Báo cáo đánh giá hiệu quả của thuật toán collaborative filtering

Với thuật toán Collaborative Filtering chúng ta sử dụng similarity hay "độ tương đồng" để đưa ra gợi ý.

Với user-based Collaborative Filtering thì độ tương có thể tính bằng phương pháp đo lường cosine, Pearson, Euclidean, …

Similarity Measures	**Computational formulae's**
Pearson Correlation(PCC)	$sim\left(u,u'\right)^{PCC} = \frac{\sum_{i \in I}\left(r_{u,i} - \bar{r}_u\right) - \left(r_{u',i} - \bar{r}_{u'}\right)}{\sqrt{\sum_{i \in I}\left(r_{u,i} - \bar{r}_u\right)^2} \cdot \sqrt{\sum_{i \in I}\left(r_{u',i} - \bar{r}_{u'}\right)^2}}$ Where I is the set of items, $r_{u,i}$ rating of given to item i by user u, $\bar{r}_u$ average rating of user u
Cosine (COS)	$sim\left(u,u'\right)^{COS} = \frac{\sum_{i \in I}\left(r_{u,i}\right) \cdot \left(r_{u',i}\right)}{\sqrt{\sum_{i \in I}\left(r_{u,i}\right)^2 \cdot \sum_{i \in I}\left(r_{u',i}\right)^2}}$
Adjusted Cosine (ACOS)	$sim\left(i,i'\right)^{ACOS} = \frac{\sum_{u \in U}\left(r_{u,i} - \bar{r}_i\right) \cdot \left(r_{u,i'} - \bar{r}_{i'}\right)}{\sqrt{\left(r_{u,i} - \bar{r}_i\right)^2} \cdot \sqrt{\left(r_{u,i'} - \bar{r}_{i'}\right)^2}}$ Where U is the set of users rated both items i.e. i and i'
Spearman's Rank Correlation similarity	$sim\left(u,u'\right)^{SRCC} = 1 - \frac{6\sum_{i \in I} \text{rank}\left(r_{u,i}\right)^2 - \text{rank}\left(r_{u',i}\right)^2}{\left\|I\right\| \cdot \left(\left\|I\right\|^2 - 1\right)}$ Where $\left\|I\right\|$ is the cardinality of co-rated items.

Hình 1. Trích - A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment

Hiệu quả và độ chính xác của các dự đoán được xác định dựa trên tham số đó chính là RMSE (Root Mean Squared Error) hoặc MAE (Mean Absolute Error).

MAE sẽ được tính bằng công thức sau:

$$MAE = \frac{1}{|S|}\sum_{i=1}^{S}|Pred_i - r_i|$$

RMSE sẽ được tính bằng công thức sau:

$$RMSE = \sqrt{\frac{1}{|S|}\sum_{i=1}^{S}(Pred_i - r_i)^2}$$

$Pred_i$ là dự đoán điểm $rating$ cho đối tượng i, r_i là $rating$ thực tế của đối tượng i và $|S|$ là bản chất của $test\ ratings$.

RMSE là một số thực không âm sẽ có giá trị từ 0.0 và 0.0 là trường hợp tốt nhất khi mà dự đoán không có sai lệch.

Để thực hiện tính toán 2 tham số này trong python chúng ta có thể sử dụng 2 thư viện là

`from sklearn.metrics import mean_absolute_error` để tính MAE
và `from sklearn.metrics import mean_squared_error và set squared = False` để tính RMSE

Bằng cách áp dụng vào bài demo chúng ta thu được kết quả như sau:

1. Với độ tương đồng cosine:

```
-------
[('0002005018', 3)]
Dự đoán với độ tương đồng cosine
--------------------------
0002005018    0.500000
0452264464    0.164902
0679425608    0.160418
0060973129    0.159440
1881320189    0.151920
dtype: float64
RMSE: 0.9466964501285277
MSE: 0.9098900151643848
```

