

Pig

Prior Work
On Datasets

Data
Processing

HDFS

Improved
evaluation
model

Estimate
Performance
using new
model

Summary

Mahout

user-based
collaborative
filtering
Recommender
Systems

Item-based
collaborative
filtering
Recommender
Systems

2020/08/13

Prior Work

- [https://grouplens.org/datasets/
personality-2018/](https://grouplens.org/datasets/personality-2018/)
- Enhancing User Experience with
Recommender Systems Beyond Prediction
Accuracies --by Joseph A. Konstan and Loren
Terveen August, 2016

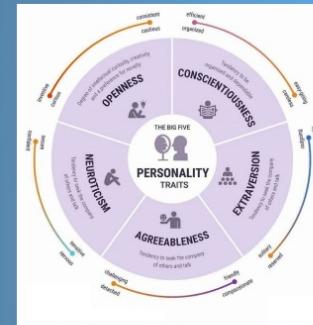
personality.csv

rating.csv

Goal!

personality-data.csv

1. User ID
 2. User_personality: OCEAN five character model
 3. Predicted Result: 12 movies and predictions
 4. Evaluation for Prediction: is_personalized and enjoy_watching



Prior Work

- [https://grouplens.org/datasets/
personality-2018/](https://grouplens.org/datasets/personality-2018/)
- Enhancing User Experience with
Recommender Systems Beyond Prediction
Accuracies --by Joseph A. Konstan and Loren
Terveen August, 2016

personality.csv

rating.csv

Goal!

rating.csv

1. User ID
2. Movie ID
3. Rating
4. Time Stamp

	user	movie	rate	cam	cam
1	1	1	1	1	1
2	2	1	1	1	1
3	3	1	1	1	1
4	4	1	1	1	1

	A	B	C	D	E
1	user	movie_id	rating	tstamp	
2	8e7cebf9a2	1	5	2001-09-10 17:19:56	
3	8e7cebf9a2	2	4	2001-09-28 11:34:55	
4	8e7cebf9a2	3	4	2001-09-28 11:42:50	
5	8e7cebf9a2	5	5	2001-09-28 11:27:30	
6	8e7cebf9a2	6	4	2002-01-07 18:12:02	
7	8e7cebf9a2	7	4	2001-09-21 14:09:24	
8	8e7cebf9a2	10	4	2001-09-29 18:13:15	
9	8e7cebf9a2	11	4	2001-09-21 11:54:21	
10	8e7cebf9a2	12	4	2001-09-29 18:58:15	
11	8e7cebf9a2	13	4	2001-09-14 15:29:46	

Prior Work

- [https://grouplens.org/datasets/
personality-2018/](https://grouplens.org/datasets/personality-2018/)
- Enhancing User Experience with
Recommender Systems Beyond Prediction
Accuracies --by Joseph A. Konstan and Loren
Terveen August, 2016

personality.csv

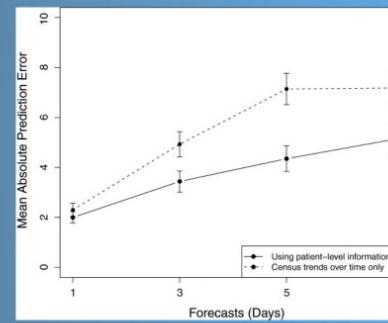
rating.csv

Goal!

AbsoluteAverageDifference

- I will evaluate the performance of this enhanced user-based recommender by the absolute average difference between system recognition and user's feedback.

Why not use the predication data of 12 movies?



Prior Work

- [https://grouplens.org/datasets/
personality-2018/](https://grouplens.org/datasets/personality-2018/)
- Enhancing User Experience with
Recommender Systems Beyond Prediction
Accuracies --by Joseph A. Konstan and Loren
Terveen August, 2016

personality.csv

rating.csv

Goal!

