

5.26总结：数据清洗、csv导入MySQL数据库

数据清洗

数据清洗即是爬下来未处理的数据文件（txt、json等格式），转化为csv文件，一般需要通过sed把里面不必要的字符删除，再以通过awk以一定规律分割，最后将脚本的处理结果重定向到csv文件中。

一般处理思路：
sed掐头-去尾-换行-置空值-删符号-删元素（列名）-awk分隔、添加列名 > 重定向到csv

操作类型	sed/awk 实现方案	示例
去除结构标记	删除JSON头尾	sed 's/^\{"data":\[//; s/\]\}\$//'
字段分隔	替换分隔符为换行符	sed 's/},{/\n/g'
特殊字符处理	移除转义字符	sed 's/\\//g'
空值标准化	统一空值表示	sed 's/"/null/g'
引号处理	删除所有引号	tr -d '"'
字段提取	按位置重组字段	awk '{print \$2,\$1,\$3}'

导出的csv文件用excel打开可能中文显示会乱码（wps不会），需正常显示可如下操作：

- 右键csv文件以txt方式打开-另存为-设置编码为ANSI-确定保存
此后用excel打开此文件即可正常显示中文

通过mysql对数据库服务器操作

```
mysql -h127.0.0.1 -P3306 -uroot -proot123456 test -e "select * from student"
```

- h：MySQL服务器的ip地址或主机名；
- u：连接MySQL服务器的用户名；
- e：执行mysql内部命令；
- p：连接MySQL服务器的密码。
- P：连接MySQL服务器的端口

作业练习

#上机练习15

#1.清洗数据《infotest.txt》保存成《infotest.csv》

```
sed 's/\[{\}/g' /root/sh/infotest.txt | #掐头
```

```
sed 's/}\]/g' | #去尾
```

```
sed 's/"/null/g' | #置空值
```

```
sed 's/"/g' | #删引号
```

#删除元素说明

```
sed 's/catchTime://g' |
```

```
sed 's/commentCount://g' |
```

```
sed 's/content://g' |
```

```
sed 's/createTime://g' |
```

```
sed 's/pic_list://g' |
```

```
sed 's/praiseCount://g' |
```

```
sed 's/reportCount://g' |
```

```
sed 's/source://g' |
```

#以逗号分割，添加列名，保存到csv

```
awk -v FS="," -v OFS="," 'BEGIN{print
```

```
"catchTime,commentCount,content,createTime,pic_list,praiseCount,reportCount,source"}{print $1,$2,$3,$4,$5,$6,$7,$8}' \
```

```
> /root/sh/infotest.csv
```

	A	B	C	D	E	F	G	H	I	J
1	catchTime	commentCount	content	createTime	pic_list	praiseCount	reportCount	source		
2	1.39E+09	1419	分享图片	1.39E+09	[http://www	5265	1285	iPad客户端		
3	1.39E+09	91	行走: #去	1.39E+09	[http://www	1	721	null		
4										
5										

#2.清洗数据《douban.txt》保存成《douban.csv》

```
sed 's/{\"subjects\":\[{\}/g' /root/sh/douban.txt | #掐头
```

```
sed 's/}\]/g' | #去尾
```

```
sed 's/},{/\\n/g' | #换行
```

```
sed 's/\\//g' | #删\\号
```

```
sed 's/"/null/g' | #置空值
```

```
sed 's/"/g' | #删引号
```

#删除元素说明

```
sed 's/id://g' |
```

```
sed 's/title://g' |
sed 's/episodes_info://g' |
sed 's/rate://g' |
sed 's/cover_x://g' |
sed 's/url://g' |
sed 's/playable://g' |
sed 's/cover://g' |
sed 's/cover_y://g' |
sed 's/is_new://g' |
#以逗号分割，添加列名，保存到csv
awk -v FS="," -v OFS="," 'BEGIN{print
"id,title,episodes_info,rate,cover_x,url,playable,cover,cover_y,is_new"}{print
$8,$4,$1,$2,$3,$5,$6,$7,$9,$10}' \
>/root/sh/douban.csv
```

