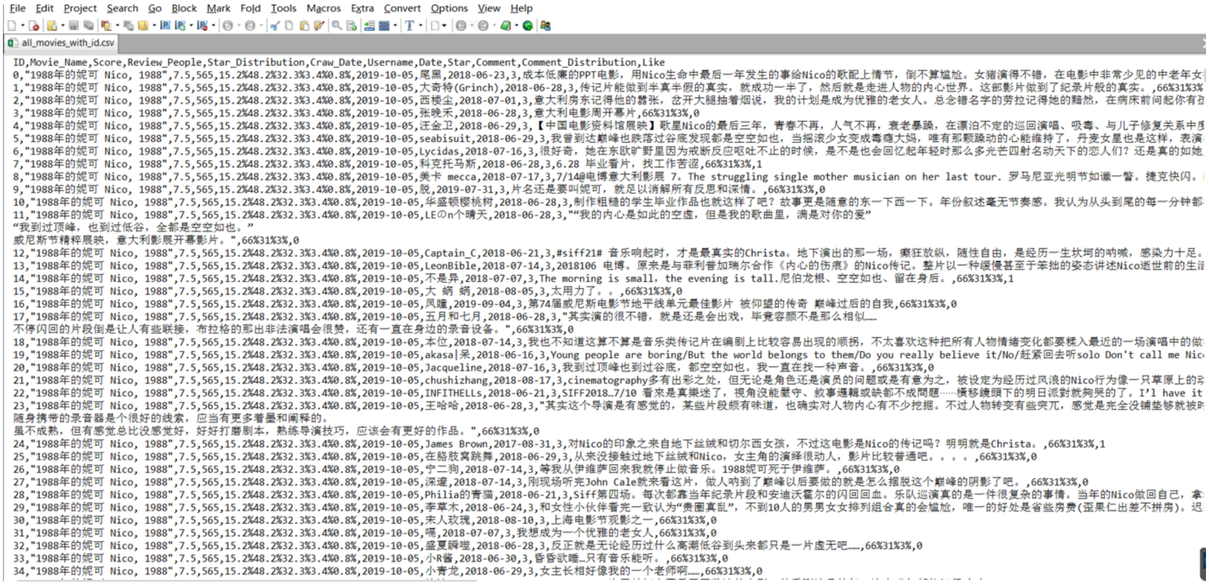
Status report

Group: DATA SMILE

- What we have done:

Crawled around 15000 movies and finally got 10269 unique movies after processing. All of the movie information and the corresponding comments have been written into one csv file. We upload this file to kaggle.com and it is now publicly available at: *https://www.kaggle.com/liujt14/dou-ban-movie-short-comments-10377movies*. Each record in the dataset has 12 attributes, including movie name, overall score, stars associated with each comment and so on.


