Recommendations_with_IBM

November 3, 2019

1 Recommendations with IBM

" In this notebook, you will be putting your recommendation skills to use on real data from the IBM Watson Studio platform.

You may either submit your notebook through the workspace here, or you may work from your local machine and submit through the next page. Either way assure that your code passes the project RUBRIC. **Please save regularly.**

By following the table of contents, you will build out a number of different methods for making recommendations that can be used for different situations.

1.1 Table of Contents

At the end of the notebook, you will find directions for how to submit your work. Let's get started by importing the necessary libraries and reading in the data.

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import project_tests as t
import pickle

%matplotlib inline

df = pd.read_csv('data/user-item-interactions.csv')
df_content = pd.read_csv('data/articles_community.csv')
del df['Unnamed: 0']
del df_content['Unnamed: 0']

# Show df to get an idea of the data
df.head()
```

[1]:
```
   article_id                                             title  \
0      1430.0  using pixiedust for fast, flexible, and easier...
1      1314.0       healthcare python streaming application demo
2      1429.0         use deep learning for image classification
3      1338.0          ml optimization using cognitive assistant
4      1276.0          deploy your python model as a restful api
```

```
                                      email
0  ef5f11f77ba020cd36e1105a00ab868bbdbf7fe7
1  083cbdfa93c8444beaa4c5f5e0f5f9198e4f9e0b
2  b96a4f2e92d8572034b1e9b28f9ac673765cd074
3  06485706b34a5c9bf2a0ecdac41daf7e7654ceb7
4  f01220c46fc92c6e6b161b1849de11faacd7ccb2
```

[2]:
```
# Show df_content to get an idea of the data
df_content.head()
```

[2]:
```
                                            doc_body  \
0  Skip navigation Sign in SearchLoading...\r\n\r...
1  No Free Hunch Navigation * kaggle.com\r\n\r\n ...
2   * Login\r\n * Sign Up\r\n\r\n * Learning Pat...
3  DATALAYER: HIGH THROUGHPUT, LOW LATENCY AT SCA...
4  Skip navigation Sign in SearchLoading...\r\n\r...

                                     doc_description  \
0  Detect bad readings in real time using Python ...
1  See the forest, see the trees. Here lies the c...
2  Heres this weeks news in Data Science and Bi...
3  Learn how distributed DBs solve the problem of...
4  This video demonstrates the power of IBM DataS...

                                       doc_full_name doc_status  article_id
0  Detect Malfunctioning IoT Sensors with Streami...       Live           0
1  Communicating data science: A guide to present...       Live           1
2         This Week in Data Science (April 18, 2017)       Live           2
3  DataLayer Conference: Boost the performance of...       Live           3
4      Analyze NY Restaurant data using Spark in DSX       Live           4
```

1.1.1 Part I : Exploratory Data Analysis

Use the dictionary and cells below to provide some insight into the descriptive statistics of the data.

1. What is the distribution of how many articles a user interacts with in the dataset? Provide a visual and descriptive statistics to assist with giving a look at the number of times each user interacts with an article.

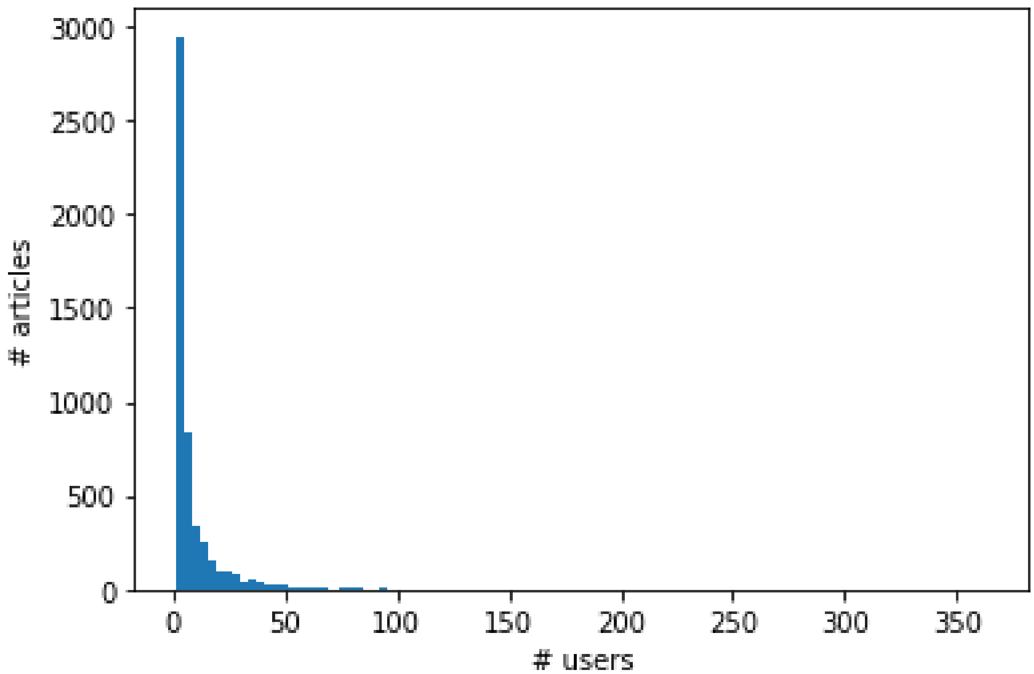
[3]:
```
# Descriptive statictics
df["email"].value_counts().describe()
```

[3]:
```
count    5148.000000
mean        8.930847
std        16.802267
min         1.000000
25%         1.000000
50%         3.000000
```

```
75%          9.000000
max        364.000000
Name: email, dtype: float64
```

[4]:
```
# Plot the number of times each user interacts with an article
plt.hist(df["email"].value_counts(), bins=100)
plt.ylabel("# articles")
plt.xlabel("# users")
```

[4]:
```
Text(0.5, 0, '# users')
```



[5]:
```
# Fill in the median and maximum number of user_article interactios below

median_val = 3 # 50% of individuals interact with ____ number of articles or␣
 ↪fewer.
max_views_by_user = 364 # The maximum number of user-article interactions by any␣
 ↪1 user is ______.
```

2. Explore and remove duplicate articles from the **df_content** dataframe.

[6]:
```
# Find and explore duplicate articles
df_content[df_content["article_id"].duplicated()]
```

[6]:
```
                                              doc_body  \
365  Follow Sign in / Sign up Home About Insight Da...
692  Homepage Follow Sign in / Sign up Homepage * H...
761  Homepage Follow Sign in Get started Homepage *...
970  This video shows you how to construct queries ...
```

```
971  Homepage Follow Sign in Get started * Home\r\n...