Pig

Prior Work
On Datasets

Data
Processing

HDFS

Improved
evaluation
model

Estimate
Performance
using new
model

Summary

Mahout

user-based
collaborative
filtering
Recommender
Systems

Item-based
collaborative
filtering
Recommender
Systems

2020/08/13

Pig operation

- Obtain Original data from questionnaire - researchData.csv
- Average value of each users' (is_personalized) - is_personalized.csv
- Average value of each users' (enjoy_watching) - enjoy_watching.csv

Assume each user in the survey offers fair and reasonable evaluation.

we have the AbsoluteAverageDifference is

$3.606783 - 3.524451 =$
0.082332

Join
personality.cs
v and table.csv

MapReduce(userid,
is_personalized)

MapReduce(userid,
enjoy_watchingg)

Join operation

researchData = JOIN personality BY userid, rating BY user;

```
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.7.3 0.17.0 keyulu 2020-08-12 09:14:32 2020-08-12 09:14:49 HASH_JOIN

Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime     MinMapTime   AvgMapTime   MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime  Alias  Feature
job_local431138236_0016 3      1          n/a          n/a          n/a          n/a          n/a          n/a          n/a          n/a          HASH_JOIN          file:///home/ke

Input(s):
Successfully read 1835 records from: "file:///home/keyulu/Downloads/pig-0.17.0/bin/personality.csv"
Successfully read 1028752 records from: "file:///home/keyulu/Downloads/pig-0.17.0/bin/ratings.csv"

Output(s):
Successfully stored 1070203 records in: "file:///home/keyulu/Downloads/pig-0.17.0/bin/output/researchData"

Counters:
Total records written : 1070203
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local431138236_0016

2020-08-12 09:14:49,139 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-12 09:14:49,139 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-12 09:14:49,139 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2020-08-12 09:14:49,141 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
(005fe8678214011d7f92e51f9a546d40, 5.5, 6.0, 2.0, 4.5, 1.0, popularity, low, 94466, 4.17266086321, 77658, 4.19960847544, 92259, 4.14083397435, 112556, 4
.12552693679, 3435, 4.22468194276, 44555, 4.23534206643, 27773, 4.09204931497, 1, 3 ,005fe8678214011d7f92e51f9a546d40, 1, 4, 2015-05-18 21:52:01 )
(005fe8678214011d7f92e51f9a546d40, 5.5, 6.0, 2.0, 4.5, 1.0, popularity, low, 94466, 4.17266086321, 77658, 4.19960847544, 92259, 4.14083397435, 112556, 4
.12552693679, 3435, 4.22468194276, 44555, 4.23534206643, 27773, 4.09204931497, 1, 3 ,005fe8678214011d7f92e51f9a546d40, 50, 4, 2015-05-19 17:58:26 )

researchData | personality::userid:bytearray | personality::openness:bytearray | personality::a
array | personality::assignedmetric:bytearray | personality::assignedcondition:bytearray | pers
array | personality::movie_3:bytearray | personality::predicted_rating_3:bytearray | personality
| personality::movie_6:bytearray | personality::predicted_rating_6:bytearray | personality::movie
personality::movie_9:bytearray | personality::predicted_rating_9:bytearray | personality::movie_10:byt
sonality::movie_12:bytearray | personality::predicted_rating_12:bytearray | personality::ts_personaliz
y | rating:::timestamp:bytearray |
```

Pig operation

- Obtain Original data from questionnaire - researchData.csv
- Average value of each users' (is_personalized) - is_personalized.csv
- Average value of each users' (enjoy_watching) - enjoy_watching.csv

Assume each user in the survey offers fair and reasonable evaluation.

we have the AbsoluteAverageDifference is

$3.606783 - 3.524451 =$
0.082332

Join
personality.cs
v and table.csv

MapReduce(userid,
is_personalized)

MapReduce(userid,
enjoy_watchingg)

enjoy_watching.csv

```
grpds = GROUP userinfo BY userid;
enjoy_watching = FOREACH grpds GENERATE group, AVG(userinfo.enjoy_watching);
STORE enjoy_watching INTO 'output/enjoy_watching' USING PigStorage(',');
```

```
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.7.3 0.17.0 keyulu 2020-08-12 12:47:17 2020-08-12 12:47:26 GROUP_BY

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime      MinMapTime    AvgMapTime   MedianMapTime  MaxReduceTime  MinReduceTime
job_local362269726_0001 3          1           n/a          n/a          n/a          n/a          n/a          n/a          enjoy_watching

Input(s):
Successfully read 1070204 records from: "file:///home/keyulu/Downloads/pig-0.17.0/bin/userinfo.csv"

Output(s):
Successfully stored 1821 records in: "file:///home/keyulu/Downloads/pig-0.17.0/bin/output/enjoy_watching"

Counters:
Total records written : 1821
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local362269726_0001
```

C1	A	B	C	D	E	F
1	userid		3524451			
2	005fe878		3			
3	0066fa81		4			
4	00fa91e20		4			
5	011aedbe		4			
6	01332544		4			
7	01609dc56		4			
8	01a88049t		2			
9	01aedde5		3			