x 自动保存 关 123 456 789 1011 1213 1415 1617 1819 2021 2223 2425 2627 2829 3031 3233 3435 3637 3839 4041 4243 4445 4647 4849 5051 5253 5455 5657 5859 6061 6263 6465 6667 6869 7071 7273 7475 7677 7879 8081 8283 8485 8687 8889 9091 9293 9495 9697 9899 100101 102103 104105 106107 108109 110111 112113 114115 116117 118119 120121 122123 124125 126127 128129 130131 132133 134135 136137 138139 140141 142143 144145 146147 148149 150151 152153 154155 156157 158159 160161 162163 164165 166167 168169 170171 172173 174175 176177 178179 180181 182183 184185 186187 188189 190191 192193 194195 196197 198199 200201 202203 204205 206207 208209 210211 212213 214215 216217 218219 220221 222223 224225 226227 228229 230231 232233 234235 236237 238239 240241 242243 244245 246247 248249 250251 252253 254255 256257 258259 260261 262263 264265 266267 268269 270271 272273 274275 276277 278279 280281 282283 284285 286287 288289 290291 292293 294295 296297 298299 300301 302303 304305 306307 308309 310311 312313 314315 316317 318319 320321 322323 324325 326327 328329 330331 332333 334335 336337 338339 340341 342343 344345 346347 348349 350351 352353 354355 356357 358359 360361 362363 364365 366367 368369 370371 372373 374375 376377 378379 380381 382383 384385 386387 388389 390391 392393 394395 396397 398399 400401 402403 404405 406407 408409 410411 412413 414415 416417 418419 420421 422423 424425 426427 428429 430431 432433 434435 436437 438439 440441 442443 444445 446447 448449 450451 452453 454455 456457 458459 460461 462463 464465 466467 468469 470471 472473 474475 476477 478479 480481 482483 484485 486487 488489 490491 492493 494495 496497 498499 500501 502503 504505 506507 508509 510511 512513 514515 516517 518519 520521 522523 524525 526527 528529 530531 532533 534535 536537 538539 540541 542543 544545 546547 548549 550551 552553 554555 556557 558559 560561 562563 564565 566567 568569 570571 572573 574575 576577 578579 580581 582583 584585 586587 588589 590591 592593 594595 596597 598599 600601 602603 604605 606607 608609 610611 612613 614615 616617 618619 620621 622623 624625 626627 628629 630631 632633 634635 636637 638639 640641 642643 644645 646647 648649 650651 652653 654655 656657 658659 660661 662663 664665 666667 668669 670671 672673 674675 676677 678679 680681 682683 684685 686687 688689 690691 692693 694695 696697 698699 700701 702703 704705 706707 708709 710711 712713 714715 716717 718719 720721 722723 724725 726727 728729 730731 732733 734735 736737 738739 740741 742743 744745 746747 748749 750751 752753 754755 756757 758759 760761 762763 764765 766767 768769 770771 772773 774775 776777 778779 780781 782783 784785 786787 788789 790791 792793 794795 796797 798799 800801 802803 804805 806807 808809 810811 812813 814815 816817 818819 820821 822823 824825 826827 828829 830831 832833 834835 836837 838839 840841 842843 844845 846847 848849 850851 852853 854855 856857 858859 860861 862863 864865 866867 868869 870871 872873 874875 876877 878879 880881 882883 884885 886887 888889 890891 892893 894895 896897 898899 900901 902903 904905 906907 908909 910911 912913 914915 916917 918919 920921 922923 924925 926927 928929 930931 932933 934935 936937 938939 940941 942943 944945 946947 948949 950951 952953 954955 956957 958959 960961 962963 964965 966967 968969 970971 972973 974975 976977 978979 980981 982983 984985 986987 988989 990991 992993 994995 996997 998999 10001001 10021003 10041005 10061007 10081009 10101011 10121013 10141015 10161017 10181019 10201021 10221023 10241025 10261027 10281029 10301031 10321033 10341035 10361037 10381039 10401041 10421043 10441045 10461047 10481049 10501051 10521053 10541055 10561057 10581059 10601061 10621063 10641065 10661067 10681069 10701071 10721073 10741075 10761077 10781079 10801081 10821083 10841085 10861087 10881089 10901091 10921093 10941095 10961097 10981099 11001101 11021103 11041105 11061107 11081109 11101111 11121113 11141115 11161117 11181119 11201121 11221123 11241125 11261127 11281129 11301131 11321133 11341135 11361137 11381139 11401141 11421143 11441145 11461147 11481149 11501151 11521153 11541155 11561157 11581159 