                                      doc_description  \
365  During the seven-week Insight Data Engineering...
692  One of the earliest documented catalogs was co...
761  Todays world of data science leverages data f...
970  This video shows you how to construct queries ...
971  If you are like most data scientists, you are ...

                                        doc_full_name doc_status  article_id
365                      Graph-based machine learning       Live          50
692  How smart catalogs can turn the big data flood...       Live         221
761  Using Apache Spark as a parallel processing fr...       Live         398
970                             Use the Primary Index       Live         577
971  Self-service data preparation with IBM Data Re...       Live         232
```

[7]:
```
# Remove any rows that have the same article_id - only keep the first
df_content.drop_duplicates(subset="article_id", inplace = True)
```

3. Use the cells below to find:
a. The number of unique articles that have an interaction with a user.
b. The number of unique articles in the dataset (whether they have any interactions or not). **c.** The number of unique users in the dataset. (excluding null values) **d.** The number of user-article interactions in the dataset.

[8]:
```
unique_articles = df["article_id"].nunique() # The number of unique articles␣
↪that have at least one interaction
unique_articles
```

[8]:
```
714
```

[9]:
```
total_articles = df_content.shape[0] # The number of unique articles on the IBM␣
↪platform
total_articles
```

[9]:
```
1051
```

[10]:
```
unique_users = df["email"].nunique()# The number of unique users
unique_users
```

[10]:
```
5148
```

[11]:
```
user_article_interactions = df.shape[0]# The number of user-article interactions
user_article_interactions
```

[11]:
```
45993
```

4. Use the cells below to find the most viewed **article_id**, as well as how often it was viewed. After talking to the company leaders, the `email_mapper` function was deemed a reasonable way to map users to ids. There were a small number of null values, and it was found that all of these null values likely belonged to a single user (which is how they are stored using the function below).

[12]:
```
df["article_id"].value_counts().head()
```

[12]:
```
1429.0    937
1330.0    927
1431.0    671
1427.0    643
1364.0    627
Name: article_id, dtype: int64
```

[13]:
```
most_viewed_article_id = "1429.0" # The most viewed article in the dataset as a␣
 ↪string with one value following the decimal
max_views = 937 # The most viewed article in the dataset was viewed how many␣
 ↪times?
```

[14]:
```
## No need to change the code here - this will be helpful for later parts of the␣
 ↪notebook
# Run this cell to map the user email to a user_id column and remove the email␣
 ↪column

def email_mapper():
    coded_dict = dict()
    cter = 1
    email_encoded = []

    for val in df['email']:
        if val not in coded_dict:
            coded_dict[val] = cter
            cter+=1

        email_encoded.append(coded_dict[val])
    return email_encoded

email_encoded = email_mapper()
del df['email']
df['user_id'] = email_encoded

# show header
df.head()
```

[14]:

	article_id	title	user_id
0	1430.0	using pixiedust for fast, flexible, and easier...	1
1	1314.0	healthcare python streaming application demo	2
2	1429.0	use deep learning for image classification	3
3	1338.0	ml optimization using cognitive assistant	4
4	1276.0	deploy your python model as a restful api	5

[15]:
```
## If you stored all your results in the variable names above,
## you shouldn't need to change anything in this cell

sol_1_dict = {
    '`50% of individuals have _____ or fewer interactions.`': median_val,
```

```
    '`The total number of user-article interactions in the dataset is ______.`':␣
↪user_article_interactions,
    '`The maximum number of user-article interactions by any 1 user is ______.`':
↪ max_views_by_user,
    '`The most viewed article in the dataset was viewed _____ times.`':␣
↪max_views,
    '`The article_id of the most viewed article is ______.`':␣
↪most_viewed_article_id,
    '`The number of unique articles that have at least 1 rating ______.`':␣
↪unique_articles,
    '`The number of unique users in the dataset is ______`': unique_users,
    '`The number of unique articles on the IBM platform`': total_articles
}

# Test your dictionary against the solution
t.sol_1_test(sol_1_dict)
```

```
It looks like you have everything right here! Nice job!
```

1.1.2 Part II: Rank-Based Recommendations

Unlike in the earlier lessons, we don't actually have ratings for whether a user liked an article or not. We only know that a user has interacted with an article. In these cases, the popularity of an article can really only be based on how often an article was interacted with.

1. Fill in the function below to return the **n** top articles ordered with most interactions as the top. Test your function using the tests below.

[16]:
```
def get_top_articles(n, df=df):
    '''
    INPUT:
    n - (int) the number of top articles to return
    df - (pandas dataframe) df as defined at the top of the notebook

    OUTPUT:
    top_articles - (list) A list of the top 'n' article titles

    '''
    top_ids = df["article_id"].value_counts().head(n).index.tolist()
    top_articles = df.set_index("article_id").loc[top_ids, :].
↪reset_index(inplace=False)
    top_articles = list(top_articles["title"].unique())
    return top_articles # Return the top article titles from df (not df_content)

def get_top_article_ids(n, df=df):
    '''
    INPUT:
    n - (int) the number of top articles to return
    df - (pandas dataframe) df as defined at the top of the notebook
```