Pig operation

- Obtain Original data from questionnaire - researchData.csv
- Average value of each users' (is_personalized) - is_personalized.csv
- Average value of each users' (enjoy_watching) - enjoy_watching.csv

Assume each user in the survey offers fair and reasonable evaluation.

we have the AbsoluteAverageDifference is

$$3.606783 - 3.524451 = \\ \boxed{0.082332}$$

Join
personality.cs
v and table.csv

MapReduce(userid,
is_personalized)

MapReduce(userid,
enjoy_watchingg)

rating.csv

```
grpds = GROUP userinfo BY userid;
rating = FOREACH grpds GENERATE group, AVG(userinfo.rating);
STORE rating INTO 'output/userinfo.rating' USING PigStorage('');
```

```
HadoopVersion    PigVersion      UserId  StartedAt    FinishedAt      Features
2.7.3    0.17.0   keyuliu  2020-08-12 12:47:17  2020-08-12 12:47:26  GROUP_BY

Success!

Job Stats (time in seconds):
JobId  Maps   Reduces  MaxMapTime   MinMapTime   AvgMapTime   MedianMapTime   MaxReduceTime   MinRe
job_local362269726_0001 3       1          n/a          n/a          n/a          n/a          n/a          n/a          enjoy

Input(s):
Successfully read 1070284 records from: "file:///home/keyuliu/Downloads/pig-0.17.0/bin/userinfo.csv"

Output(s):
Successfully stored 1821 records in: "file:///home/keyuliu/Downloads/pig-0.17.0/bin/output/enjoy_watching"

Counters:
Total records written : 1821
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local362269726_0001
```

	A	B	C
1	userid		3.606783
2	005fe8678	4.111111	
3	0066fa81	3.971545	
4	00fa91e20	3.327285	
5	011aedbe	3.387879	

Pig operation

- Obtain Original data from questionnaire - researchData.csv
- Average value of each users' (is_personalized) - is_personalized.csv
- Average value of each users' (enjoy_watching) - enjoy_watching.csv

Assume each user in the survey offers fair and reasonable evaluation.

we have the AbsoluteAverageDifference is

$3.606783 - 3.524451 =$
0.082332

Join
personality.cs
v and table.csv

MapReduce(userid,
is_personalized)

MapReduce(userid,
enjoy_watchingg)

Pig

Prior Work
On Datasets

Data
Processing

HDFS

Improved
evaluation
model

Estimate
Performance
using new
model

Summary

Mahout

user-based
collaborative
filtering
Recommender
Systems

Item-based
collaborative
filtering
Recommender
Systems

2020/08/13

Limitation

In fact, people with different personality will give different comment for same things even they have similarity taste.



1. Separate people by personality

2. Predicted Result Symbol

3. Realistic FeedBack Symbol

Using Personality To Distinguish User

In MapReduce, the OCEAN five factor will be used to decide how partitioner works.

userid	openness	agreeable	emotional	conscientious	extraversive
8e7cebf9a2	5	2	3	2.5	6.5

Limitation

In fact, people with different personality will give different comment for same things even they have similarity taste.



1. Separate people by personality

2. Predicted Result Symbol

3. Realistic FeedBack Symbol

assignedcondition

As we have used enjoy_watching as predicted result from this enhanced user-based recommender system

Now the assignedcondition (high, midterm, low) will be assumed as the objective benchmark for movies.

To transfer these new benckmark to numbers in a resonable way. I will use quartering.

For example, there are totally 11 numbers from 0 to 10.
the value of

$$\text{High} = (11+1) * 0.75$$

$$\text{midterm} = (1+1) * 0.5$$

$$\text{low} + (11+1) * 0.25$$

Limitation

In fact, people with different personality will give different comment for same things even they have similarity taste.



1. Separate people by personality

2. Predicted Result Symbol

3. Realistic FeedBack Symbol

$a * \text{is_personalized} + b * \text{enjoy_watching}$
 $(a+b=2)$

`is_personalized +enjoy_watching` will represent the feedback from user from they used this recommender system.

As `is_personalized` is from (1-5) and `enjoy_watching` is also from (1-5).

And they need to sub the predicted result-assigned condition. So in the model I decide to multiple the assigned condition to make them balanced when calculating absolute average difference.

The representation of realistic feedback will be:

$a * \text{is_personalized} + b * \text{enjoy_watching} (a+b=2)$

Limitation

In fact, people with different personality will give different comment for same things even they have similarity taste.