11601161 11621163 11641165 11661167 11681169 11701171 11721173 11741175 11761177 11781179 11801181 11821183 11841185 11861187 11881189 11901191 11921193 11941195 11961197 11981199 12001201 12021203 12041205 12061207 12081209 12101211 12121213 12141215 12161217 12181219 12201221 12221223 12241225 12261227 12281229 12301231 12321233 12341235 12361237 12381239 12401241 12421243 12441245 12461247 12481249 12501251 12521253 12541255 12561257 12581259 12601261 12621263 12641265 12661267 12681269 12701271 12721273 12741275 12761277 12781279 12801281 12821283 12841285 12861287 12881289 12901291 12921293 12941295 12961297 12981299 13001301 13021303 13041305 13061307 13081309 13101311 13121313 13141315 13161317 13181319 13201321 13221323 13241325 13261327 13281329 13301331 13321333 13341335 13361337 13381339 13401341 13421343 13441345 13461347 13481349 13501351 13521353 13541355 13561357 13581359 13601361 13621363 13641365 13661367 13681369 13701371 13721373 13741375 13761377 13781379 13801381 13821383 13841385 13861387 13881389 13901391 13921393 13941395 13961397 13981399 14001401 14021403 14041405 14061407 14081409 14101411 14121413 14141415 14161417 14181419 14201421 14221423 14241425 14261427 14281429 14301431 14321433 14341435 14361437 14381439 14401441 14421443 14441445 14461447 14481449 14501451 14521453 14541455 14561457 14581459 14601461 14621463 14641465 14661467 14681469 14701471 14721473 14741475 14761477 14781479 14801481 14821483 14841485 14861487 14881489 14901491 14921493 14941495 14961497 14981499 15001501 15021503 15041505 15061507 15081509 15101511 15121513 15141515 15161517 15181519 15201521 15221523 15241525 15261527 15281529 15301531 15321533 15341535 15361537 15381539 15401541 15421543 15441545 15461547 15481549 15501551 15521553 15541555 15561557 15581559 15601561 15621563 15641565 15661567 15681569 15701571 15721573 15741575 15761577 15781579 15801581 15821583 15841585 15861587 15881589 15901591 15921593 15941595 15961597 15981599 16001601 16021603 16041605 16061607 16081609 16101611 16121613 16141615 16161617 16181619 16201621 16221623 16241625 16261627 16281629 16301631 16321633 16341635 16361637 16381639 16401641 16421643 16441645 16461647 16481649 16501651 16521653 16541655 16561657 16581659 16601661 16621663 16641665 16661667 16681669 16701671 16721673 16741675 16761677 16781679 16801681 16821683 16841685 16861687 16881689 16901691 16921693 16941695 16961697 16981699 17001701 17021703 17041705 17061707 17081709 17101711 17121713 17141715 17161717 17181719 17201721 17221723 17241725 17261727 17281729 17301731 17321733 17341735 17361737 17381739 17401741 17421743 17441745 17461747 17481749 17501751 17521753 17541755 17561757 17581759 17601761 17621763 17641765 17661767 17681769 17701771 17721773 17741775 17761777 17781779 17801781 17821783 17841785 17861787 17881789 17901791 17921793 17941795 17961797 17981799 18001801 18021803 18041805 18061807 18081809 18101811 18121813 18141815 18161817 18181819 18201821 18221823 18241825 18261827 18281829 18301831 18321833 18341835 18361837 18381839 18401841 18421843 18441845 18461847 18481849 18501851 18521853 18541855 18561857 18581859 18601861 18621863 18641865 18661867 18681869 18701871 18721873 18741875 18761877 18781879 18801881 18821883 18841885 18861887 18881889 18901891 18921893 18941895 18961897 18981899 19001901 19021903 19041905 19061907 19081909 19101911 19121913 19141915 19161917 19181919 19201921 19221923 19241925 19261927 19281929 19301931 19321933 19341935 19361937 19381939 19401941 19421943 19441945 19461947 19481949 19501951 19521953 19541955 19561957 19581959 19601961 19621963 19641965 19661967 19681969 19701971 19721973 19741975 19761977 19781979 19801981 19821983 19841985 19861987 19881989 19901991 19921993 19941995 19961997 19981999 20002001 20022003 20042005 20062007 20082009 20102011 20122013 20142015 20162017 20182019 20202021 20222023 20242025 20262027 20282029 20302031 20322033 20342035 20362037 20382039 20402041 20422043 20442045 20462047 20482049 20502051 20522053 20542055 20562057 20582059 20602061 20622063 20642065 20662067 20682069 20702071 20722073 