# CSE 258 - HW 3

Jin Dai / A92408103

## Task (Cook/Make prediction)

1
```
import gzip
import random
from collections import defaultdict
import csv
```

2
```
def read_gz(path):
    for l in gzip.open(path, 'rt'):
        yield eval(l)

def read_csv(path):
    f = gzip.open(path, 'rt')
    c = csv.reader(f)
    header = next(c)
    print(header)
    for l in c:
        yield l
```

3
```
dataset = list(read_csv("trainInteractions.csv.gz"))
dataset[:2]
```

```
['user_id', 'recipe_id', 'date', 'rating']
```

3
```
[['88348277', '03969194', '2004-12-23', '5'],
 ['86699739', '27096427', '2002-01-12', '4']]
```

4
```
# shuffle
random.shuffle(dataset)
```

5
```
train_size = 400000
data_train = dataset[:train_size]
data_valid = dataset[train_size:]
print('data_train size = %d\tdata_valid size = %d' % (len(data_train), len(data_valid)))
X_train = [d[:2] for d in data_train]
X_valid = [d[:2] for d in data_valid]
```

```
data_train size = 400000        data_valid size = 100000
```

### 1.

6
```
# compute user-recipes dict and all recipes set
user_recipes = defaultdict(set)
all_recipes = set()
for d in X_train:
    usr = d[0]
```

```
    r = d[1]
    all_recipes.add(r)
    user_recipes[usr].add(r)
```

7

```
# get a negative sample per each entry in the validation set
def random_sample(from_list, exclusions):
    s = random.choice(from_list)
    while s in exclusions:
        s = random.choice(from_list)
    return s

all_recipes_list = list(all_recipes)
not_made_samples = []
for d in X_valid:
    usr = d[0]
    r_cooked = d[1]
    r_uncooked = random_sample(all_recipes_list, user_recipes[usr].union({r_cooked}))
    not_made_samples.append([usr, r_uncooked])
```

8

```
# the baseline model
recipe_count = defaultdict(int)
total_cooked = 0

for d in X_train:
    r = d[1]
    recipe_count[r] += 1
    total_cooked += 1

most_popular = [(recipe_count[x], x) for x in recipe_count]
most_popular.sort()
most_popular.reverse()

def fit(most_popular, total_cooked, threshold=0.5):
    popular_set = set()
    count = 0
    for ic, i in most_popular:
      count += ic
      popular_set.add(i)
      if count > total_cooked * threshold: break
    return popular_set

return1 = fit(most_popular, total_cooked)
```

9

```
def predict_accuracy_1(made, not_made, popular_set):
    # count the true positives
    TP_count = 0
    for d in made:
        r = d[1]
        if r in popular_set:
            TP_count += 1
    # count the true negatives
    TN_count = 0
    for d in not_made:
        r = d[1]
        if r not in popular_set:
            TN_count += 1
    print('true positive = %d' % TP_count)
    print('true negative = %d' % TN_count)
    return (TP_count + TN_count) / (1.0 * (len(made) + len(not_made)))
```

```
10 accuracy = predict_accuracy_1(X_valid, not_made_samples, return1)
   print('given threshold=0.5, overall accuracy on the validation set + negative sample set is %.5f

   true positive = 44358
   true negative = 88156
   given threshold=0.5, overall accuracy on the validation set + negative sample set is 0.66257
```

## 2.

First we perform random sampling on all user-recipe pairs to get the "non-made" examples.

```
11 recipe_users = defaultdict(set)
   all_users = set()
   for d in X_train:
       usr = d[0]
       r = d[1]
       all_users.add(usr)
       recipe_users[r].add(usr)
```

```
12 def random_sample_2(all_users, all_recipes, exclusions):
       usr = random.choice(all_users)
       r = random.choice(all_recipes)
       s = (usr, r)
       while s in exclusions:
           usr = random.choice(all_users)
           r = random.choice(all_recipes)
           s = (usr, r)
       return s

   all_cooked_user_recipe_pair = set([(d[0], d[1]) for d in X_train + X_valid])
   all_users_list = list(all_users)

   not_made_samples = set()
   while len(not_made_samples) < len(X_valid):
       not_made_samples.add(random_sample_2(all_users_list, all_recipes_list, all_cooked_user_recip
   not_made_samples = list(not_made_samples)
```

