

# 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.

## Table of Contents

- I. [Exploratory Data Analysis](#)
- II. [Rank Based Recommendations](#)
- III. [User-User Based Collaborative Filtering](#)
- IV. [Content Based Recommendations \(EXTRA - NOT REQUIRED\)](#)
- V. [Matrix Factorization](#)
- VI. [Extras & Concluding](#)

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.

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import project_tests as t
import pickle
import seaborn as sns
from collections import defaultdict

%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()
```

```
Out[1]:
```

|   | article_id | title   | email                                    |
|---|------------|---|--|
| 0 | 1430.0     | using pixiedust for fast, flexible, and easier... | ef5f11f77ba020cd36e1105a00ab868bbdbf7fe7 |
| 1 | 1314.0     | healthcare python streaming application demo      | 083cbdfa93c8444beaa4c5f5e0f5f9198e4f9e0b |

|   | article_id | title                                      | email                                    |
|---|------------|--|--|
| 2 | 1429.0     | use deep learning for image classification | b96a4f2e92d8572034b1e9b28f9ac673765cd074 |
| 3 | 1338.0     | ml optimization using cognitive assistant  | 06485706b34a5c9bf2a0ecdac41daf7e7654ceb7 |

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

```
Out[2]:
```

|   | doc_body  | doc_description   | doc_full_name   | doc_status | article_id |
|---|---|---|---|------------|------------|
| 0 | Skip navigation Sign in<br>SearchLoading...\r\n\r...    | Detect bad readings in<br>real time using Python ...    | Detect Malfunctioning<br>IoT Sensors with<br>Streami... | Live       | 0          |
| 1 | No Free Hunch Navigation<br>* kaggle.com\r\n\r\n ...    | See the forest, see the<br>trees. Here lies the c...    | Communicating data<br>science: A guide to<br>present... | Live       | 1          |
| 2 | ≡ * Login\r\n * Sign<br>Up\r\n\r\n * Learning Pat...    | Here's this week's news<br>in Data Science and Bi...    | This Week in Data<br>Science (April 18, 2017)           | Live       | 2          |
| 3 | DATALAYER: HIGH<br>THROUGHPUT, LOW<br>LATENCY AT SCA... | Learn how distributed<br>DBs solve the problem<br>of... | DataLayer Conference:<br>Boost the performance<br>of... | Live       | 3          |
| 4 | Skip navigation Sign in<br>SearchLoading...\r\n\r...    | This video demonstrates<br>the power of IBM<br>DataS... | Analyze NY Restaurant<br>data using Spark in DSX        | Live       | 4          |

## 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.

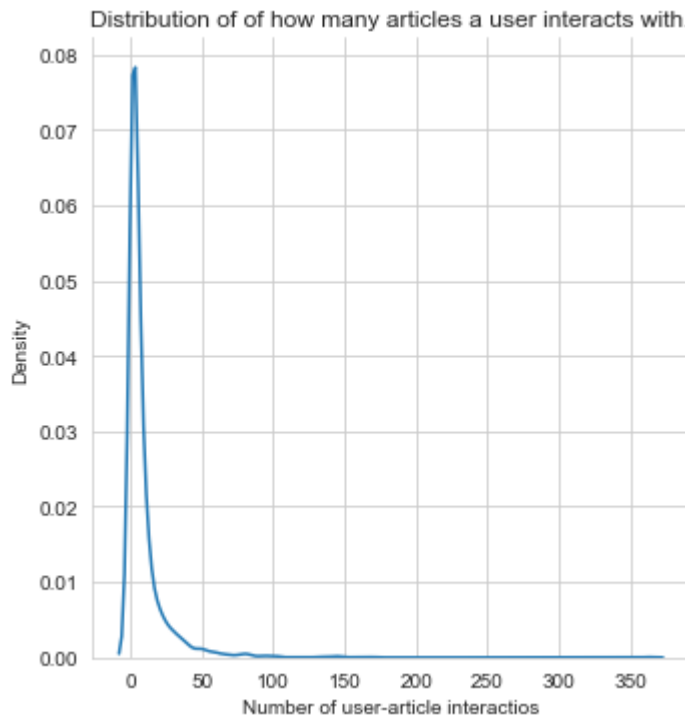
```
In [3]: # Fill in the median and maximum number of user_article interactios below