```
    OUTPUT:
    top_articles - (list) A list of the top 'n' article titles

    '''
    # Your code here
    top_articles = df["article_id"].value_counts().head(n).index.astype(str).
 ↪tolist()


    return top_articles # Return the top article ids
```

[17]:
```
print(get_top_articles(10))
print(get_top_article_ids(10))
```

```
['use deep learning for image classification', 'insights from new york car
accident reports', 'visualize car data with brunel', 'use xgboost, scikit-learn
& ibm watson machine learning apis', 'predicting churn with the spss random tree
algorithm', 'healthcare python streaming application demo', 'finding optimal
locations of new store using decision optimization', 'apache spark lab, part 1:
basic concepts', 'analyze energy consumption in buildings', 'gosales
transactions for logistic regression model']
['1429.0', '1330.0', '1431.0', '1427.0', '1364.0', '1314.0', '1293.0', '1170.0',
'1162.0', '1304.0']
```

[18]:
```
# Test your function by returning the top 5, 10, and 20 articles
top_5 = get_top_articles(5)
top_10 = get_top_articles(10)
top_20 = get_top_articles(20)

# Test each of your three lists from above
t.sol_2_test(get_top_articles)
```

```
Your top_5 looks like the solution list! Nice job.
Your top_10 looks like the solution list! Nice job.
Your top_20 looks like the solution list! Nice job.
```

1.1.3 Part III: User-User Based Collaborative Filtering

1. Use the function below to reformat the **df** dataframe to be shaped with users as the rows and articles as the columns.

- Each **user** should only appear in each **row** once.
- Each **article** should only show up in one **column**.
- **If a user has interacted with an article, then place a 1 where the user-row meets for that article-column**. It does not matter how many times a user has interacted with the article, all entries where a user has interacted with an article should be a 1.

- **If a user has not interacted with an item, then place a zero where the user-row meets for that article-column**.

Use the tests to make sure the basic structure of your matrix matches what is expected by the solution.

[19]:
```
dummies = pd.get_dummies(df["article_id"])
user_item = pd.concat([df, dummies], axis=1)
user_item = user_item.drop([user_item.columns[0] , user_item.columns[1]], axis=1)
user_item = user_item.groupby("user_id",as_index=False).sum()
user_item = user_item.set_index("user_id")
```

[20]:
```
# create the user-article matrix with 1's and 0's

def create_user_item_matrix(df):
    '''
    INPUT:
    df - pandas dataframe with article_id, title, user_id columns

    OUTPUT:
    user_item - user item matrix

    Description:
    Return a matrix with user ids as rows and article ids on the columns with 1␣
 ↪values where a user interacted with
    an article and a 0 otherwise
    '''
    # Fill in the function here
    user_item = df.groupby(["user_id", "article_id"])["title"].max().unstack()
    user_item = user_item.notnull().astype(int)

    return user_item # return the user_item matrix

user_item = create_user_item_matrix(df)
user_item.head()
```

[20]:
```
article_id  0.0     2.0     4.0     8.0     9.0     12.0    14.0    15.0    \
user_id
1                0       0       0       0       0       0       0       0
2                0       0       0       0       0       0       0       0
3                0       0       0       0       0       1       0       0
4                0       0       0       0       0       0       0       0
5                0       0       0       0       0       0       0       0

article_id  16.0    18.0    ...  1434.0  1435.0  1436.0  1437.0  1439.0  \
user_id                     ...
1                0       0  ...       0       0       1       0       1
2                0       0  ...       0       0       0       0       0
3                0       0  ...       0       0       1       0       0
4                0       0  ...       0       0       0       0       0
```

```
5                0      0  ...      0      0      0      0      0

article_id  1440.0  1441.0  1442.0  1443.0  1444.0
user_id
1                0       0       0       0       0
2                0       0       0       0       0
3                0       0       0       0       0
4                0       0       0       0       0
5                0       0       0       0       0

[5 rows x 714 columns]
```

[21]:
```
## Tests: You should just need to run this cell.  Don't change the code.
assert user_item.shape[0] == 5149, "Oops!  The number of users in the␣
 ↪user-article matrix doesn't look right."
assert user_item.shape[1] == 714, "Oops!  The number of articles in the␣
 ↪user-article matrix doesn't look right."
assert user_item.sum(axis=1)[1] == 36, "Oops!  The number of articles seen by␣
 ↪user 1 doesn't look right."
print("You have passed our quick tests!  Please proceed!")
```

```
You have passed our quick tests!  Please proceed!
```

2. Complete the function below which should take a user_id and provide an ordered list of the most similar users to that user (from most similar to least similar). The returned result should not contain the provided user_id, as we know that each user is similar to him/herself. Because the results for each user here are binary, it (perhaps) makes sense to compute similarity as the dot product of two users.

Use the tests to test your function.

[22]:
```
def find_similar_users(user_id, user_item=user_item):
    '''
    INPUT:
    user_id - (int) a user_id
    user_item - (pandas dataframe) matrix of users by articles:
                1's when a user has interacted with an article, 0 otherwise

    OUTPUT:
    similar_users - (list) an ordered list where the closest users (largest dot␣
 ↪product users)
                    are listed first

    Description:
    Computes the similarity of every pair of users based on the dot product
    Returns an ordered

    '''
    # compute similarity of each user to the provided user
```

```
    similarity_vector = user_item[user_item.index == user_id].dot(np.
 ↪transpose(user_item))
    # sort by similarity
    similarity_vector_sorted = similarity_vector.sort_values(axis=1, by=user_id,␣
 ↪ascending=False)
    # create list of just the ids
    most_similar_users = similarity_vector_sorted.columns.tolist()
    # remove the own user's id
    most_similar_users.remove(user_id)
    return most_similar_users # return a list of the users in order from most to␣
 ↪least similar
```

[23]:
```
# Do a spot check of your function
print("The 10 most similar users to user 1 are: {}".
 ↪format(find_similar_users(1)[:10]))
print("The 5 most similar users to user 3933 are: {}".
 ↪format(find_similar_users(3933)[:5]))
print("The 3 most similar users to user 46 are: {}".
 ↪format(find_similar_users(46)[:3]))
```