For this question, we can try calculating accuracies using a few different thresholds like 10th, 20th,...,80th, 90th percentiles. We also want to compare against the accuracy using all the training data.

```
13 max_accuracy = 0
   best_threshold = 0.5
   for t in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]:
       return_n = fit(most_popular, total_cooked, t)
       accuracy = predict_accuracy_1(X_valid, not_made_samples, return_n)
       print('given threshold=%.2f, overall accuracy on the validation set + negative sample set is
       if accuracy > max_accuracy:
           max_accuracy = accuracy
           best_threshold = t

   true positive = 9556
   true negative = 99701
```

```
given threshold=0.10, overall accuracy on the validation set + negative sample set is 0.54629
true positive = 19256
true negative = 98722
given threshold=0.20, overall accuracy on the validation set + negative sample set is 0.58989
true positive = 28145
true negative = 96704
given threshold=0.30, overall accuracy on the validation set + negative sample set is 0.62425
true positive = 36550
true negative = 93285
given threshold=0.40, overall accuracy on the validation set + negative sample set is 0.64917
true positive = 44358
true negative = 88136
given threshold=0.50, overall accuracy on the validation set + negative sample set is 0.66247
true positive = 52090
true negative = 80996
given threshold=0.60, overall accuracy on the validation set + negative sample set is 0.66543
true positive = 59634
true negative = 70604
given threshold=0.70, overall accuracy on the validation set + negative sample set is 0.65119
true positive = 66955
true negative = 55844
given threshold=0.80, overall accuracy on the validation set + negative sample set is 0.61399
true positive = 74519
true negative = 29791
given threshold=0.90, overall accuracy on the validation set + negative sample set is 0.52155
true positive = 82225
true negative = 0
given threshold=1.00, overall accuracy on the validation set + negative sample set is 0.41113
```

```
14 print("When we set the threshold to the %dth percentile, the max accuracy is reached at %.5f." %
```

```
When we set the threshold to the 60th percentile, the max accuracy is reached at 0.66543.
```

## 3.

```
15 def Jaccard(s1, s2):
       numer = len(s1.intersection(s2))
       denom = len(s1.union(s2))
       if denom == 0:
           return 0
       return (1.0 * numer) / denom
```

```
44 max_sim_dict = defaultdict(float)

   def get_max_sim(usr, r):
       if (usr, r) in max_sim_dict:
           return max_sim_dict[(usr, r)]
       max_sim = 0.0
       for r2 in user_recipes[usr]:
           if r2 == r: continue
           max_sim = max(max_sim, Jaccard(recipe_users[r], recipe_users[r2]))
       max_sim_dict[(usr, r)] = max_sim
       return max_sim

   def predict_accuracy_3(made, not_made, sim_threshold):
       TP_count = 0
       for d in made:
           if get_max_sim(d[0], d[1]) > sim_threshold:
               TP_count += 1
```

```
    TN_count = 0
    for d in not_made:
        if get_max_sim(d[0], d[1]) <= sim_threshold:
            TN_count += 1
    print('true positive = %d' % TP_count)
    print('true negative = %d' % TN_count)
    print(len(max_sim_dict))
    return (TP_count + TN_count) / (1.0 * (len(made) + len(not_made)))
```

```
accuracy = predict_accuracy_3(X_valid, not_made_samples, 0.5)
print('given threshold=0.5, overall accuracy on the validation set + negative sample set is %.5f
```

```
true positive = 22
true negative = 100000
200000
given threshold=0.5, overall accuracy on the validation set + negative sample set is 0.50011
```

```
import statistics
# get the median of all Jaccard similarities in max_sim_dict
max_sims = [v for _, v in max_sim_dict.items()]
max_sims.sort()
statistics.median(max_sims)
```

```
0.0
```

```
len(max_sims)
```

```
200000
```

```
for t in [0, 0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.075, 0.1, 0.15, 0.2]:
    accuracy = predict_accuracy_3(X_valid, not_made_samples, t)
    print('given threshold=%.3f, overall accuracy on the validation set + negative sample set is
    if accuracy > max_accuracy:
        max_accuracy = accuracy
        best_threshold = t
```

```
true positive = 64885
true negative = 78243
200000
given threshold=0.000, overall accuracy on the validation set + negative sample set is 0.71564
true positive = 64885
true negative = 78243
200000
given threshold=0.001, overall accuracy on the validation set + negative sample set is 0.71564
true positive = 63653
true negative = 80820
200000
given threshold=0.005, overall accuracy on the validation set + negative sample set is 0.72237
true positive = 61524
true negative = 83429
200000
given threshold=0.010, overall accuracy on the validation set + negative sample set is 0.72476
true positive = 55932
true negative = 87203
200000
given threshold=0.020, overall accuracy on the validation set + negative sample set is 0.71567
true positive = 48531
true negative = 89424
200000
given threshold=0.030, overall accuracy on the validation set + negative sample set is 0.68978
```

```
true positive = 40860
true negative = 91170
200000
given threshold=0.040, overall accuracy on the validation set + negative sample set is 0.66015
true positive = 35537
true negative = 92389
200000
given threshold=0.050, overall accuracy on the validation set + negative sample set is 0.63963
true positive = 28619
true negative = 94367
200000
given threshold=0.075, overall accuracy on the validation set + negative sample set is 0.61493
true positive = 23078
true negative = 96058
200000
given threshold=0.100, overall accuracy on the validation set + negative sample set is 0.59568
true positive = 16454
true negative = 97579
200000
given threshold=0.150, overall accuracy on the validation set + negative sample set is 0.57017
true positive = 9788
true negative = 98725
200000
given threshold=0.200, overall accuracy on the validation set + negative sample set is 0.54256
```