# 50% of individuals interact with ____ number of articles or fewer.
median_val = df.groupby('email')['article_id'].count().median()
# The maximum number of user-article interactions by any 1 user is _____.
max_views_by_user = df.groupby('email')['article_id'].count().max()
```

```
In [4]: # Distribution plot of number of user-article interactios

sns.set_style('whitegrid')
sns.displot(df.groupby('email')['article_id'].count(), kind='kde')
plt.xlabel('Number of user-article interactios')
plt.title('Distribution of of how many articles a user interacts with.')
```

Out[4]: Text(0.5, 1.0, 'Distribution of of how many articles a user interacts with.')



2. Explore and remove duplicate articles from the **df\_content** dataframe.

```
In [5]: # shape of df_content dataframe
df_content.shape
```

Out[5]: (1056, 5)

```
In [6]: # number of duplicate rows
df_content.duplicated('article_id').sum()
```

Out[6]: 5

```
In [7]: # Find and explore duplicate articles
df_content[df_content.duplicated('article_id', keep=False)]
```

Out[7]:

|     | doc_body  | doc_description                                   | doc_full_name                                     | doc_status | article_id |
|-----|---|---|---|------------|------------|
| 50  | Follow Sign in / Sign up Home About Insight Da...     | Community Detection at Scale                      | Graph-based machine learning                      | Live       | 50         |
| 221 | * United States\r\n\r\nIBM® * Site map\r\n\r\n\r\n... | When used to make sense of huge amounts of con... | How smart catalogs can turn the big data flood... | Live       | 221        |
| 232 | Homepage Follow Sign in Get started Homepage * ...    | If you are like most data scientists, you are ... | Self-service data preparation with IBM Data Re... | Live       | 232        |

|     | doc_body  | doc_description                                   | doc_full_name                                     | doc_status | article_id |
|-----|---|---|---|------------|------------|
| 365 | Follow Sign in / Sign up Home About Insight Da... | During the seven-week Insight Data Engineering... | Graph-based machine learning                      | Live       | 50         |
| 399 | Homepage Follow Sign in Get started * Home\r\n... | Today's world of data science leverages data f... | Using Apache Spark as a parallel processing fr... | Live       | 398        |
| 578 | This video shows you how to construct queries ... | This video shows you how to construct queries ... | Use the Primary Index                             | Live       | 577        |
| 692 | Homepage Follow Sign in / Sign up Homepage * H... | One of the earliest documented catalogs was co... | How smart catalogs can turn the big data flood... | Live       | 221        |
| 761 | Homepage Follow Sign in Get started Homepage *... | Today's world of data science leverages data f... | Using Apache Spark as a parallel processing fr... | Live       | 398        |
| 970 | This video shows you how to construct queries ... | This video shows you how to construct queries ... | Use the Primary Index                             | Live       | 577        |

```
In [8]: # Remove any rows that have the same article_id - only keep the first
df_content.drop_duplicates('article_id', inplace=True)
```

```
In [9]: # Check shape of df_content after dropping duplicates
df_content.shape
```

```
Out[9]: (1051, 5)
```

3. Use the cells below to find:

- The number of unique articles that have an interaction with a user.
- The number of unique articles in the dataset (whether they have any interactions or not).
- The number of unique users in the dataset. (excluding null values)
- The number of user-article interactions in the dataset.

```
In [10]: # Number of unique articles that have an interaction with a user
df.groupby('article_id')['email'].count().sort_values()
```

```
Out[10]: article_id
1092.0      1
984.0       1
417.0       1
1237.0      1
1233.0      1
...
1364.0     627
1427.0     643
1431.0     671
1330.0     927
1429.0     937
```

Name: email, Length: 714, dtype: int64

```
In [11]: # The number of unique articles on the IBM platform
df_content['article_id'].nunique()
```

Out[11]: 1051

```
In [12]: # The number of unique users
df['email'].nunique()
```

Out[12]: 5148

```
In [13]: # The number of user-article interactions
df.shape
```

Out[13]: (45993, 3)

```
In [14]: # The number of user-article interactions after dropping duplicates
df.drop_duplicates().shape[0]
```

Out[14]: 33682

```
In [15]: unique_articles = 714 # The number of unique articles that have at least one
total_articles = 1051 # The number of unique articles on the IBM platform
unique_users = 5148 # The number of unique users
user_article_interactions = 45993 # The number of user-article interactions
```

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).