Với mỗi rating được dự đoán sẽ sai lệch khoảng 0.94 nếu người dùng chỉ đánh giá một cuốn sách điều nay cho chúng ta thấy hiệu suất ở lần dự đoán này là thấp.
Tương tự nếu chúng ta thực hiện dự đoán tiếp:

```
[('0002005018', 5), ('0060973129', 3)]
Dự đoán với độ tương đồng cosine
--------------------------
0002005018    2.659440
0060973129    1.297202
0679425608    0.964205
1881320189    0.954087
0452264464    0.788365
dtype: float64
RMSE: 0.8313380144718675
MSE: 0.7776357838780574
----------------------------
```

Với mỗi rating được dự đoán sẽ sai lệch khoảng 0.83 nếu người dùng chỉ đánh giá một cuốn sách điều nay cho chúng ta thấy hiệu suất ở lần dự đoán này đã tốt hơn so với lần trước nhưng vẫn là thấp.

```
[('0002005018', 5), ('0060973129', 3), ('0061076031', 5)]
Dự đoán với độ tương đồng cosine
--------------------------
0002005018    1.961019
0061076031    1.740911
0060973129    0.993866
0679425608    0.898729
0195153448    0.799883
dtype: float64
RMSE: 0.6203587924751143
MSE: 0.5352728218358906
------------------------------
```

Với mỗi rating được dự đoán sẽ sai lệch khoảng 0.62 nếu người dùng chỉ đánh giá một cuốn sách điều nay cho chúng ta thấy hiệu suất ở lần dự đoán này tốt hơn hẳn so với lần đầu tiên.

Qua 3 input thì chúng ta có thể đưa ra nhận xét như sau: Người dùng thực hiện đánh giá càng nhiều thì độ sai lệch với mỗi rating sẽ giảm xuống.

```
[('0002005018', 5), ('0060973129', 3), ('0061076031', 5), ('1567407781', 5), ('1881320189', 5)]
Dự đoán với độ tương đồng cosine
--------------------------
1881320189    2.737386
0002005018    1.973889
0060973129    1.712031
1567407781    1.660636
0609804618    1.033104
dtype: float64
RMSE: 0.40153432112782567
MSE: 0.32864302006606333
```

2. Với độ tương đồng PCC (Pearson correclation coefficient

```
[('0002005018', 5)]
Dự đoán với độ tương đồng cosine
---------------------------
0002005018    2.500000
0452264464    0.824512
0679425608    0.802090
0060973129    0.797202
1881320189    0.759599
dtype: float64
RMSE: 0.9466964501285277
MSE: 0.9098900151643848
------------------------------
Dự đoán với độ tương đồng PCC(Pearson Correclation Coefficient)
---------------------
0002005018    2.500000
0452264464    0.824512
0679425608    0.802090
0060973129    0.797202
1881320189    0.759599
dtype: float64
RMSE: 0.9466964501285278
MSE: 0.9098900151643848
```

```
[('0002005018', 5), ('0061076031', 5)]
Dự đoán với độ tương đồng cosine
--------------------------
0061076031    1.801578
0002005018    1.801578
0679425608    0.736614
0195153448    0.603792
0671870432    0.585507
dtype: float64
RMSE: 0.6390959305401092
MSE: 0.5522133542803627
------------------------------
Dự đoán với độ tương đồng PCC(Pearson Correclation Coefficient)
--------------------
0002005018    1.801578
0061076031    1.801578
0679425608    0.736614
0195153448    0.603792
0671870432    0.585507
dtype: float64
RMSE: 0.6390959305401092
MSE: 0.5522133542803627
```

Sau khi thực hiện các dự đoán với cùng input như độ tương đồng cosine, ta thấy cả 2 độ tương đồng này có độ sai lệch tương tự nhau.