69
```
print("When we set the similarity threshold to the %dth percentile, the max accuracy is reached
```

```
When we set the similarity threshold to the 1th percentile, the max accuracy is reached at 0.724
```

## 4

83
```
# define the predict function
def single_predict(user, recipe, popular_set, sim_threshold):
    if recipe in popular_set or get_max_sim(user, recipe) > sim_threshold:
        return 1
    return 0

def predict(X, popular_set, sim_threshold):
    return [single_predict(d[0], d[1], popular_set, sim_threshold) for d in X]
```

84
```
def predict_accuracy_4(made, not_made, popular_set, sim_threshold=0.01):
    pred_made = predict(made, popular_set, sim_threshold)
    TP_count = pred_made.count(1)
    pred_not_made = predict(not_made, popular_set, sim_threshold)
    TN_count = pred_not_made.count(0)
    print('true positive = %d' % TP_count)
    print('true negative = %d' % TN_count)
    return (TP_count + TN_count) / (1.0 * (len(made) + len(not_made)))
```

93
```
max_accuracy = 0
best_t1 = 0.6
best_t2 = 0.01
for t1 in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]:
    pop_set_perc_t1 = fit(most_popular, total_cooked, t1)
    for t2 in [0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75]:
        accuracy = predict_accuracy_4(X_valid, not_made_samples, pop_set_perc_t1, t2)
        print('Using popularity threshold=%.3f and similarity threshold=%.3f, we can achieve ove
```

```
        if accuracy > max_accuracy:
            max_accuracy = accuracy
            best_t1 = t1
            best_t2 = t2

true positive = 61972
true negative = 83363
Using popularity threshold=0.100 and similarity threshold=0.010, we can achieve overall accuracy
true positive = 55482
true negative = 88321
Using popularity threshold=0.100 and similarity threshold=0.025, we can achieve overall accuracy
true positive = 44464
true negative = 92093
Using popularity threshold=0.100 and similarity threshold=0.050, we can achieve overall accuracy
true positive = 38139
true negative = 94068
Using popularity threshold=0.100 and similarity threshold=0.075, we can achieve overall accuracy
true positive = 32614
true negative = 95759
Using popularity threshold=0.100 and similarity threshold=0.100, we can achieve overall accuracy
true positive = 15397
true negative = 98953
Using popularity threshold=0.100 and similarity threshold=0.250, we can achieve overall accuracy
true positive = 9578
true negative = 99701
Using popularity threshold=0.100 and similarity threshold=0.500, we can achieve overall accuracy
true positive = 9556
true negative = 99701
Using popularity threshold=0.100 and similarity threshold=0.750, we can achieve overall accuracy
true positive = 62756
true negative = 83019
Using popularity threshold=0.200 and similarity threshold=0.010, we can achieve overall accuracy
true positive = 58586
true negative = 87633
Using popularity threshold=0.200 and similarity threshold=0.025, we can achieve overall accuracy
true positive = 52542
true negative = 91153
Using popularity threshold=0.200 and similarity threshold=0.050, we can achieve overall accuracy
true positive = 47413
true negative = 93097
Using popularity threshold=0.200 and similarity threshold=0.075, we can achieve overall accuracy
true positive = 42145
true negative = 94780
Using popularity threshold=0.200 and similarity threshold=0.100, we can achieve overall accuracy
true positive = 25097
true negative = 97974
Using popularity threshold=0.200 and similarity threshold=0.250, we can achieve overall accuracy
true positive = 19278
true negative = 98722
Using popularity threshold=0.200 and similarity threshold=0.500, we can achieve overall accuracy
true positive = 19256
true negative = 98722
Using popularity threshold=0.200 and similarity threshold=0.750, we can achieve overall accuracy
true positive = 64035
true negative = 82034
Using popularity threshold=0.300 and similarity threshold=0.010, we can achieve overall accuracy
true positive = 60912
true negative = 86294
Using popularity threshold=0.300 and similarity threshold=0.025, we can achieve overall accuracy
true positive = 57313
true negative = 89392
Using popularity threshold=0.300 and similarity threshold=0.050, we can achieve overall accuracy
true positive = 54415
true negative = 91121
```

```
Using popularity threshold=0.300 and similarity threshold=0.075, we can achieve overall accuracy=
true positive = 50052
true negative = 92769
Using popularity threshold=0.300 and similarity threshold=0.100, we can achieve overall accuracy=
true positive = 33916
true negative = 95956
Using popularity threshold=0.300 and similarity threshold=0.250, we can achieve overall accuracy=
true positive = 28167
true negative = 96704
Using popularity threshold=0.300 and similarity threshold=0.500, we can achieve overall accuracy=
true positive = 28145
true negative = 96704