```
In [16]: df.groupby('article_id')['email'].count().sort_values(ascending=False)
```

```
Out[16]: article_id
1429.0    937
1330.0    927
1431.0    671
1427.0    643
1364.0    627
...
1113.0      1
1119.0      1
984.0       1
1127.0      1
1266.0      1
Name: email, Length: 714, dtype: int64
```

```
In [17]: df.groupby('article_id')['email'].count().idxmax()
```

Out[17]: 1429.0

```
In [18]: # The most viewed article in the dataset as a string with one value following
most_viewed_article_id = "1429.0"
# The most viewed article in the dataset was viewed how many times?
max_views = df.groupby('article_id')['email'].count().max()
```

```
In [19]: # No need to change the code here - this will be helpful for later parts of ti
# Run this cell to map the user email to a user_id column and remove the email.

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()
```

Out[19]:

|   | 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       |

```
In [20]: # 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 ____.',
    'The maximum number of user-article interactions by any 1 user is ____',
    'The most viewed article in the dataset was viewed ____ times.': max_v,
    'The article_id of the most viewed article is ____.': most_viewed_art,
    'The number of unique articles that have at least 1 rating ____.': un,
    '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!

## 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.

```
In [21]: df.head()
```

```
Out[21]:
```

|   | 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       |

```

In [22]: 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_article_ids = get_top_article_ids(n)
top_articles = list(df.groupby('article_id')[
    'title'].unique().apply(lambda t: t[0])[top_article_ids])
# alternately
# top_articles = list(df[df['article_id'].isin(top_article_ids)].drop_duplicates())

# Return the top article titles from df (not df_content)
return top_articles

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_article_ids - (list) A list of the top 'n' article ids
'''

top_article_ids = list(df.groupby('article_id')[
    'user_id'].count().sort_values(ascending=False).index[:n])
# alternative method
# top_article_ids = list(df['article_id'].value_counts().index[:5])

return top_article_ids # Return the top article ids

```

```

In [23]: 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]

```



In [24]:

```
# 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.

## 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.

```
In [25]: # 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
    an article and a 0 otherwise
    """

    user_item = df.groupby(['user_id', 'article_id']
                           ).count().unstack().replace(np.nan, 0)
    user_item[user_item > 0] = 1
    user_item.columns = user_item.columns.droplevel()
    # alternatively
    # user_item.columns = user_item.columns.get_level_values(1)

    return user_item # return the user_item matrix

user_item = create_user_item_matrix(df)
```

```
In [26]: # 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-art
assert user_item.shape[1] == 714, "Oops! The number of articles in the user-a
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.

```
In [27]: 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)
                are listed first

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

'''

# compute similarity of each user to the provided user
user_similarity = user_item.loc[user_id].dot(user_item.transpose())

# sort by similarity
user_similarity_sorted = user_similarity.sort_values(ascending=False)

# create list of just the ids
most_similar_users = user_similarity_sorted.index

# remove the own user's id
most_similar_users = most_similar_users.drop(user_id)

# return a list of the users in order from most to least similar
return list(most_similar_users)
```

```
In [28]: # 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, 13, 1, 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.

```
In [29]: 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
                 (this is identified by the title column)
'''

article_ids = list(np.array(article_ids).astype(float))
article_names = list(df[df['article_id'].isin(
    article_ids)].drop_duplicates('article_id')['title'])

return article_names # Return the article names associated with list of a
```

```
In [30]: 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
                 (this is identified by the doc_full_name column in df_cont

Description:
Provides a list of the article_ids and article titles that have been seen
'''

user_items = user_item.loc[user_id]
article_ids = list(user_items[user_items > 0].index)
article_ids = list(np.array(article_ids).astype(str))
article_names = get_article_names(article_ids)

return article_ids, article_names # return the ids and names
```

```
In [31]: 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
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

'''

recs = []

articles_seen_ids, articles_seen_names = get_user_articles(user_id)

most_similar_users = find_similar_users(user_id)

for user in most_similar_users:
    similar_article_ids, similar_article_names = get_user_articles(user)
    rec_article_ids = np.setdiff1d(
        similar_article_ids, articles_seen_ids, assume_unique=True)
    recs.extend(rec_article_ids)
    recs = list(np.unique(recs))

    if len(recs) >= m:
        break

return recs[:m] # return your recommendations for this user_id
```

```
In [32]: # Check Results
# Return 10 recommendations for user 1
get_article_names(user_user_recs(1, 10))
```