3. Với độ tương đồng SRC (Spearman Rank Coefficient))

```
Dự đoán với độ tương đồng PCC(Pearson Correclation Coefficient)
--------------------
0002005018    1.801578
0061076031    1.801578
0679425608    0.736614
0195153448    0.603792
0671870432    0.585507
dtype: float64
RMSE: 0.6390959305401092
MSE: 0.5522133542803627
Dự đoán với độ tương đồng SRC(Spearman Rank Coefficient)
--------------------
0002005018    1.787532
0061076031    1.787532
0195153448    0.876361
0679425608    0.855429
0689821166    0.774918
dtype: float64
RMSE: 0.6315683340132255
MSE: 0.5508606709935151
```

```
Dự đoán với độ tương đồng PCC(Pearson Correclation Coefficient)
--------------------
0061076031    1.781995
1567407781    1.733689
0002005018    1.054850
0684823802    0.641626
0374157065    0.564566
dtype: float64
RMSE: 0.504831208231473
MSE: 0.41959512358691936
Dự đoán với độ tương đồng SRC(Spearman Rank Coefficient)
--------------------
0061076031    1.817076
1567407781    1.714624
0002005018    0.972611
0609804618    0.843956
0684823802    0.712005
dtype: float64
RMSE: 0.49421807210189167
MSE: 0.4125748423366379
```

```
Dự đoán với độ tương đồng PCC(Pearson Correclation Coefficient)
--------------------
1881320189    2.542898
0002005018    1.814448
1567407781    1.711491
0060973129    1.212031
0061076031    1.087493
dtype: float64
RMSE: 0.4075818436719067
MSE: 0.3336448406108849
Dự đoán với độ tương đồng SRC(Spearman Rank Coefficient)
--------------------
1881320189    2.757455
1567407781    1.873067
0002005018    1.578839
0609804618    1.431042
0060973129    1.323137
dtype: float64
RMSE: 0.3780485427125643
MSE: 0.3157113815284893
```

Với cùng input như nhau ta thấy khi sử dụng độ tương đồng SRC thì với mỗi rating độ sai lệch thấp hơn so với sử dụng Cosine hoặc PCC

4. Với độ tương đồng KCC (Kendall Tau correlation coefficient)

```
Dự đoán với độ tương đồng SRC(Spearman Rank Coefficient)
---------------------
0002005018    1.787532
0061076031    1.787532
0195153448    0.876361
0679425608    0.855429
0689821166    0.774918
dtype: float64
RMSE: 0.6315683340132255
MSE: 0.5508606709935151
Dự đoán với độ tương đồng KCC(Kendall Tau correlation coefficient)
---------------------
0002005018    1.860724
0061076031    1.860724
0679425608    0.762170
0195153448    0.758330
0345402871    0.668047
dtype: float64
RMSE: 0.6330496024441216
MSE: 0.5551268665605337

Dự đoán với độ tương đồng SRC(Spearman Rank Coefficient)
---------------------
0061076031    1.817076
1567407781    1.714624
0002005018    0.972611
0609804618    0.843956
0684823802    0.712005
dtype: float64
RMSE: 0.49421807210189167
MSE: 0.4125748423366379
Dự đoán với độ tương đồng KCC(Kendall Tau correlation coefficient)
---------------------
0061076031    1.893860
1567407781    1.807467
0002005018    1.135055
0609804618    0.814230
0684823802    0.646749
dtype: float64
RMSE: 0.49712119654913767
MSE: 0.4179735709120508
```

```
Dự đoán với độ tương đồng SRC(Spearman Rank Coefficient)
---------------------
1881320189    2.757455
1567407781    1.873067
0002005018    1.578839
0609804618    1.431042
0060973129    1.323137
dtype: float64
RMSE: 0.3780485427125643
MSE: 0.3157113815284893
Dự đoán với độ tương đồng KCC(Kendall Tau correlation coefficient)
---------------------
1881320189    2.743395
1567407781    1.973146
0002005018    1.694421
0061076031    1.412208
0609804618    1.337916
dtype: float64
RMSE: 0.3834968219422942
MSE: 0.3214912887250032
```

Cũng với input như trên, ta thấy rằng độ đo KCC có sai lệch cao hơn so với SRC.

Kết luận: Khi demo với dataset gồm 30 cuốn sách và 30 người dùng qua các phép thử trên ta thấy rằng sử dụng độ sai lệch SRC (Spearman Rank Coefficient) cho ra các kết quả dự đoán có sai lệch thấp nhất kế tiếp là KCC và cuối cùng là Cosine và PCC. Cùng với đó RMSE và MAE sẽ thay đổi tùy vào số lượng record.