Using popularity threshold=0.300 and similarity threshold=0.750, we can achieve overall accuracy=
true positive = 65605
true negative = 79973
Using popularity threshold=0.400 and similarity threshold=0.010, we can achieve overall accuracy=
true positive = 63224
true negative = 83833
Using popularity threshold=0.400 and similarity threshold=0.025, we can achieve overall accuracy=
true positive = 60703
true negative = 86536
Using popularity threshold=0.400 and similarity threshold=0.050, we can achieve overall accuracy=
true positive = 58889
true negative = 88019
Using popularity threshold=0.400 and similarity threshold=0.075, we can achieve overall accuracy=
true positive = 56414
true negative = 89422
Using popularity threshold=0.400 and similarity threshold=0.100, we can achieve overall accuracy=
true positive = 42238
true negative = 92538
Using popularity threshold=0.400 and similarity threshold=0.250, we can achieve overall accuracy=
true positive = 36572
true negative = 93285
Using popularity threshold=0.400 and similarity threshold=0.500, we can achieve overall accuracy=
true positive = 36550
true negative = 93285
Using popularity threshold=0.400 and similarity threshold=0.750, we can achieve overall accuracy=
true positive = 67343
true negative = 76525
Using popularity threshold=0.500 and similarity threshold=0.010, we can achieve overall accuracy=
true positive = 65503
true negative = 79869
Using popularity threshold=0.500 and similarity threshold=0.025, we can achieve overall accuracy=
true positive = 63623
true negative = 82167
Using popularity threshold=0.500 and similarity threshold=0.050, we can achieve overall accuracy=
true positive = 62344
true negative = 83410
Using popularity threshold=0.500 and similarity threshold=0.075, we can achieve overall accuracy=
true positive = 60711
true negative = 84556
Using popularity threshold=0.500 and similarity threshold=0.100, we can achieve overall accuracy=
true positive = 49812
true negative = 87392
Using popularity threshold=0.500 and similarity threshold=0.250, we can achieve overall accuracy=
true positive = 44380
true negative = 88136
Using popularity threshold=0.500 and similarity threshold=0.500, we can achieve overall accuracy=
true positive = 44358
true negative = 88136
Using popularity threshold=0.500 and similarity threshold=0.750, we can achieve overall accuracy=
true positive = 69319
true negative = 71257
Using popularity threshold=0.600 and similarity threshold=0.010, we can achieve overall accuracy=
```

```
true positive = 67930
true negative = 74033
Using popularity threshold=0.600 and similarity threshold=0.025, we can achieve overall accuracy=
true positive = 66587
true negative = 75915
Using popularity threshold=0.600 and similarity threshold=0.050, we can achieve overall accuracy=
true positive = 65668
true negative = 76944
Using popularity threshold=0.600 and similarity threshold=0.075, we can achieve overall accuracy=
true positive = 64567
true negative = 77867
Using popularity threshold=0.600 and similarity threshold=0.100, we can achieve overall accuracy=
true positive = 57178
true negative = 80257
Using popularity threshold=0.600 and similarity threshold=0.250, we can achieve overall accuracy=
true positive = 52111
true negative = 80996
Using popularity threshold=0.600 and similarity threshold=0.500, we can achieve overall accuracy=
true positive = 52090
true negative = 80996
Using popularity threshold=0.600 and similarity threshold=0.750, we can achieve overall accuracy=
true positive = 71645
true negative = 63019
Using popularity threshold=0.700 and similarity threshold=0.010, we can achieve overall accuracy=
true positive = 70660
true negative = 65160
Using popularity threshold=0.700 and similarity threshold=0.025, we can achieve overall accuracy=
true positive = 69750
true negative = 66577
Using popularity threshold=0.700 and similarity threshold=0.050, we can achieve overall accuracy=
true positive = 69125
true negative = 67371
Using popularity threshold=0.700 and similarity threshold=0.075, we can achieve overall accuracy=
true positive = 68403
true negative = 68076
Using popularity threshold=0.700 and similarity threshold=0.100, we can achieve overall accuracy=
true positive = 64262
true negative = 69881
Using popularity threshold=0.700 and similarity threshold=0.250, we can achieve overall accuracy=
true positive = 59655
true negative = 70604
Using popularity threshold=0.700 and similarity threshold=0.500, we can achieve overall accuracy=
true positive = 59634
true negative = 70604
Using popularity threshold=0.700 and similarity threshold=0.750, we can achieve overall accuracy=
true positive = 74418
true negative = 50555
Using popularity threshold=0.800 and similarity threshold=0.010, we can achieve overall accuracy=
true positive = 73785
true negative = 52033
