Recommendations_with_IBM

May 8, 2023

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

I. Section **??** II. Section **??** III. Section **??** IV. Section **??** V. Section **??** VI. Section **??**

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 [2]:
```
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()
```

Out[2]:

	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
```

In [3]:
```
# change article_id to String data type
df['article_id'] = df['article_id'].astype(str)
```

In [4]:
```
# Show df_content to get an idea of the data
df_content.head()
```

Out[4]:
```
                                            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.

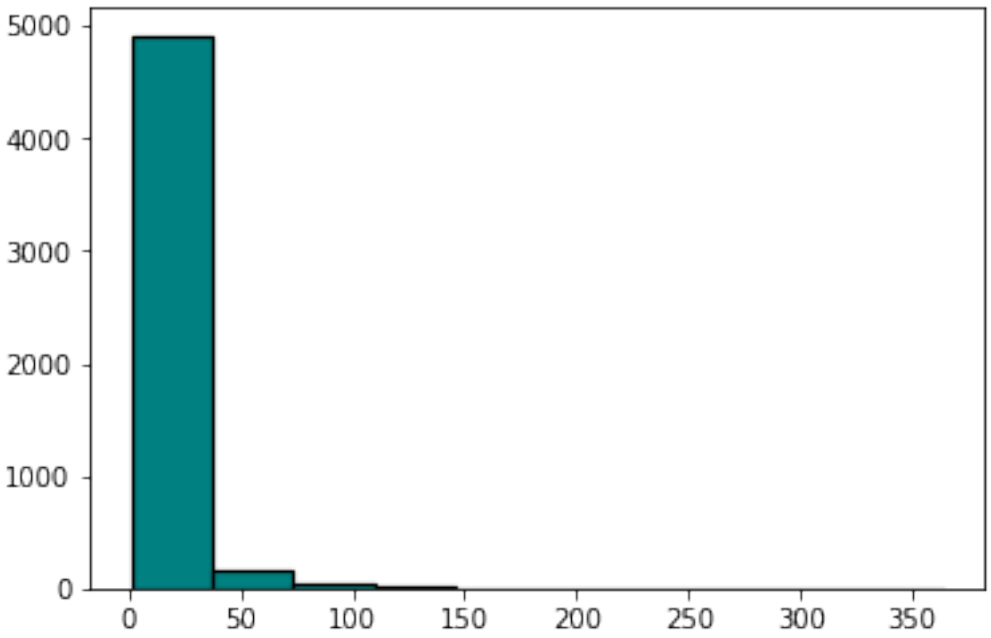
In [5]:
```
data = df.groupby('email')['article_id'].count()
data.describe()
```

Out[5]:
```
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: article_id, dtype: float64
```

```
In [6]: plt.hist(data, bins = 10,color='#008080', edgecolor='black', linewidth=1.2)
```

```
Out[6]: (array([  4.91500000e+03,   1.67000000e+02,   4.60000000e+01,
                  1.10000000e+01,   7.00000000e+00,   0.00000000e+00,
                  0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
                  2.00000000e+00]),
         array([   1. ,   37.3,   73.6,  109.9,  146.2,  182.5,  218.8,  255.1,
                 291.4,  327.7,  364. ]),
         <a list of 10 Patch objects>)
```



- The distribution of user activities is highly imbalanced. In other words, a large majority of users, specifically 75%, only engage with less than 10 articles.

```
In [7]: # 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
```

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

In [8]:
```
# Find and explore duplicate articles
df_content[df_content['article_id'].duplicated(keep=False)].sort_values(by='article_id')
```

Out[8]:

	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
365	Follow Sign in / Sign up Home About Insight Da...	During the seven-week Insight Data Engineering...	Graph-based machine learning	Live	50
221	* United States\r\n\r\nIBMő * Site map\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
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
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
971	Homepage Follow Sign in Get started * Home\r\n...	If you are like most data scientists, you are ...	Self-service data preparation with IBM Data Re...	Live	232
399	Homepage Follow Sign in Get started * Home\r\n...	Todays world of data science leverages data f...	Using Apache Spark as a parallel processing fr...	Live	398
761	Homepage Follow Sign in Get started Homepage *...	Todays 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
970	This video shows you how to construct queries ...	This video shows you how to construct queries ...	Use the Primary Index	Live	577