```
Out[32]: ['analyze energy consumption in buildings',
'analyze accident reports on amazon emr spark',
'520 using notebooks with pixiedust for fast, flexi...\nName: title, dtype: object',
'1448 i ranked every intro to data science course on...\nName: title, dtype: object',
'data tidying in data science experience',
'airbnb data for analytics: vancouver listings',
'recommender systems: approaches & algorithms',
'airbnb data for analytics: mallorca reviews',
'analyze facebook data using ibm watson and watson studio',
'a tensorflow regression model to predict house values']
```

In [33]:

```
# 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', '1432.0', '1433.0', '1434.0', '1435.0',
                              '1436.0', '1437.0', '1438.0', '1439.0', '1440.0',
                              '1441.0', '1442.0', '1443.0', '1444.0', '1445.0',
                              '1446.0', '1447.0', '1448.0', '1449.0', '1450.0',
                              '1451.0', '1452.0', '1453.0', '1454.0', '1455.0',
                              '1456.0', '1457.0', '1458.0', '1459.0', '1460.0',
                              '1461.0', '1462.0', '1463.0', '1464.0', '1465.0',
                              '1466.0', '1467.0', '1468.0', '1469.0', '1470.0',
                              '1471.0', '1472.0', '1473.0', '1474.0', '1475.0',
                              '1476.0', '1477.0', '1478.0', '1479.0', '1480.0',
                              '1481.0', '1482.0', '1483.0', '1484.0', '1485.0',
                              '1486.0', '1487.0', '1488.0', '1489.0', '1490.0',
                              '1491.0', '1492.0', '1493.0', '1494.0', '1495.0',
                              '1496.0', '1497.0', '1498.0', '1499.0'])) == set(['housing
                              'use the c

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 de
                              'self-service data preparation w

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 reconstru
                              'healthcare python streaming appl

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.

```
In [34]: 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
                num_interactions - the number of articles viewed by the user

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

'''

similarity = user_item.loc[user_id].dot(user_item.transpose())
similarity.rename('similarity', inplace=True)

num_interactions = df.groupby('user_id')['article_id'].count()
num_interactions.rename('num_interactions', inplace=True)

neighbors_df = similarity.to_frame().join(num_interactions)
neighbors_df.index.rename('neighbor_id', inplace=True)

neighbors_df.sort_values(
    by=['similarity', 'num_interactions'], ascending=False, inplace=True)
neighbors_df.drop(user_id, inplace=True)

return neighbors_df # Return the dataframe specified in the doc_string
```

```
In [35]: # top articles
top_articles = df.groupby('article_id')[
    'user_id'].count().sort_values(ascending=False)
top_articles
```

```
Out[35]: article_id
1429.0    937
1330.0    927
1431.0    671
1427.0    643
1364.0    627
...
1113.0     1
1119.0     1
984.0      1
1127.0     1
1266.0     1
Name: user_id, Length: 714, dtype: int64
```

```
In [36]: 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
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.

'''
recs = []

articles_seen_ids, articles_seen_names = get_user_articles(user_id)

most_similar_users = get_top_sorted_users(user_id).index

for user in most_similar_users:
    similar_article_ids, similar_article_names = get_user_articles(user)
    rec_article_ids = np.setdiff1d(
        similar_article_ids, articles_seen_ids, assume_unique=True)
    # arrange article ids by total interactions
    rec_article_ids = list(
        top_articles[rec_article_ids.astype(float)].sort_values(ascending=False).index[:m])
    recs.extend(rec_article_ids)
    recs = list(np.unique(recs))

    if len(recs) >= m:
        break

recs = recs[:m]
rec_names = get_article_names(recs)

return recs, rec_names
```

```
In [37]: # Quick spot check - don't change this code - just use it to test your function
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:



```
[12.0, 109.0, 125.0, 142.0, 164.0, 205.0, 302.0, 336.0, 362.0, 465.0]
```

The top 10 recommendations for user 20 are the following article names:  
['timeseries data analysis of iot events by using jupyter notebook', 'dsx: hybrid mode', 'accelerate your workflow with dsx', 'learn tensorflow and deep learning together and now!', 'a beginner's guide to variational methods', 'tensorflow quick tips', 'challenges in deep learning', 'neural networks for beginners: popular types and applications', 'statistics for hackers', 'introduction to neural networks: advantages and applications']

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 comments below.