Using popularity threshold=0.800 and similarity threshold=0.025, we can achieve overall accuracy=
true positive = 73247
true negative = 53033
Using popularity threshold=0.800 and similarity threshold=0.050, we can achieve overall accuracy=
true positive = 72866
true negative = 53574
Using popularity threshold=0.800 and similarity threshold=0.075, we can achieve overall accuracy=
true positive = 72427
true negative = 54040
Using popularity threshold=0.800 and similarity threshold=0.100, we can achieve overall accuracy=
true positive = 70110
true negative = 55235
Using popularity threshold=0.800 and similarity threshold=0.250, we can achieve overall accuracy=
true positive = 66957
```

```
true negative = 55844
Using popularity threshold=0.800 and similarity threshold=0.500, we can achieve overall accuracy
true positive = 66955
true negative = 55844
Using popularity threshold=0.800 and similarity threshold=0.750, we can achieve overall accuracy
true positive = 78084
true negative = 27078
Using popularity threshold=0.900 and similarity threshold=0.010, we can achieve overall accuracy
true positive = 77779
true negative = 27833
Using popularity threshold=0.900 and similarity threshold=0.025, we can achieve overall accuracy
true positive = 77513
true negative = 28336
Using popularity threshold=0.900 and similarity threshold=0.050, we can achieve overall accuracy
true positive = 77340
true negative = 28626
Using popularity threshold=0.900 and similarity threshold=0.075, we can achieve overall accuracy
true positive = 77128
true negative = 28857
Using popularity threshold=0.900 and similarity threshold=0.100, we can achieve overall accuracy
true positive = 76069
true negative = 29461
Using popularity threshold=0.900 and similarity threshold=0.250, we can achieve overall accuracy
true positive = 74519
true negative = 29791
Using popularity threshold=0.900 and similarity threshold=0.500, we can achieve overall accuracy
true positive = 74519
true negative = 29791
Using popularity threshold=0.900 and similarity threshold=0.750, we can achieve overall accuracy
```

```
94 print("When we set the popularity threshold to the %.2fth percentile and similarity threshold to
```

```
When we set the popularity threshold to the 40.00th percentile and similarity threshold to the 5
the max accuracy is reached at 0.73620.
```

```
100 max_accuracy = 0
    best_t1 = 0.6
    best_t2 = 0.01
    for t1 in [0.3, 0.31, 0.32, 0.33, 0.34, 0.35, 0.36, 0.37, 0.38, 0.39, 0.4]:
        pop_set_perc_t1 = fit(most_popular, total_cooked, t1)
        for t2 in [0.03, 0.035, 0.036, 0.037, 0.038, 0.039, 0.04, 0.041, 0.042, 0.043, 0.044, 0.045]
            accuracy = predict_accuracy_4(X_valid, not_made_samples, pop_set_perc_t1, t2)
            print('Using popularity threshold=%.3f and similarity threshold=%.3f, we can achieve ove
            if accuracy > max_accuracy:
                max_accuracy = accuracy
                best_t1 = t1
                best_t2 = t2
```

```
true positive = 60205
true negative = 86991
Using popularity threshold=0.300 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 59443
true negative = 87752
Using popularity threshold=0.300 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 59266
true negative = 87927
Using popularity threshold=0.300 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 59257
true negative = 87932
Using popularity threshold=0.300 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 59069
```

```
true negative = 88099
Using popularity threshold=0.300 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 58866
true negative = 88239
Using popularity threshold=0.300 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 58657
true negative = 88394
Using popularity threshold=0.300 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 58656
true negative = 88399
Using popularity threshold=0.300 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 58437
true negative = 88585
Using popularity threshold=0.300 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 58431
true negative = 88587
Using popularity threshold=0.300 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 58191
true negative = 88780
Using popularity threshold=0.300 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 58186
true negative = 88789
Using popularity threshold=0.300 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 60472
true negative = 86776
Using popularity threshold=0.310 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 59730
true negative = 87525
Using popularity threshold=0.310 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 59560
true negative = 87697
Using popularity threshold=0.310 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 59552
true negative = 87701
Using popularity threshold=0.310 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 59371