File Edit Project Search Go Block Mark Fold Tools Macros Extra Convert Options View Help
all_movies_with_id.csv

```
ID,Movie_Name,Score,Review_People,Star_Distribution,Craw_Date,Username,Date,Star,Comment,Comment_Distribution,Like
0,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,尾黑,2018-06-23,3,成本低廉的PPT电影，用Nico生命中最后一年发生的事给Nico的歌配上情节，倒不算尴尬。女猪演得不错，在电影中非常少见的中老年女
1,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,大奇特(Grinch),2018-06-28,3,传记片能做到半真半假的真实，就成功一半了，然后就是走进人物的内心世界。这部影片做到了纪录片般的真实。,66%31%3%
2,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,西楼尘,2018-07-01,3,意大利房东记得他的嚣张，岔开大腿抽着烟说，我的计划是成为优雅的老女人。总念错名字的劳拉记得她的黯然，在病床前问起你有
3,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,张晚禾,2018-06-28,3,意大利电影周开幕片,66%31%3%,0
4,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,汪金卫,2018-06-29,3,【中国电影资料馆展映】歌星Nico的最后三年，青春不再，人气不再，衰老暴躁，在漂泊不定的巡回演唱、吸毒、与儿子修复关系中度
5,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,seabisuit,2018-06-29,3,我曾到达巅峰也跌落过谷底发现都是空空如也，当摇滚少女变成毒瘾大妈，唯有那颗躁动的心能维持了，丹麦女星也是这样，表演
6,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,Lycidas,2018-07-16,3,很好奇，她在东欧旷野里因为戒断反应呕吐不止的时候，是不是也会回忆起年轻时那么多光芒四射名动天下的恋人们？还是真的如她
7,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,科克托马斯,2018-06-28,3,6.28 毕业看片，找工作苦涩,66%31%3%,1
8,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,美卡 mecca,2018-07-17,3,7/14@电博意大利影展 7. The struggling single mother musician on her last tour. 罗马尼亚光明节如谶一瞥。捷克快闪。
9,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,脱,2019-07-31,3,片名还是要叫妮可，就足以消解所有反思和深情。,66%31%3%,0
10,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,华盛顿樱桃树,2018-06-28,3,制作粗糙的学生毕业作品也就这样了吧？故事更是随意的东一下西一下。年份叙述毫无节奏感。我认为从头到尾的每一分钟都
11,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,LEのn个晴天,2018-06-28,3,"“我的内心是如此的空虚，但是我的歌曲里，满是对你的爱”
“我到过顶峰，也到过低谷，全都是空空如也。”
威尼斯节精粹展映，意大利影展开幕影片。",66%31%3%,0
12,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,Captain_C,2018-06-21,3,#siff21# 音乐响起时，才是最真实的Christa。地下演出的那一场，癫狂放纵，随性自由，是经历一生坎坷的呐喊，感染力十足。
13,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,LeonBible,2018-07-14,3,2018106 电博。原来是与菲利普加瑞尔合作《内心的伤痕》的Nico传记。整片以一种缓慢甚至于笨拙的姿态讲述Nico逝世前的生活
14,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,不是异,2018-07-07,3,The morning is small, the evening is tall.尼伯龙根、空空如也、留在身后。,66%31%3%,1
15,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,大 螃 螃,2018-08-05,3,太用力了。。,66%31%3%,0
16,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,凤瞳,2019-09-04,3,第74届威尼斯电影节地平线单元最佳影片 被仰望的传奇 巅峰过后的自我,66%31%3%,0
17,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,五月和七月,2018-06-28,3,"其实演的很不错，就是还是会出戏，毕竟容颜不是那么相似……
不停闪回的片段倒是让人有些联接，布拉格的那出非法演唱会很赞，还有一直在身边的录音设备。",66%31%3%,0
18,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,本位,2018-07-14,3,我也不知道这算不算是音乐类传记片在编剧上比较容易出现的顺拐，不太喜欢这种把所有人物情绪变化都要糅入最近的一场演唱中的做
19,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,akasa|呆,2018-06-16,3,Young people are boring/But the world belongs to them/Do you really believe it/No/赶紧回去听solo Don't call me Nic
20,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,Jacqueline,2018-07-16,3,我到过顶峰也到过谷底，都空空如也。我一直在找一种声音。,66%31%3%,0
21,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,chushizhang,2018-08-17,3,cinematography多有出彩之处，但无论是角色还是演员的问题或是有意为之，被设定为经历过风浪的Nico行为像一只草原上的
22,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,INFITHELLs,2018-06-21,3,SIFF2018…7/10 看來是真樂迷了，視角沒能嚴守、敘事邏輯或缺都不成問題……橫移鏡頭下的明日派對就夠哭的了。I’l have it
23,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,王哈哈,2018-06-28,3,"其实这个导演是有感觉的，某些片段颇有味道，也确实对人物内心有不少挖掘。不过人物转变有些突兀，感觉是完全没铺垫够就被
随身携带的录音器是个很好的线索，应当有更多着墨和阐释的。
虽不成熟，但有感觉总比没感觉好，好好打磨剧本，熟练导演技巧，应该会有更好的作品。",66%31%3%,0
24,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,James Brown,2017-08-31,3,对Nico的印象之来自地下丝绒和切尔西女孩，不过这电影是Nico的传记吗？明明就是Christa。,66%31%3%,1
25,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,在胳肢窝跳舞,2018-06-29,3,从来没接触过地下丝绒和Nico，女主角的演绎很动人，影片比较普通吧。。。。,66%31%3%,0
26,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,宁二狗,2018-07-14,3,等我从伊维萨回来我就停止做音乐。1988妮可死于伊维萨。,66%31%3%,0
27,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,深邃,2018-07-14,3,刚现场听完John Cale就来看这片，做人呐到了巅峰以后要做的就是怎么摆脱这个巅峰的阴影了吧。,66%31%3%,0
28,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,Philia的青猫,2018-06-21,3,Siff第四场。每次都靠当年纪录片段和安迪沃霍尔的闪回回血。乐队巡演真的是一件很复杂的事情。当年的Nico做回自己，拿
29,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,李草木,2018-06-24,3,和女性小伙伴看完一致认为“贵圈真乱”，不到10人的男男女女排列组合真的会尴尬，唯一的好处是省些房费(歪果仁出差不拼房)。迟
30,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,宋人玫瑰,2018-08-10,3,上海电影节观影之一,66%31%3%,0
31,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,嗝,2018-07-07,3,我想成为一个优雅的老女人,66%31%3%,0
32,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,盛夏瞬噬,2018-06-28,3,反正就是无论经历过什么高潮低谷到头来都只是一片虚无吧……,66%31%3%,0
33,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,小R酱,2018-06-30,3,昏昏欲睡…只有音乐能听。,66%31%3%,0
34,"1988年的妮可 Nico, 1988",7.5,565,15.2%48.2%32.3%3.4%0.8%,2019-10-05,小青龙,2018-06-29,3,女主长相好像我的一个老师啊……,66%31%3%,0
```