```
The 10 most similar users to user 1 are: [3933, 23, 3782, 203, 4459, 3870, 131,
4201, 46, 5041]
The 5 most similar users to user 3933 are: [1, 23, 3782, 203, 4459]
The 3 most similar users to user 46 are: [4201, 3782, 23]
```

3. Now that you have a function that provides the most similar users to each user, you will want to use these users to find articles you can recommend. Complete the functions below to return the articles you would recommend to each user.

[24]:
```
def get_article_names(article_ids, df=df):
    '''
    INPUT:
    article_ids - (list) a list of article ids
    df - (pandas dataframe) df as defined at the top of the notebook

    OUTPUT:
    article_names - (list) a list of article names associated with the list of␣
 ↪article ids
                    (this is identified by the title column)
    '''
    # Your code here
    df_selected = df.loc[df["article_id"].isin(article_ids)]
    article_names = df_selected["title"].unique().tolist()

    return article_names # Return the article names associated with list of␣
 ↪article ids
```

```
def get_user_articles(user_id, user_item = user_item):
    '''
    INPUT:
    user_id - (int) a user id
    user_item - (pandas dataframe) matrix of users by articles:
                1's when a user has interacted with an article, 0 otherwise

    OUTPUT:
    article_ids - (list) a list of the article ids seen by the user
    article_names - (list) a list of article names associated with the list of␣
 ↪article ids
                    (this is identified by the doc_full_name column in␣
 ↪df_content)

    Description:
    Provides a list of the article_ids and article titles that have been seen by␣
 ↪a user
    '''
    # Your code here
    article_ids = user_item.loc[user_id]
    article_ids = article_ids[article_ids == 1].index.tolist()
    for item in range(len(article_ids)):
        article_ids[item] = str(article_ids[item])
    article_names = get_article_names(article_ids)
    return article_ids, article_names # return the ids and names


def user_user_recs(user_id, m=10):
    '''
    INPUT:
    user_id - (int) a user id
    m - (int) the number of recommendations you want for the user

    OUTPUT:
    recs - (list) a list of recommendations for the user

    Description:
    Loops through the users based on closeness to the input user_id
    For each user - finds articles the user hasn't seen before and provides them␣
 ↪as recs
    Does this until m recommendations are found

    Notes:
    Users who are the same closeness are chosen arbitrarily as the 'next' user

    For the user where the number of recommended articles starts below m
```

```
    and ends exceeding m, the last items are chosen arbitrarily

    '''
    # Your code here
    user_ids = find_similar_users(user_id)
    recs = df[df["user_id"].isin(user_ids)]["article_id"]
    recs = list(set(recs))
    recs = recs[:m]
    return recs # return your recommendations for this user_id
```

[25]:
```
# Check Results
get_article_names(user_user_recs(1, 10)) # Return 10 recommendations for user 1
```

[25]:
```
['got zip code data? prep it for analytics.  ibm watson data lab  medium',
 'timeseries data analysis of iot events by using jupyter notebook',
 'the greatest public datasets for ai  startup grind',
 '3992    using apache spark to predict attack vectors a...\nName: title, dtype:
object',
 'detect malfunctioning iot sensors with streaming analytics',
 'this week in data science (april 18, 2017)',
 'higher-order logistic regression for large datasets',
 'apache spark 2.0: extend structured streaming for spark ml',
 'data science bowl 2017',
 'analyze ny restaurant data using spark in dsx']
```

[26]:
```
# Test your functions here - No need to change this code - just run this cell
assert set(get_article_names(['1024.0', '1176.0', '1305.0', '1314.0', '1422.0',␣
 ↪'1427.0'])) == set(['using deep learning to reconstruct high-resolution␣
 ↪audio', 'build a python app on the streaming analytics service', 'gosales␣
 ↪transactions for naive bayes model', 'healthcare python streaming application␣
 ↪demo', 'use r dataframes & ibm watson natural language understanding', 'use␣
 ↪xgboost, scikit-learn & ibm watson machine learning apis']), "Oops! Your the␣
 ↪get_article_names function doesn't work quite how we expect."
assert set(get_article_names(['1320.0', '232.0', '844.0'])) == set(['housing␣
 ↪(2015): united states demographic measures','self-service data preparation␣
 ↪with ibm data refinery','use the cloudant-spark connector in python␣
 ↪notebook']), "Oops! Your the get_article_names function doesn't work quite how␣
 ↪we expect."
assert set(get_user_articles(20)[0]) == set(['1320.0', '232.0', '844.0'])
assert set(get_user_articles(20)[1]) == set(['housing (2015): united states␣
 ↪demographic measures', 'self-service data preparation with ibm data␣
 ↪refinery','use the cloudant-spark connector in python notebook'])
assert set(get_user_articles(2)[0]) == set(['1024.0', '1176.0', '1305.0', '1314.
 ↪0', '1422.0', '1427.0'])
```

```
assert set(get_user_articles(2)[1]) == set(['using deep learning to reconstruct␣
↪high-resolution audio', 'build a python app on the streaming analytics␣
↪service', 'gosales transactions for naive bayes model', 'healthcare python␣
↪streaming application demo', 'use r dataframes & ibm watson natural language␣
↪understanding', 'use xgboost, scikit-learn & ibm watson machine learning␣
↪apis'])
print("If this is all you see, you passed all of our tests!  Nice job!")
```

```
If this is all you see, you passed all of our tests!  Nice job!
```

4. Now we are going to improve the consistency of the **user_user_recs** function from above.

- Instead of arbitrarily choosing when we obtain users who are all the same closeness to a given user - choose the users that have the most total article interactions before choosing those with fewer article interactions.
- Instead of arbitrarily choosing articles from the user where the number of recommended articles starts below m and ends exceeding m, choose articles with the articles with the most total interactions before choosing those with fewer total interactions. This ranking should be what would be obtained from the **top_articles** function you wrote earlier.

[27]:
```
def get_top_sorted_users(user_id, df=df, user_item=user_item):
    '''
    INPUT:
    user_id - (int)
    df - (pandas dataframe) df as defined at the top of the notebook
    user_item - (pandas dataframe) matrix of users by articles:
            1's when a user has interacted with an article, 0 otherwise