20742075 20762077 20782079 20802081 20822083 20842085 20862087 20882089 20902091 20922093 20942095 20962097 20982099 21002101 21022103 21042105 21062107 21082109 21102111 21122113 21142115 21162117 21182119 21202121 21222123 21242125 21262127 21282129 21302131 21322133 21342135 21362137 21382139 21402141 21422143 21442145 21462147 21482149 21502151 21522153 21542155 21562157 21582159 21602161 21622163 21642165 21662167 21682169 21702171 21722173 21742175 21762177 21782179 21802181 21822183 21842185 21862187 21882189 21902191 21922193 21942195 21962197 21982199 22002201 22022203 22042205 22062207 22082209 22102211 22122213 22142215 22162217 22182219 22202221 22222223 22242225 22262227 22282229 22302231 22322233 22342235 22362237 22382239 22402241 22422243 22442245 22462247 22482249 22502251 22522253 22542255 22562257 22582259 22602261 22622263 22642265 22662267 22682269 22702271 22722273 22742275 22762277 22782279 22802281 22822283 22842285 22862287 22882289 22902291 22922293 22942295 22962297 22982299 23002301 23022303 23042305 23062307 23082309 23102311 23122313 23142315 23162317 23182319 23202321 23222323 23242325 23262327 23282329 23302331 23322333 23342335 23362337 23382339 23402341 23422343 23442345 23462347 23482349 23502351 23522353 23542355 23562357 23582359 23602361 23622363 23642365 23662367 23682369 23702371 23722373 23742375 23762377 23782379 23802381 23822383 23842385 23862387 23882389 23902391 23922393 23942395 23962397 23982399 24002401 24022403 24042405 24062407 24082409 24102411 24122413 24142415 24162417 24182419 24202421 24222423 24242425 24262427 24282429 24302431 24322433 24342435 24362437 24382439 24402441 24422443 24442445 24462447 24482449 24502451 24522453 24542455 24562457 24582459 24602461 24622463 24642465 24662467 24682469 24702471 24722473 24742475 24762477 24782479 24802481 24822483 24842485 24862487 24882489 24902491 24922493 24942495 24962497 24982499 25002501 25022503 25042505 25062507 25082509 25102511 25122513 25142515 25162517 25182519 25202521 25222523 25242525 25262527 25282529 25302531 25322533 25342535 25362537 25382539 25402541 25422543 25442545 25462547 25482549 25502551 25522553 25542555 25562557 25582559 25602561 25622563 25642565 25662567 25682569 25702571 25722573 25742575 25762577 25782579 25802581 25822583 25842585 25862587 25882589 25902591 25922593 25942595 25962597 25982599 26002601 26022603 26042605 26062607 26082609 26102611 26122613 26142615 26162617 26182619 26202621 26222623 26242625 26262627 26282629 26302631 26322633 26342635 26362637 26382639 26402641 26422643 26442645 26462647 26482649 26502651 26522653 26542655 26562657 26582659 26602661 26622663 26642665 26662667 26682669 26702671 26722673 26742675 26762677 26782679 26802681 26822683 26842685 26862687 26882689 26902691 26922693 26942695 26962697 26982699 27002701 27022703 27042705 27062707 27082709 27102711 27122713 27142715 27162717 27182719 27202721 27222723 27242725 27262727 27282729 27302731 27322733 27342735 27362737 27382739 27402741 27422743 27442745 27462747 27482749 27502751 27522753 27542755 27562757 27582759 27602761 27622763 27642765 27662767 27682769 27702771 27722773 27742775 27762777 27782779 27802781 27822783 27842785 27862787 27882789 27902791 27922793 27942795 27962797 27982799 28002801 28022803 28042805 28062807 28082809 28102811 28122813 28142815 28162817 28182819 28202821 28222823 28242825 28262827 28282829 28302831 28322833 28342835 28362837 28382839 28402841 28422843 28442845 28462847 28482849 28502851 28522853 28542855 28562857 28582859 28602861 28622863 28642865 28662867 28682869 28702871 28722873 28742875 28762877 28782879 28802881 28822883 28842885 28862887 28882889 28902891 28922893 28942895 28962897 28982899 29002901 29022903 29042905 29062907 29082909 29102911 29122913 29142915 29162917 29182919 29202921 29222923 29242925 29262927 29282929 29302931 29322933 29342935 29362937 29382939 29402941 29422943 29442945 29462947 29482949 29502951 29522953 29542955 29562957 29582959 29602961 29622963 29642965 29662967 29682969 29702971 29722973 29742975 29762977 29782979 29802981 29822983 29842985 29862987 298