All_movies_with_id.csv, 2.14GB

Processed dataset. We have used common Chinese stopwords and punctuations list to remove less useful elements. We utilize jieba segmentation tool to split each review as a list of words and find different parts of a Chinese sentence. With the calculation of TF-IDF, we found the most important words in one sentence and analyse its sentiment representation for the whole comment.

Model Design. The design of the overall framework is still in progress. Currently, we are discussing how to model the review word sequence. Typically, for a movie review containing several review sentences, there are two possible solutions: (1) regard the whole document as a long word sequence; (2) reserve the sentence boundaries and capture the hierarchical word-sentence & sentence-document relation. Also, in order to find more effective feature extractor, we are now comparing the performance of transformer with that of CNN/RNN on review representation learning.

- What we will do:

Emotion analysis modelling. Assume that there are few useless comments, which are issued by Internet marketers, inside Top 250 movies of Douban, we train one uniform-helpfulness model to analysis emotion of comments. The training dataset testing dataset are all Top 250 movies. The reference values will be the rank of each comment and the final rank. K-means clustering and dimension reduction algorithm(SVD) will be utilized to cluster movies according to reviewers' emotions per dimension. The main feature, or the bright spot of each movie, will also be presented in our output results.

Helpfulness evaluation modelling via RNN and CNN. After obtaining a sequence of review representations from our enhanced self-attention networks, we can borrow the recurrent neural networks (RNN) and convolutional neural networks (CNN) to calculate the helpfulness-aware overall representations. The training dataset and testing dataset include all movies information, i.e. ratings, reviews, descriptions and audiences, meanwhile, the related information with the long comments or from moviegoer reviewers will be considered as major reference for the evaluation of helpfulness. The useless comments will be given a lower weight and one new rank will be outputted. It is expected that this model can illustrate the “reality” of the rank, comparing the output of these two self-built models. The reality of Top 250 movies should approach 1 according to our assumptions.

Baseline model testing. Compare our emotion analysis results with other deep learning-based models and traditional models.

Online real-time testing with new movies. This is to demonstrate our ability to process big data with high velocity. Data streaming algorithm could be applied to analyse the new movie quality and propagation trend with the trained model. Ranks based on uniform-helpfulness model and helpfulness evaluation model will be given, compared to the final rank given by Douban.

Data visualization. The movies emotion map will be illustrated as a special heatmap to visualize the aggregated per-dimension emotion distribution. Clustering movies based on evoked emotion and emotion vector similarities might be also demonstrated by graphs. Besides, the estimated ratings, keywords and evaluation calculated by our models will be shown up to compare with original information.