    OUTPUT:
    neighbors_df - (pandas dataframe) a dataframe with:
                    neighbor_id - is a neighbor user_id
                    similarity - measure of the similarity of each user to the␣
↪provided user_id
                    num_interactions - the number of articles viewed by the user␣
↪- if a user has vieved an item

    Other Details - sort the neighbors_df by the similarity and then by number␣
↪of interactions where
                    highest of each is higher in the dataframe

    '''
    # Create empty dataframe
    neighbor_id = []
    similarity = []
    num_interactions = []
    neighbors_df = pd.DataFrame({
```

```python
                                "neighbor_id": neighbor_id,
                                "similarity": similarity,
                                "num_interactions": num_interactions
                                })
    #get user interactions
    user_interactions = df.groupby(['user_id'])['article_id'].count()


    for item in user_item.index:
        if item != user_id:
            neighbor_id = item
            # Uses similarty from the find_similar_users function above
            similarity = user_item[user_item.index == user_id].dot(np.
 ↪transpose(user_item.loc[item])).values[0]
            num_interactions = user_interactions.loc[item]
            neighbors_df.loc[neighbor_id] =␣
 ↪[neighbor_id,similarity,num_interactions]

    neighbors_df = neighbors_df.sort_values(by = ["similarity", "neighbor_id"],␣
 ↪ascending = [False, True])
    return neighbors_df # Return the dataframe specified in the doc_string


def user_user_recs_part2(user_id, m=10):
    '''
    INPUT:
    user_id - (int) a user id
    m - (int) the number of recommendations you want for the user

    OUTPUT:
    recs - (list) a list of recommendations for the user by article id
    rec_names - (list) a list of recommendations for the user by article title

    Description:
    Loops through the users based on closeness to the input user_id
    For each user - finds articles the user hasn't seen before and provides them␣
 ↪as recs
    Does this until m recommendations are found

    Notes:
    * Choose the users that have the most total article interactions
    before choosing those with fewer article interactions.

    * Choose articles with the articles with the most total interactions
    before choosing those with fewer total interactions.

    '''
```

```
    # Your code here
    neighbors_df = get_top_sorted_users(user_id)
    user_article_ids, user_article_names = get_user_articles(user_id)

    recs = []

    top_m_neighbor_ids = neighbors_df["neighbor_id"].head(m).tolist()
    for neighbor in top_m_neighbor_ids:
        neighbor_article_ids, neighbor_article_names = 
 ↪get_user_articles(neighbor)
        recs.extend(neighbor_article_ids)

    recs = list(set(recs) - set(user_article_ids))
    recs = list(set(recs[:m]))
    rec_names = get_article_names(recs)

    return recs, rec_names
```

[28]:
```
# Quick spot check - don't change this code - just use it to test your functions
rec_ids, rec_names = user_user_recs_part2(20, 10)
print("The top 10 recommendations for user 20 are the following article ids:")
print(rec_ids)
print()
print("The top 10 recommendations for user 20 are the following article names:")
print(rec_names)
```

```
The top 10 recommendations for user 20 are the following article ids:
['1164.0', '651.0', '1175.0', '1014.0', '1328.0', '686.0', '1409.0', '164.0',
'1433.0', '1436.0']

The top 10 recommendations for user 20 are the following article names:
['analyze open data sets with pandas dataframes', 'welcome to pixiedust',
'income (2015): united states demographic measures', 'learn tensorflow and deep
learning together and now!', 'analyzing streaming data from kafka topics', '1448
i ranked every intro to data science course on...\nName: title, dtype: object',
'score a predictive model built with ibm spss modeler, wml & dsx', 'uci: red
wine quality', 'visualize the 1854 london cholera outbreak', 'breast cancer
detection with xgboost, wml and scikit']
```

5. Use your functions from above to correctly fill in the solutions to the dictionary below. Then test your dictionary against the solution. Provide the code you need to answer each following the comments below.

[29]:
```
### Tests with a dictionary of results
user1_most_sim =  get_top_sorted_users(1).iloc[0].neighbor_id # Find the user
 ↪that is most similar to user 1
user131_10th_sim = get_top_sorted_users(131).iloc[9].neighbor_id # Find the 10th
 ↪most similar user to user 131
```

[30]:
```
## Dictionary Test Here
sol_5_dict = {
    'The user that is most similar to user 1.': user1_most_sim,
    'The user that is the 10th most similar to user 131': user131_10th_sim,
}

t.sol_5_test(sol_5_dict)
```

```
This all looks good!  Nice job!
```

6. If we were given a new user, which of the above functions would you be able to use to make recommendations? Explain. Can you think of a better way we might make recommendations? Use the cell below to explain a better method for new users.

Our first approach is a recommendation engine based on collaborative filtering, which requires having enough interactions between users and articles. A new user would not have any interactions with articles yet, so collaborative filtering would not be useful. A better way to make recommendations for new users would be recommending the most popular articles for them, until new users have enough interactions. With enough interaction data we can start using collaborative filtering for recommendations.

7. Using your existing functions, provide the top 10 recommended articles you would provide for the a new user below. You can test your function against our thoughts to make sure we are all on the same page with how we might make a recommendation.

[31]:
```
new_user = '0.0'

# What would your recommendations be for this new user '0.0'?  As a new user,␣
↪they have no observed articles.
# Provide a list of the top 10 article ids you would give to
new_user_recs = get_top_article_ids(10) # Your recommendations here
```

[32]:
```
assert set(new_user_recs) == set(['1314.0','1429.0','1293.0','1427.0','1162.
↪0','1364.0','1304.0','1170.0','1431.0','1330.0']), "Oops!  It makes sense that␣
↪in this case we would want to recommend the most popular articles, because we␣
↪don't know anything about these users."