true negative = 87865
Using popularity threshold=0.310 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 59181
true negative = 88003
Using popularity threshold=0.310 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 58984
true negative = 88155
Using popularity threshold=0.310 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 58983
true negative = 88160
Using popularity threshold=0.310 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 58772
true negative = 88343
Using popularity threshold=0.310 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 58766
true negative = 88345
Using popularity threshold=0.310 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 58535
true negative = 88534
Using popularity threshold=0.310 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 58530
true negative = 88543
Using popularity threshold=0.310 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 60730
true negative = 86575
Using popularity threshold=0.320 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 60014
true negative = 87314
```

```
Using popularity threshold=0.320 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 59848
true negative = 87484
Using popularity threshold=0.320 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 59840
true negative = 87488
Using popularity threshold=0.320 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 59668
true negative = 87649
Using popularity threshold=0.320 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 59489
true negative = 87785
Using popularity threshold=0.320 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 59299
true negative = 87935
Using popularity threshold=0.320 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 59298
true negative = 87940
Using popularity threshold=0.320 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 59096
true negative = 88120
Using popularity threshold=0.320 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 59090
true negative = 88122
Using popularity threshold=0.320 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 58869
true negative = 88308
Using popularity threshold=0.320 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 58864
true negative = 88317
Using popularity threshold=0.320 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 60974
true negative = 86327
Using popularity threshold=0.330 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 60282
true negative = 87061
Using popularity threshold=0.330 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 60120
true negative = 87229
Using popularity threshold=0.330 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 60113
true negative = 87233
Using popularity threshold=0.330 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 59948
true negative = 87392
Using popularity threshold=0.330 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 59778
true negative = 87527
Using popularity threshold=0.330 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 59594
true negative = 87675
Using popularity threshold=0.330 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 59593
true negative = 87679
Using popularity threshold=0.330 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 59397
true negative = 87856
Using popularity threshold=0.330 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 59391
true negative = 87858
Using popularity threshold=0.330 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 59178
true negative = 88044
Using popularity threshold=0.330 and similarity threshold=0.044, we can achieve overall accuracy
```

```
true positive = 59173
true negative = 88052
Using popularity threshold=0.330 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 61240
true negative = 86075
Using popularity threshold=0.340 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 60580
true negative = 86796
Using popularity threshold=0.340 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 60428
true negative = 86958
Using popularity threshold=0.340 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 60423
true negative = 86961
Using popularity threshold=0.340 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 60265
true negative = 87119
Using popularity threshold=0.340 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 60102
true negative = 87251
Using popularity threshold=0.340 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 59921
true negative = 87395
Using popularity threshold=0.340 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 59920
true negative = 87398
Using popularity threshold=0.340 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 59729
true negative = 87569
Using popularity threshold=0.340 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 59723
true negative = 87570