1. Separate people by personality

2. Predicted Result Symbol

3. Realistic FeedBack Symbol

Pig

Prior Work
On Datasets

Data
Processing

HDFS

Improved
evaluation
model

Estimate
Performance
using new
model

Summary

Mahout

user-based
collaborative
filtering
Recommender
Systems

Item-based
collaborative
filtering
Recommender
Systems

2020/08/13

Map Reduce

Discrete Function of AAD and personality

openness	agreeableness	emotional_stability	conscientiousness	extraversion	counter	absoluteDifference	averageAbsoluteDifference	1070203
20	30	25	60	15	1537	15370	10	
35	35	25	65	15	84	840	10	
45	35	50	40	15	1007	10070	10	
50	35	45	30	25	688	6880	10	
55	40	65	25	25	27	240	10	
55	50	30	50	40	67	670	10	
55	50	65	55	55	1210	12100	10	
55	55	45	35	30	591	5910	10	
55	70	50	60	10	357	3570	10	
65	45	60	40	60	488	4880	10	

Converting to PDF using distributed linear algebra or knn/SVM. Then, building a model to analysis the relationship between personality and AAD.

ScreenShoot

bin/hadoop fs -mkdir /
EvaluationModel

bin/hadoop fs -copyFromLocal ~/
Downloads/userinfo.csv /input

bin/hadoop jar ~/Downloads/
SecondarySort-1.0-SNAPSHOT.jar
DriverClass /input/userinfo.csv /
result/output/Evaluation
bin/hadoop fs -copyToLocal /result/
output/Evaluation/part-r-00000 ~/
Downloads/Project

```
2020-08-13 21:31:08,494 INFO mapreduce.Job: Counters: 53
   File System Counters
      File output bytes: 49201962
      FILE1: Number of bytes written:94638807
      FILE1: Number of bytes written:94638807
      FILE1: Number of read operations=0
      FILE1: Number of large read operations=0
      FILE1: Number of write operations=0
      FILE1: Number of large write operations=0
      HDFS: Number of bytes read=8778697
      HDFS: Number of bytes written:94638808
      HDFS: Number of read operations=0
      HDFS: Number of read operations=0
      HDFS: Number of large read operations=0
      HDFS: Number of large write operations=0
      Job Counters
         Launched map tasks=1
         Launched reduce tasks=1
         Data-local map tasks=1
         Total time spent by all map tasks in occupied slots (ms)=10821
         Total time spent by all reduces in occupied slots (ms)=5108
         Total time spent by all map tasks (ms)=10821
         Total time spent by all reduce tasks (ms)=5108
         Total vcore-milliseconds taken by all map tasks=10821
         Total vcore-milliseconds taken by all reduce tasks=5108
         Total megabyte-milliseconds taken by all map tasks=10874
         Total megabyte-milliseconds taken by all reduce tasks=522408
Map-Reduce Framework
   Map input records=1070203
   Map output records=1070203
   Map output bytes=1070203
   Map output materialized bytes=48201962
   Input split bytes=107
   Combine input records=0
   Reduce input records=1778
   Reduce input groups=1778
   Reduce input bytes=1070203
   Reduce input records=1070203
   Reduce output records=1778
   Spills Local=0, Remote=0, External=0
   Shuffled Maps =1
   Failed Shuffles=0
   Merged Map outputs=1
   GC time elapsed (ms)=886
   CPU time spent (ms)=1000
   Physical memory (bytes) snapshot:778875392
   Virtual memory (bytes) snapshot:119811584
   Total physical memory (bytes)=463622144
   Peak Map Physical memory (bytes)=463622144
   Peak Map Virtual memory (bytes)=2556987520
   Peak Reduce Physical memory (bytes)=307233248
   Peak Reduce Virtual memory (bytes)=252298864
Shuffle Errors
   SASL Error Count=0
   CONNECTIONs=0
   IO_Error_Count=0
   NETWORK_ERRORs=0
   WRONG_MAP=0
   WRONG_PARTITION=0
File Input Format Counters
  Bytes Read=8678594
  File Output Format Counters
  Bytes Written=5108
```

Map Reduce

Discrete Function of AAD and personality

openness	agreeableness	emotional_stability	conscientiousness	extraversion	counter	absoluteDifference	averageAbsoluteDifference	1070203
20	30	40	60	15	1537	15370	10	
35	35	25	65	15	84	840	10	
45	35	50	40	15	1007	10070	10	
50	35	45	30	25	688	6880	10	
55	40	65	25	25	247	2470	10	
55	50	30	50	40	67	670	10	
55	50	65	55	55	1210	12100	10	
55	55	45	35	30	591	5910	10	
55	70	50	60	10	357	3570	10	
65	45	60	40	60	488	4880	10	

Converting to PDF using distributed linear algebra or knn/SVM. Then, building a model to analysis the relationship between personality and AAD.

