] = new_data3
new_dataframe['score'] = score_list
new_dataframe['sentiment'] = pos_neg_list

new_dataframe
```

	내용	score	sentiment
42	난 이영화 점줬습니다그리고 이글은 영화자체에 대해서 혹평하는게 아닙니다제발좀 말입니...	91.440383	0
72	확실한결말다들 아시겠지만 피아노를 정석대로 치면 년미래로 오구 빨리치면 년전으로 갑...	76.026699	0
39	일단 오랜만에 상당히 좋은 영화가 나왔고 이런 영화를 봐서 감수성이 풍부해지는 어떠...	73.751724	0
26	꼭 옛날 영화 가 위여야 한다는생각을 버려야 한다 언제까지 쇼생크탈출레옹타이타닉등 ...	53.242242	1
96	아마 이 영화를 보신후에 이런저런 결말을 생각 하시리라 생각합니다저 같은 경우도 이...	55.307782	0
...	...	...	...
25	나는 미술학도다 그것도 순수미술학도 내가 배우는 내용과 그리는 것들은 공감대를 형성...	95.436102	1
1	말할 수 없는 비밀우리나라에서 개봉 됐을때 부터 숱하게 올라온 리뷰와 칭찬글들그 유...	97.361827	1
82	저우제륜은 천재 맞다 무려 년생이 이런 영화의 각본 감독 연출을 했다니어찌 보면 애...	99.588180	1
20	알바가지고 ㅈㄹ하는것들내가이말만몇번되풀이하는지모르겠다 이영화 개봉관이관이다 한국에 ...	58.712506	1
51	영화 보신 분이니 규칙부터 잡고 시작하겠습니다이 영화의 내용 전개에서 가장 중요한 ...	95.489150	1

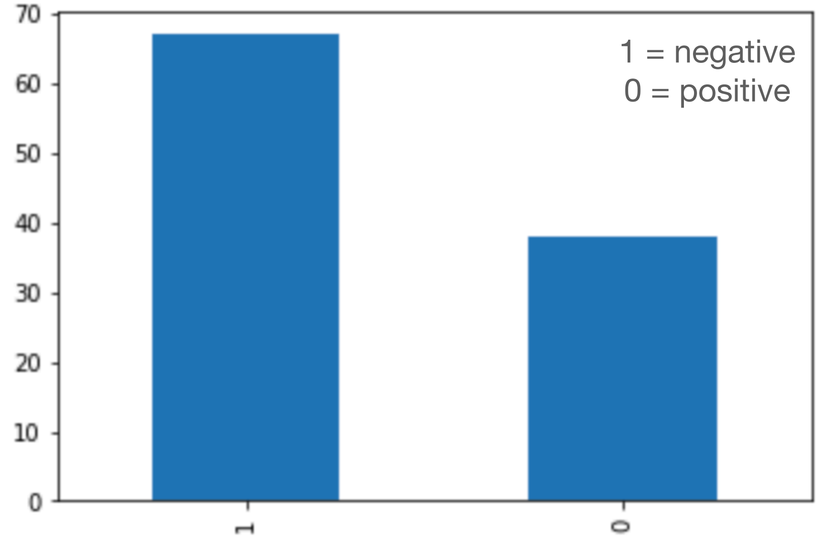
105 rows × 3 columns

visualization

visualization

```
new_dataframe['sentiment'].value_counts().plot(kind = 'bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2a948dcd50>
```



thank you