print("That's right!  Nice job!")
```

```
That's right!  Nice job!
```

1.1.4 Part IV: Content Based Recommendations (EXTRA - NOT REQUIRED)

Another method we might use to make recommendations is to perform a ranking of the highest ranked articles associated with some term. You might consider content to be the **doc_body**, **doc_description**, or **doc_full_name**. There isn't one way to create a content based recommendation, especially considering that each of these columns hold content related information.

1. Use the function body below to create a content based recommender. Since there isn't one right answer for this recommendation tactic, no test functions are provided. Feel free to change the function inputs if you decide you want to try a method that requires more input values. The

input values are currently set with one idea in mind that you may use to make content based recommendations. One additional idea is that you might want to choose the most popular recommendations that meet your 'content criteria', but again, there is a lot of flexibility in how you might make these recommendations.

1.1.5 This part is NOT REQUIRED to pass this project. However, you may choose to take this on as an extra way to show off your skills.

[33]:
```
def make_content_recs():
    '''
    INPUT:

    OUTPUT:

    '''
```

2. Now that you have put together your content-based recommendation system, use the cell below to write a summary explaining how your content based recommender works. Do you see any possible improvements that could be made to your function? Is there anything novel about your content based recommender?

1.1.6 This part is NOT REQUIRED to pass this project. However, you may choose to take this on as an extra way to show off your skills.

Write an explanation of your content based recommendation system here.

3. Use your content-recommendation system to make recommendations for the below scenarios based on the comments. Again no tests are provided here, because there isn't one right answer that could be used to find these content based recommendations.

1.1.7 This part is NOT REQUIRED to pass this project. However, you may choose to take this on as an extra way to show off your skills.

[34]:
```
# make recommendations for a brand new user


# make a recommendations for a user who only has interacted with article id␣
↪'1427.0'
```

1.1.8 Part V: Matrix Factorization

In this part of the notebook, you will build use matrix factorization to make article recommendations to the users on the IBM Watson Studio platform.

1. You should have already created a **user_item** matrix above in **question 1** of **Part III** above. This first question here will just require that you run the cells to get things set up for the rest of **Part V** of the notebook.

[35]:
```
# Load the matrix here
user_item_matrix = pd.read_pickle('user_item_matrix.p')
#user_item_matrix = user_item
```

[36]:
```
# quick look at the matrix
user_item_matrix.head()
```

[36]:
```
article_id  0.0  100.0  1000.0  1004.0  1006.0  1008.0  101.0  1014.0  1015.0  \
user_id
1           0.0    0.0     0.0     0.0     0.0     0.0    0.0     0.0     0.0
2           0.0    0.0     0.0     0.0     0.0     0.0    0.0     0.0     0.0
3           0.0    0.0     0.0     0.0     0.0     0.0    0.0     0.0     0.0
4           0.0    0.0     0.0     0.0     0.0     0.0    0.0     0.0     0.0
5           0.0    0.0     0.0     0.0     0.0     0.0    0.0     0.0     0.0

article_id  1016.0  ...  977.0  98.0  981.0  984.0  985.0  986.0  990.0  \
user_id             ...
1              0.0  ...    0.0   0.0    1.0    0.0    0.0    0.0    0.0
2              0.0  ...    0.0   0.0    0.0    0.0    0.0    0.0    0.0
3              0.0  ...    1.0   0.0    0.0    0.0    0.0    0.0    0.0
4              0.0  ...    0.0   0.0    0.0    0.0    0.0    0.0    0.0
5              0.0  ...    0.0   0.0    0.0    0.0    0.0    0.0    0.0

article_id  993.0  996.0  997.0
user_id
1             0.0    0.0    0.0
2             0.0    0.0    0.0
3             0.0    0.0    0.0
4             0.0    0.0    0.0
5             0.0    0.0    0.0

[5 rows x 714 columns]
```

2. In this situation, you can use Singular Value Decomposition from numpy on the user-item matrix. Use the cell to perform SVD, and explain why this is different than in the lesson.

[37]:
```
# Perform SVD on the User-Item Matrix Here
u, s, vt = np.linalg.svd(user_item_matrix) # use the built in to get the three␣
↪matrices
```

[38]:
```
print(u.shape)
print(s.shape)
print(vt.shape)
```

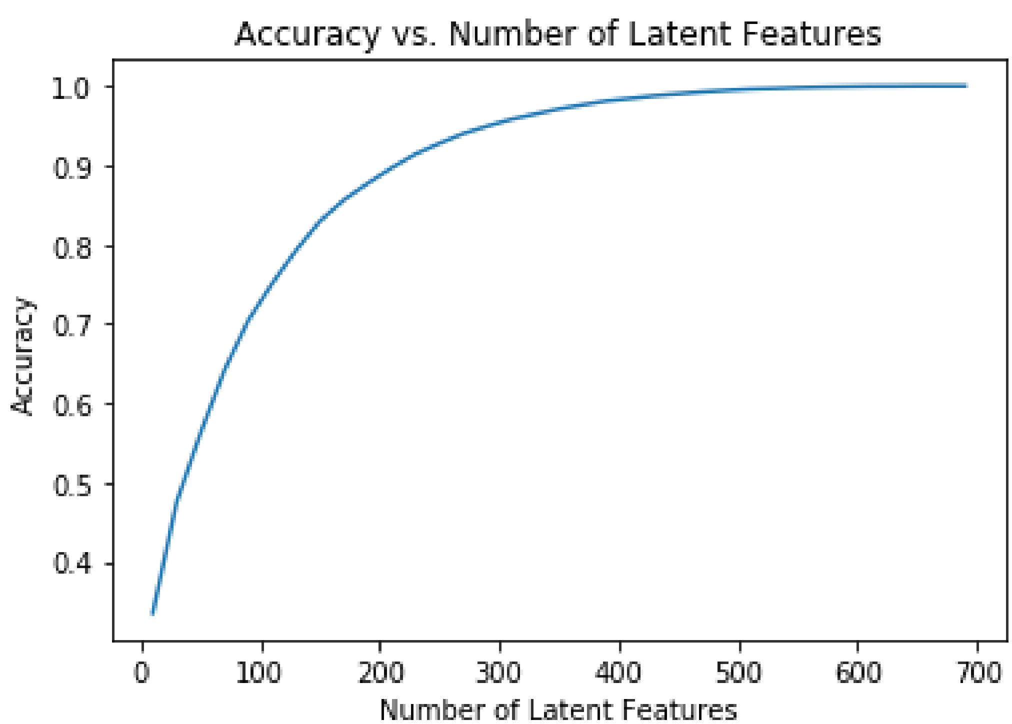
```
(5149, 5149)
(714,)
(714, 714)
```

Provide your response here. User_item_matrix has no missing values, so the SVD can decompsit user-item matrix to the three matrices u, s, vt. I there are any missing values, funkSVD must be used.

3. Now for the tricky part, how do we choose the number of latent features to use? Running the below cell, you can see that as the number of latent features increases, we obtain a lower error

rate on making predictions for the 1 and 0 values in the user-item matrix. Run the cell below to get an idea of how the accuracy improves as we increase the number of latent features.

[39]:
```
num_latent_feats = np.arange(10,700+10,20)
sum_errs = []