#上机练习16

#编写shell脚本/root/shell/mysqlcsv.sh 实现:

#1.如果douban表存在则删除

#2.如果douban表不存在则新建

#3.导入douban.csv数据到douban表中,建表过程为根据导入的csv文件自动创建表

#4.查询douban表验证结果

#设置mysql连接参数

dbhost="127.0.0.1" #主机名

dbprot=3306 #端口

dbuser="root" #用户名

dbpass="root123456" #密码

db="test" #表空间

#如果douban表存在则删除

sql1="drop table if exists douban"

mysql -h\$dbhost -P\$dbprot -u\$dbuser -p\$dbpass \$db -e "\$sql1"

#获取第一行(列名)

names=`head -n +1 /root/sh/douban.csv | sed 's/,/ /g'`

#找出最后一个列名

lastname=`echo "\$names" | awk '{print \$NF}'`

#如果douban表不存在则新建

sql2="create table if not exists douban("

for i in \$names

do

if [\$i == "\$lastname"]

then

sql2="\$sql2\$i varchar(200))"

else

sql2="\$sql2\$i varchar(200),"

fi

done

mysql -h\$dbhost -P\$dbprot -u\$dbuser -p\$dbpass \$db -e "\$sql2"

#导入数据

sql3="LOAD DATA INFILE '/usr/local/mysql/data/douban.csv' INTO TABLE douban
CHARACTER SET utf8

FIELDS TERMINATED BY ','

LINEs TERMINATED BY '\n'

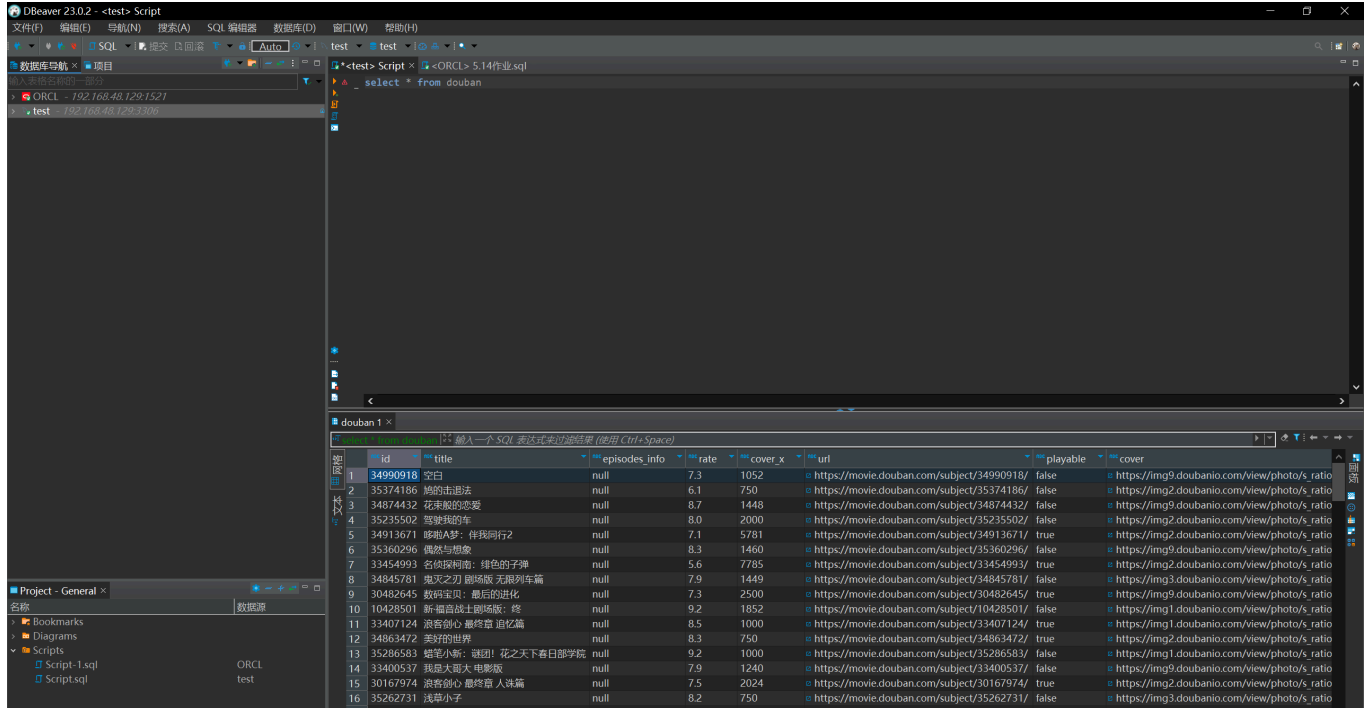
IGNORE 1 LINEs"