Using popularity threshold=0.340 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 59517
true negative = 87754
Using popularity threshold=0.340 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 59515
true negative = 87762
Using popularity threshold=0.340 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 61461
true negative = 85826
Using popularity threshold=0.350 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 60816
true negative = 86537
Using popularity threshold=0.350 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 60670
true negative = 86696
Using popularity threshold=0.350 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 60666
true negative = 86698
Using popularity threshold=0.350 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 60514
true negative = 86855
Using popularity threshold=0.350 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 60359
true negative = 86984
Using popularity threshold=0.350 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 60181
true negative = 87127
Using popularity threshold=0.350 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 60180
true negative = 87130
Using popularity threshold=0.350 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 59994
```

```
true negative = 87301
Using popularity threshold=0.350 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 59988
true negative = 87302
Using popularity threshold=0.350 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 59786
true negative = 87485
Using popularity threshold=0.350 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 59785
true negative = 87493
Using popularity threshold=0.350 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 61703
true negative = 85563
Using popularity threshold=0.360 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 61076
true negative = 86266
Using popularity threshold=0.360 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 60936
true negative = 86424
Using popularity threshold=0.360 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 60932
true negative = 86426
Using popularity threshold=0.360 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 60789
true negative = 86580
Using popularity threshold=0.360 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 60636
true negative = 86706
Using popularity threshold=0.360 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 60465
true negative = 86846
Using popularity threshold=0.360 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 60464
true negative = 86849
Using popularity threshold=0.360 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 60282
true negative = 87017
Using popularity threshold=0.360 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 60276
true negative = 87018
Using popularity threshold=0.360 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 60079
true negative = 87198
Using popularity threshold=0.360 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 60078
true negative = 87205
Using popularity threshold=0.360 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 61985
true negative = 85289
Using popularity threshold=0.370 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 61375
true negative = 85988
Using popularity threshold=0.370 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 61238
true negative = 86145
Using popularity threshold=0.370 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 61234
true negative = 86147
Using popularity threshold=0.370 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 61098
true negative = 86298
Using popularity threshold=0.370 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 60952
true negative = 86424
```

```
Using popularity threshold=0.370 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 60785
true negative = 86562
Using popularity threshold=0.370 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 60784
true negative = 86565
Using popularity threshold=0.370 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 60609
true negative = 86733
Using popularity threshold=0.370 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 60603
true negative = 86733
Using popularity threshold=0.370 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 60412
true negative = 86910
Using popularity threshold=0.370 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 60412
true negative = 86917
Using popularity threshold=0.370 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 62214
true negative = 85050
Using popularity threshold=0.380 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 61616
true negative = 85744
Using popularity threshold=0.380 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 61485
true negative = 85899
Using popularity threshold=0.380 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 61481
true negative = 85901