```
In [38]: # Tests with a dictionary of results

# Find the user that is most similar to user 1
user1_most_sim = get_top_sorted_users(1).index[0]
# Find the 10th most similar user to user 131
user131_10th_sim = get_top_sorted_users(131).index[9]
```

```
In [39]: # 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.

**Since the new user id is not present in the user\_item dataframe, we cannot make recommendations using collaborative filtering. In other words, we have no information about the user to find other similar users. Hence, we will use a rank based recommender system to recommend articles to the new user.**

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.

```
In [40]: new_user = '0.0'

# What would your recommendations be for this new user '0.0'? As a new user,
# Provide a list of the top 10 article ids you would give to
# Your recommendations here
new_user_recs = list(np.array(get_top_article_ids(10)).astype(str))
```

```
In [41]: assert set(new_user_recs) == set(['1314.0', '1429.0', '1293.0', '1427.0', '116
                                             ), "Oops! It makes sense that in this case v

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

That's right! Nice job!

## 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.

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.

```
In [42]: 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?

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.

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.

```
In [43]: # make recommendations for a brand new user

# make a recommendations for a user who only has interacted with article id '1'
```

## 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.

```
In [44]: # Load the matrix here
user_item_matrix = pd.read_pickle('user_item_matrix.p')
```

```
In [45]: # quick look at the matrix
user_item_matrix.head()
```

```
Out[45]: article_id  0.0  100.0  1000.0  1004.0  1006.0  1008.0  101.0  1014.0  1015.0  1016.0  ...  977.0  98.0
user_id
1  0.0   0.0   0.0   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   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   0.0  ...   1.0   0.0
4  0.0   0.0   0.0   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   0.0  ...   0.0   0.0
```

5 rows × 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.

```
In [46]: # Perform SVD on the User-Item Matrix Here

# use the built in to get the three matrices
u, s, vt = np.linalg.svd(user_item_matrix)
```

**Numpy's SVD works in this case cause we updated all null values to 0. If we didn't we would have to use FunkSVD and find the optimal values for the matrices using gradient**

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.

```
In [47]: num_latent_feats = np.arange(10, 700+10, 20)
sum_errs = []

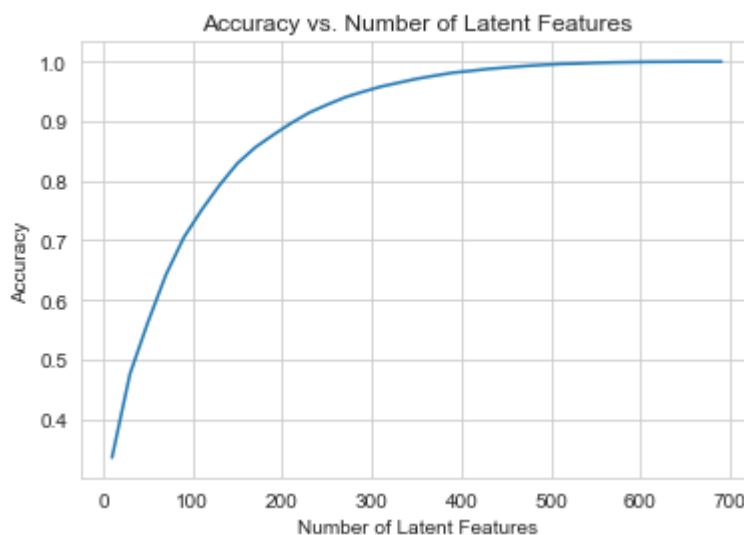
for k in num_latent_feats:
    # restructure with k latent features
    s_new, u_new, vt_new = np.diag(s[:k]), u[:, :k], vt[:, k, :]

    # 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

In [48]:

```
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 row)
    user_item_test - a user-item matrix of the testing dataframe
                    (unique users for each row and unique articles for each row)
    test_idx - all of the test user ids
    test_arts - all of the test article ids

    """

    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)
```

In [49]:

```
# How many users can we make predictions for in the test set?
len(np.intersect1d(test_idx, user_item_train.index))
```

Out[49]: 20

In [50]:

```
# How many users in the test set are we not able to make predictions for because of the cold start problem?
len(test_idx) - len(np.intersect1d(test_idx, user_item_train.index))
```

Out[50]: 662

In [51]:

```
# How many movies can we make predictions for in the test set?
len(np.intersect1d(test_arts, user_item_train.columns))
```

Out[51]: 574

```
In [52]: # How many movies in the test set are we not able to make predictions for because  
len(test_arts) - len(np.intersect1d(test_arts, user_item_train.columns))
```

```
Out[52]: 0
```

```
In [53]: # 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': d,  
    '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': a  
}  
  
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.

```
In [54]: # fit SVD on the user_item_train matrix  
# fit svd similar to above then use the cells below  
u_train, s_train, vt_train = np.linalg.svd(user_item_train)
```

```
In [55]: # Use these cells to see how well you can use the training  
# decomposition to predict on test data
```

```
In [56]: # Users can we make predictions for in the test set  
pred_users = np.intersect1d(test_idx, user_item_train.index)
```

```
In [57]: # Test user indices  
idx = np.where(user_item_train.index.isin(pred_users))[0]
```

```
In [58]: # Test item indices
col = np.where(user_item_train.columns.isin(user_item_test.columns))[0]
```

```
In [59]: 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_train[:k]), u_train[:, :k], vt_train[:k, :]

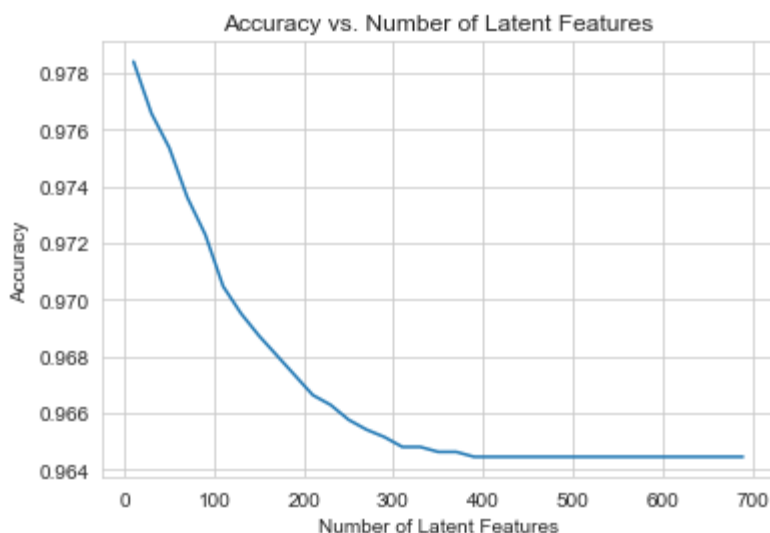
    # take dot product
    user_item_est = np.around(
        np.dot(np.dot(u_new[idx, :], s_new), vt_new[:, col]))

    # compute error for each prediction to actual value
    diffs = np.subtract(user_item_test.loc[pred_users], 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)/(len(idx) * len(col)))
plt.xlabel('Number of Latent Features')
plt.ylabel('Accuracy')
plt.title('Accuracy vs. Number of Latent Features')
```

```
Out[59]: Text(0.5, 1.0, '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 plot above, we can see how the accuracy goes down by adding more latent features. There are 4487 user ids in the training set and 682 user ids in the test set.**

However, they only have 20 user ids in common. The observation in the training data are not representative of the population as a whole. Increasing the order of randomness before splitting the data may help improve this. Secondly, the type of ratings in the user-item matrix is binary. Have a rating system with a broader scale may help improve predictions.

We could do an A/B test online for a period of time to measure the effectiveness of any of the above recommendation systems vs. the existing system in place. We can create a cookie-based diversion to split users into 2 experimental groups and use the number of articles read by users as the performance evaluation metric.

## Extras

Using your workbook, you could now save your recommendations for each user, develop a class to make new predictions and update your results, and make a flask app to deploy your results. These tasks are beyond what is required for this project. However, from what you learned in the lessons, you are certainly capable of taking these tasks on to improve upon your work here!

## 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 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.

## 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!



```
In [1]: from subprocess import call  
        call(['python', '-m', 'nbconvert', 'Recommendations_with_IBM.ipynb'])
```

```
Out[1]: 1
```