mysql -h\$dbhost -P\$dbprot -u\$dbuser -p\$dbpass \$db -e "\$sql3"

#查看表和数据

```
sql4="select * from douban"
```

```
mysql -h$dbhost -P$dbprot -u$dbuser -p$dbpass $db -e "$sql4"
```



The screenshot shows the DBeaver 23.0.2 interface. The top toolbar includes icons for File, Edit, Run, SQL, and other database-related functions. The main editor displays a SQL script with the following content:

```
test > test
--> test > Script > <ORCL> > 5.14作业.sql
select * from douban
```

Below the editor, the 'douban 1' table is displayed with the following columns: id, title, episodes_info, rate, cover_x, url, playable, and cover. The table contains 16 rows of data, each representing a movie entry from Douban.

id	title	episodes_info	rate	cover_x	url	playable	cover
34990918	空白	null	7.3	1052	https://movie.douban.com/subject/34990918/	false	https://img9.doubanio.com/view/photo/s_ratio
35374186	她的未闻法	null	6.1	750	https://movie.douban.com/subject/35374186/	false	https://img2.doubanio.com/view/photo/s_ratio
34874432	花束般的恋爱	null	8.7	1448	https://movie.douban.com/subject/34874432/	false	https://img9.doubanio.com/view/photo/s_ratio
35235502	驾驶我的车	null	8.0	2000	https://movie.douban.com/subject/35235502/	false	https://img2.doubanio.com/view/photo/s_ratio
34913671	哆啦A梦：伴我同行2	null	7.1	5781	https://movie.douban.com/subject/34913671/	true	https://img2.doubanio.com/view/photo/s_ratio
35360296	偶然与想象	null	8.3	1460	https://movie.douban.com/subject/35360296/	false	https://img2.doubanio.com/view/photo/s_ratio
33454993	名侦探柯南：绯色的子弹	null	5.6	7785	https://movie.douban.com/subject/33454993/	true	https://img2.doubanio.com/view/photo/s_ratio
34845781	鬼灭之刃 剧场版 无限列车篇	null	7.9	1449	https://movie.douban.com/subject/34845781/	true	https://img2.doubanio.com/view/photo/s_ratio
30482645	数码宝贝：最后的进化	null	7.3	2500	https://movie.douban.com/subject/30482645/	true	https://img9.doubanio.com/view/photo/s_ratio
10428501	新·福音战士剧场版：终	null	9.2	1852	https://movie.douban.com/subject/10428501/	false	https://img1.doubanio.com/view/photo/s_ratio
33407124	浪客剑心 最终章 追忆篇	null	8.5	1000	https://movie.douban.com/subject/33407124/	true	https://img1.doubanio.com/view/photo/s_ratio
34863472	姜子牙的世界	null	8.3	750	https://movie.douban.com/subject/34863472/	true	https://img2.doubanio.com/view/photo/s_ratio
35286583	雄狮少年	null	9.2	1000	https://movie.douban.com/subject/35286583/	false	https://img1.doubanio.com/view/photo/s_ratio
33400537	我是大哥大 电影版	null	7.9	1240	https://movie.douban.com/subject/33400537/	false	https://img9.doubanio.com/view/photo/s_ratio
30167974	浪客剑心 最终章 人诛篇	null	7.5	2024	https://movie.douban.com/subject/30167974/	true	https://img9.doubanio.com/view/photo/s_ratio
35262731	浅草小子	null	8.2	750	https://movie.douban.com/subject/35262731/	false	https://img3.doubanio.com/view/photo/s_ratio