Pig

Prior Work
On Datasets

Data
Processing

HDFS

Improved
evaluation
model

Estimate
Performance
using new
model

Summary

Mahout

user-based
collaborative
filtering
Recommender
Systems

Item-based
collaborative
filtering
Recommender
Systems

2020/08/13

AbsoluteAverageDifference from hold on test(0.9,0.1)

Outout is unstable. But as more test is proceed, the AAD tends to 0.93

```
C:\Program Files\Java\jre6\bin\java.exe" -cp
28/08/13 21:49:17 INFO file.FileDataModel: Creating FileDataModel for file E:\INFO7250\ratings.csv
28/08/13 21:49:17 INFO file.FileDataModel: Creating FileDataModel for file E:\INFO7250\ratings.csv
28/08/13 21:49:17 INFO file.FileDataModel: Reading file info...
28/08/13 21:49:17 INFO file.FileDataModel: Processed 1000000 times
28/08/13 21:49:17 INFO file.FileDataModel: Read lines: 1629751
28/08/13 21:49:17 INFO model.GenericCfModel: Processed 1820 users
28/08/13 21:49:17 INFO eval.AbstractDifferenceRecommenderEvaluator: Beginning evaluation using 0.9 of FileDataModel[dataFile:E:\INFO7250\ratings.csv]
28/08/13 21:49:17 INFO eval.AbstractDifferenceRecommenderEvaluator: Beginning evaluation of 1887 users
28/08/13 21:49:17 INFO eval.AbstractDifferenceRecommenderEvaluator: Average time per recommendation: 117ms
28/08/13 21:49:17 INFO eval.StatCallable: Average time per recommendation: 117ms
28/08/13 21:49:17 INFO eval.StatCallable: Approximate memory used: 112MB / 264MB
28/08/13 21:49:17 INFO eval.StatCallable: Unable to recommend in 22 cases
28/08/13 21:49:17 INFO eval.StatCallable: Average time per recommendation: 1117ms
28/08/13 21:49:17 INFO eval.StatCallable: Approximate memory used: 252MB / 445MB
28/08/13 21:57:41 INFO eval.StatCallable: Unable to recommend in 58248 cases
28/08/13 21:57:41 INFO eval.AbstractDifferenceRecommenderEvaluator: Evaluation result: 0.881599649139872
0.881599649139872

Process finished with exit code 0
```

Pig

Prior Work
On Datasets

Data
Processing

HDFS

Improved
evaluation
model

Estimate
Performance
using new
model

Summary

Mahout

user-based
collaborative
filtering
Recommender
Systems

Item-based
collaborative
filtering
Recommender
Systems

2020/08/13

Item-based

Using item itself as if using UserNeighborhood

```
28/08/13 22:01:50 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 139598
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 138816
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 139598
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 139669
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 141741
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 142893
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 143866
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 144168
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 162784
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 175663
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 184451
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 186491
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 191157
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 192627
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 196129
28/08/13 2 28/08/13 22:02:59 INFO eval.AbstractDifferenceRecommenderEvaluator: Item exists in test data but not training data: 197675
0.67836899
0.674816702217727
Process finished with exit code 0
```

Why output is stable and the prediction is precise ?

Pig

Prior Work
On Datasets

Data
Processing

HDFS

Improved
evaluation
model

Estimate
Performance
using new
model

Summary

Mahout

user-based
collaborative
filtering
Recommender
Systems

Item-based
collaborative
filtering
Recommender
Systems

2020/08/13

Summary

Limitation : can't use predicted movield with it predicted rating.

Future work: using KNN-item based build new recommender system instead of GenericItemBasedRecommender in API.

Pig

Prior Work
On Datasets

Data
Processing

HDFS

Improved
evaluation
model

Estimate
Performance
using new
model

Summary

Mahout

user-based
collaborative
filtering
Recommender
Systems

Item-based
collaborative
filtering
Recommender
Systems

2020/08/13