for k in num_latent_feats:
    # restructure with k latent features
    s_new, u_new, vt_new = np.diag(s[:k]), u[:, :k], vt[:k, :]

    # take dot product
    user_item_est = np.around(np.dot(np.dot(u_new, s_new), vt_new))

    # compute error for each prediction to actual value
    diffs = np.subtract(user_item_matrix, user_item_est)

    # total errors and keep track of them
    err = np.sum(np.sum(np.abs(diffs)))
    sum_errs.append(err)


plt.plot(num_latent_feats, 1 - np.array(sum_errs)/df.shape[0]);
plt.xlabel('Number of Latent Features');
plt.ylabel('Accuracy');
plt.title('Accuracy vs. Number of Latent Features');
```



4. From the above, we can't really be sure how many features to use, because simply having a better way to predict the 1's and 0's of the matrix doesn't exactly give us an indication of if we are able to make good recommendations. Instead, we might split our dataset into a training and test set of data, as shown in the cell below.

Use the code from question 3 to understand the impact on accuracy of the training and test sets of data with different numbers of latent features. Using the split below:

- How many users can we make predictions for in the test set?

- How many users are we not able to make predictions for because of the cold start problem?
- How many articles can we make predictions for in the test set?

- How many articles are we not able to make predictions for because of the cold start problem?

[40]:
```
df_train = df.head(40000)
df_test = df.tail(5993)

def create_test_and_train_user_item(df_train, df_test):
    '''
    INPUT:
    df_train - training dataframe
    df_test - test dataframe

    OUTPUT:
    user_item_train - a user-item matrix of the training dataframe
                      (unique users for each row and unique articles for each␣
 ↪column)
    user_item_test - a user-item matrix of the testing dataframe
                    (unique users for each row and unique articles for each␣
 ↪column)
    test_idx - all of the test user ids
    test_arts - all of the test article ids

    '''
    # Your code here
    user_item_train = create_user_item_matrix(df_train)
    user_item_test  = create_user_item_matrix(df_test)


    test_idx = user_item_test.index
    test_arts = user_item_test.columns
    return user_item_train, user_item_test, test_idx, test_arts

user_item_train, user_item_test, test_idx, test_arts =␣
 ↪create_test_and_train_user_item(df_train, df_test)
```

[41]:
```python
#How many users can we make predictions for in the test set?:
len(set(user_item_train.index).intersection(set(user_item_test.index)))
```

[41]: 20

[42]:
```python
#How many users in the test set are we not able to make predictions for because␣
↪of the cold start problem?
len(set(user_item_test.index).difference(set(user_item_train.index)))
```

[42]: 662

[43]:
```python
#How many articles can we make predictions for in the test set?
len(set(user_item_train.columns).intersection(set(user_item_test.columns)))
```

[43]: 574

[44]:
```python
#How many articles in the test set are we not able to make predictions for␣
↪because of the cold start problem?
len(set(user_item_test.columns).difference(set(user_item_train.columns)))
```

[44]: 0

[45]:
```python
# Replace the values in the dictionary below
a = 662
b = 574
c = 20
d = 0


sol_4_dict = {
    'How many users can we make predictions for in the test set?': c,
    'How many users in the test set are we not able to make predictions for␣
↪because of the cold start problem?': a,
    'How many movies can we make predictions for in the test set?': b,
    'How many movies in the test set are we not able to make predictions for␣
↪because of the cold start problem?': d
}
t.sol_4_test(sol_4_dict)
```

```
Awesome job!  That's right!  All of the test movies are in the training data,
but there are only 20 test users that were also in the training set.  All of the
other users that are in the test set we have no data on.  Therefore, we cannot
make predictions for these users using SVD.
```

5. Now use the **user_item_train** dataset from above to find U, S, and V transpose using SVD. Then find the subset of rows in the **user_item_test** dataset that you can predict using this matrix decomposition with different numbers of latent features to see how many features makes sense to keep based on the accuracy on the test data. This will require combining what was done in questions 2 - 4.

Use the cells below to explore how well SVD works towards making predictions for recommendations on the test data.

[46]:
```
# fit SVD on the user_item_train matrix
u_train, s_train, vt_train = np.linalg.svd(user_item_train, full_matrices=False)␣
↪# fit svd similar to above then use the cells below
```

[47]:
```
print(u_train.shape)
print(s_train.shape)
print(vt_train.shape)
```

```
(4487, 714)
(714,)
(714, 714)
```

[48]:
```
# Use these cells to see how well you can use the training
# decomposition to predict on test data
```

[49]:
```
u_test = u_train[user_item_train.index.isin(test_idx)]
vt_test = vt_train[user_item_train.columns.isin(test_arts)].T
print(u_test.shape)
print(vt_test.shape)
```

