RecommendeRangers

ML

Problem: recommend games on steam to users

Two main approaches:

1. Collaborative filtering
2. Content filtering
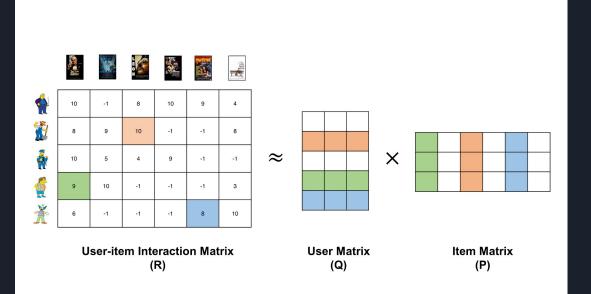
Starting point

- Collaborative filtering
- Python surprise module

Method

- Matrix factorization

User-item Interaction Matrix (R) ≈ **User Matrix (Q)** × **Item Matrix (P)**

Main methods

1. SVD (Singular value decomposition
2. SVD++
3. NMF (Non negative matrix factorization

Code

Load training data

```
In [ ]:  train_df = pd.read_csv('/kaggle/input/imp-data/train.csv')
         reader = Reader(rating_scale=(0, 1))
         train_data = Dataset.load_from_df(train_df[['steamid', 'appid', 'voted_up']], reader)
```

Code

Load training data

```
In [ ]:
train_df = pd.read_csv('/kaggle/input/imp-data/train.csv')
reader = Reader(rating_scale=(0, 1))
train_data = Dataset.load_from_df(train_df[['steamid', 'appid', 'voted_up']], reader)
```

Build a trainset and train the model

```
In [ ]:
trainset = train_data.build_full_trainset()
model = NMF(n_factors=20, n_epochs=600, verbose=True)
model.fit(trainset)
```

Code

Load training data

```
In [ ]:  train_df = pd.read_csv('/kaggle/input/imp-data/train.csv')
         reader = Reader(rating_scale=(0, 1))
         train_data = Dataset.load_from_df(train_df[['steamid', 'appid', 'voted_up']], reader)
```

Build a trainset and train the model

```
In [ ]:  trainset = train_data.build_full_trainset()
         model = NMF(n_factors=20, n_epochs=600, verbose=True)
         model.fit(trainset)
```

Load test data

```
In [ ]:  test_df = pd.read_csv('/kaggle/input/imp-data/test.csv')
```

Code

Load training data

In [ ]:
```python
train_df = pd.read_csv('/kaggle/input/imp-data/train.csv')
reader = Reader(rating_scale=(0, 1))
train_data = Dataset.load_from_df(train_df[['steamid', 'appid', 'voted_up']], reader)
```

Build a trainset and train the model

In [ ]:
```python
trainset = train_data.build_full_trainset()
model = NMF(n_factors=20, n_epochs=600, verbose=True)
model.fit(trainset)
```

Load test data

In [ ]:
```python
test_df = pd.read_csv('/kaggle/input/imp-data/test.csv')
```

Convert test data into a list of tuples for prediction

In [ ]:
```python
testset = list(zip(test_df['steamid'].values, test_df['appid'].values, [None]*len(test_df)))
# Make predictions
predictions = np.array([model.predict(uid, iid).est for uid, iid, _ in testset])
```

Tuning the model

- n_factors for scale of embeddings
- n_epochs for number of repetitions

Improvement

- Deep learning
- Content filtering
- Hybrid approach

Thanks