Using popularity threshold=0.380 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 61351
true negative = 86052
Using popularity threshold=0.380 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 61210
true negative = 86173
Using popularity threshold=0.380 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 61047
true negative = 86308
Using popularity threshold=0.380 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 61046
true negative = 86311
Using popularity threshold=0.380 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 60882
true negative = 86476
Using popularity threshold=0.380 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 60876
true negative = 86476
Using popularity threshold=0.380 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 60687
true negative = 86651
Using popularity threshold=0.380 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 60687
true negative = 86658
Using popularity threshold=0.380 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 62487
true negative = 84751
Using popularity threshold=0.390 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 61906
true negative = 85436
Using popularity threshold=0.390 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 61780
true negative = 85591
Using popularity threshold=0.390 and similarity threshold=0.036, we can achieve overall accuracy
```

```
true positive = 61777
true negative = 85593
Using popularity threshold=0.390 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 61649
true negative = 85743
Using popularity threshold=0.390 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 61515
true negative = 85862
Using popularity threshold=0.390 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 61358
true negative = 85996
Using popularity threshold=0.390 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 61357
true negative = 85999
Using popularity threshold=0.390 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 61201
true negative = 86162
Using popularity threshold=0.390 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 61195
true negative = 86162
Using popularity threshold=0.390 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 61010
true negative = 86333
Using popularity threshold=0.390 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 61010
true negative = 86340
Using popularity threshold=0.390 and similarity threshold=0.045, we can achieve overall accuracy
true positive = 62725
true negative = 84446
Using popularity threshold=0.400 and similarity threshold=0.030, we can achieve overall accuracy
true positive = 62158
true negative = 85122
Using popularity threshold=0.400 and similarity threshold=0.035, we can achieve overall accuracy
true positive = 62036
true negative = 85276
Using popularity threshold=0.400 and similarity threshold=0.036, we can achieve overall accuracy
true positive = 62033
true negative = 85278
Using popularity threshold=0.400 and similarity threshold=0.037, we can achieve overall accuracy
true positive = 61912
true negative = 85425
Using popularity threshold=0.400 and similarity threshold=0.038, we can achieve overall accuracy
true positive = 61780
true negative = 85543
Using popularity threshold=0.400 and similarity threshold=0.039, we can achieve overall accuracy
true positive = 61625
true negative = 85675
Using popularity threshold=0.400 and similarity threshold=0.040, we can achieve overall accuracy
true positive = 61624
true negative = 85678
Using popularity threshold=0.400 and similarity threshold=0.041, we can achieve overall accuracy
true positive = 61474
true negative = 85840
Using popularity threshold=0.400 and similarity threshold=0.042, we can achieve overall accuracy
true positive = 61469
true negative = 85840
Using popularity threshold=0.400 and similarity threshold=0.043, we can achieve overall accuracy
true positive = 61290
true negative = 86009
Using popularity threshold=0.400 and similarity threshold=0.044, we can achieve overall accuracy
true positive = 61290
true negative = 86016
Using popularity threshold=0.400 and similarity threshold=0.045, we can achieve overall accuracy
```

```
101 print("When we set the popularity threshold to the %.2fth percentile and similarity threshold to
```

```
When we set the popularity threshold to the 38.00th percentile and similarity threshold to the 3
the max accuracy is reached at 0.73701.
```

5

```
102 best_pop_set = fit(most_popular, total_cooked, best_t1)
    predictions = open("predictions_Made.txt", 'w')
    for l in open("stub_Made.txt"):
      if l.startswith("user_id"):
        #header
        predictions.write(l)
        continue
      u,i = l.strip().split('-')
      if single_predict(u, i, best_pop_set, best_t2):
        predictions.write(u + '-' + i + ",1\n")
      else:
        predictions.write(u + '-' + i + ",0\n")

    predictions.close()
```

Kaggle Username: Jin Dai