```
(20, 714)
(714, 574)
```

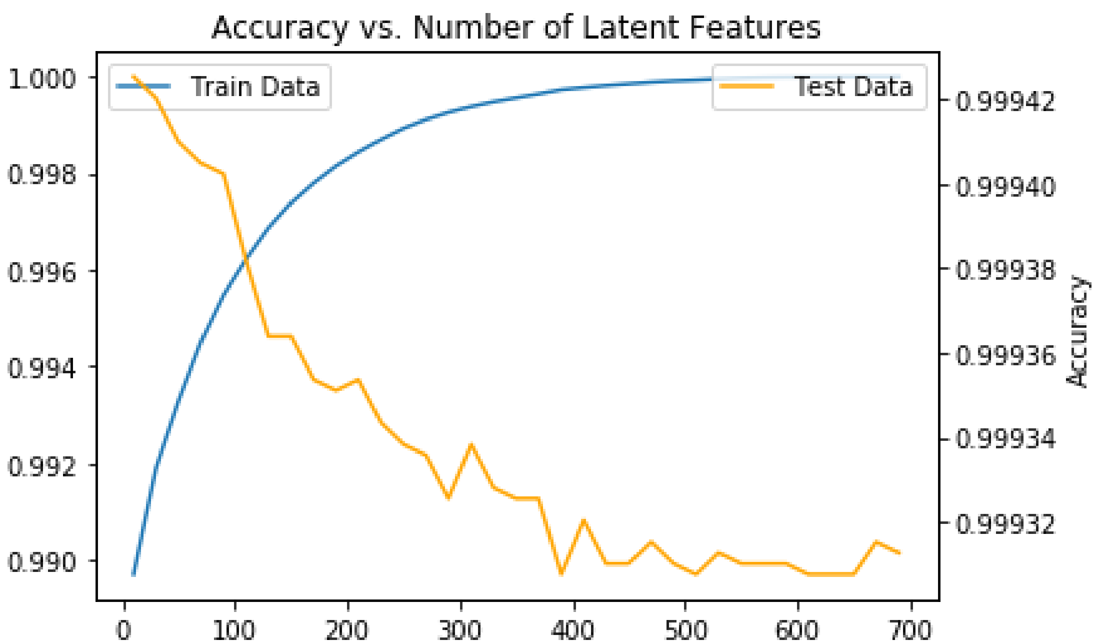
[50]:
```
test_users = set(user_item_train.index).intersection(set(user_item_test.index))
test_articles = set(user_item_train.columns).intersection(set(user_item_test.
↪columns))
```

[53]:
```
user_item_test_subset = user_item_test.loc[test_users, test_articles]
```

[54]:
```
num_latent_feats = np.arange(10,700+10,20)
sum_errs_train = []
sum_errs_test = []

for k in num_latent_feats:
    # restructure with k latent features
    s_train_new, u_train_new, vt_train_new = np.diag(s_train[:k]), u_train[:, :
↪k], vt_train[:k, :]
    u_test_new, vt_test_new = u_test[:, :k], vt_test[:k, :]

    # take dot product
    user_item_train_est = np.around(np.dot(np.dot(u_train_new, s_train_new),␣
↪vt_train_new))
    user_item_test_est = np.around(np.dot(np.dot(u_test_new, s_train_new),␣
↪vt_test_new))

    # compute error for each prediction to actual value
    diffs_train = np.subtract(user_item_train, user_item_train_est)
    diffs_test = np.subtract(user_item_test_subset, user_item_test_est)
```

```
    # total errors and keep track of them
    sum_errs_train.append(np.sum(np.sum(np.abs(diffs_train))))
    sum_errs_test.append(np.sum(np.sum(np.abs(diffs_test))))
```

[61]:
```
fig, ax1 = plt.subplots()
ax2 = ax1.twinx()
ax1.plot(num_latent_feats, 1 - np.array(sum_errs_train)/(user_item_train.
 ↪shape[0]*user_item_test.shape[1]), label="Train Data");
ax2.plot(num_latent_feats, 1 - np.array(sum_errs_test)/(user_item_test.
 ↪shape[0]*user_item_test.shape[1]), label="Test Data", color = "orange");
plt.xlabel('Number of Latent Features');
plt.ylabel('Accuracy');
ax1.legend(loc="best")
ax2.legend(loc="best")
plt.title('Accuracy vs. Number of Latent Features');
```



6. Use the cell below to comment on the results you found in the previous question. Given the circumstances of your results, discuss what you might do to determine if the recommendations you make with any of the above recommendation systems are an improvement to how users currently find articles?

From the graph above, it can bee seen, that the training accuracy increases, the more latent features are in the data. On the other hand, test accuracy decreases, the more latent feature we have. Consequently, our model does not generalize well.

This is a common problem of recommendations systems, that run on collaborative filtering. The more user actions an item has, the easier is to find recommentations. As time progresses, the system will be able to give more and more accurate recommendations. However, with our given data we can only make predictions for 20 users in the testing set, 662 users in the test set are we not able to make predictions for because of the cold start problem.

The solution would be to use content-based recommendations for new user, to overcome the cold start problem. Also to test the real performance of the recommendation system, we suggest to run a A/B test. A simple expertiment design could devide the users into two groups: a group A using the new recommendation system and a group B being the control group. The results should testify, if the new recommendation system performs well in practise.

1.2 Conclusion

Congratulations! You have reached the end of the Recommendations with IBM project!

Tip: Once you are satisfied with your work here, check over your report to make sure that it is satisfies all the areas of the rubric. You should also probably remove all of the "Tips" like this one so that the presentation is as polished as possible.

1.3 Directions to Submit

Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

[63]:
```
from subprocess import call
call(['python', '-m', 'nbconvert', 'Recommendations_with_IBM.ipynb'])
```

[63]: 0

[ ]:
```
```