In [9]:
```
# Remove any rows that have the same article_id - only keep the first
df_content_unique = df_content.drop_duplicates(subset = ['article_id'], keep = 'first')
df_content.shape, df_content_unique.shape
```

Out[9]:
```
((1056, 5), (1051, 5))
```

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.

```
In [10]: df.article_id.nunique(), df_content_unique.article_id.nunique(), df.email.nunique(), df

Out[10]: (714, 1051, 5148, 45993)

In [11]: unique_articles = 714 # The number of unique articles that have at least one interactio
         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 [12]: df.groupby(['article_id'])['email'].count().sort_values(ascending=False).head()

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

In [13]: most_viewed_article_id ='1429.0' # The most viewed article in the dataset as a string u
         max_views = 937 # The most viewed article in the dataset was viewed how many times?

In [14]: ## No need to change the code here - this will be helpful for later parts of the noteb
         # 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()
```

Out[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

In [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_a
    '`The maximum number of user-article interactions by any 1 user is ______.`': max_v
    '`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_artic
    '`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.

In [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

    '''
```

```
        # Your code here
        top_titles = df.title.value_counts().sort_values(ascending=False)
        top_articles = top_titles[:n,].index.tolist()
        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_ids = df.article_id.value_counts().sort_values(ascending=False)
        top_articles = top_ids[:n,].index.tolist()
        return top_articles # Return the top article ids
```

```
In [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', 'v
['1429.0', '1330.0', '1431.0', '1427.0', '1364.0', '1314.0', '1293.0', '1170.0', '1162.0', '1304
```

```
In [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.

In [19]:
```
# 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
    an article and a 0 otherwise
    '''
    # Fill in the function here
    subset = df[['user_id', 'article_id']]
    subset['count'] = 1
    user_item = subset.groupby(['user_id', 'article_id'])['count'].max().unstack()
    user_item.fillna(0, inplace=True)
    return user_item # return the user_item matrix

user_item = create_user_item_matrix(df)
```

In [20]:
```
## 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 matr
assert user_item.shape[1] == 714, "Oops!  The number of articles in the user-article ma
assert user_item.sum(axis=1)[1] == 36, "Oops!  The number of articles seen by user 1 do
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 [21]: 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

             '''
             # compute similarity of each user to the provided user
             similarity_matrix = user_item.dot(np.transpose(user_item))
             similarity_for_user_id = similarity_matrix.loc[user_id, :]
             # sort by similarity
             similarity_for_user_id = similarity_for_user_id.sort_values(ascending=False)
             # create list of just the ids
             most_similar_users = similarity_for_user_id.index.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


In [22]: # 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)[:
         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, 131, 3870, 46, 4201, 5041]
The 5 most similar users to user 3933 are: [1, 23, 3782, 4459, 203]
The 3 most similar users to user 46 are: [4201, 23, 3782]
```

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 [23]: 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
                    (this is identified by the title column)
    '''
    # Your code here
    article_names = np.unique(df[df.article_id.isin(article_ids)]['title']).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
                    (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
    userinfo = user_item.loc[user_id, :]
    article_ids = userinfo[userinfo.values == 1].index.astype('str')
    article_names = get_article_names(article_ids, df)
    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 rec
    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
        # get a list of articles that the given user has seen
        seen_articles = get_user_articles(user_id, user_item)[0]

        # retrieve similar user lists
        similar_users = find_similar_users(user_id, user_item)

        # loop through the users based on the closeness

        recs = np.array([])

        for user in similar_users:

            neighbor_seen_articles = get_user_articles(user, user_item)[0]

            # find the list of articles that are not seen by the user
            new_recs = np.setdiff1d(neighbor_seen_articles, seen_articles, assume_unique=Tr

            # Update recs with new recs
            recs = np.unique(np.concatenate([new_recs, recs], axis=0))

            # If we have enough recommendations exit the loop
            if len(recs) >= m:
                break

        new_recs = recs[:m].tolist()
        return recs # return your recommendations for this user_id
```

In [24]:
```
# Check Results
get_article_names(user_user_recs(1, 10)) # Return 10 recommendations for user 1
```

Out[24]:
```
['1448    i ranked every intro to data science course on...\nName: title, dtype: object
 '5 practical use cases of social network analytics: going beyond facebook and twitter'
 '502    forgetting the past to learn the future: long ...\nName: title, dtype: object'
 '520    using notebooks with pixiedust for fast, flexi...\nName: title, dtype: object'
 '54174    detect potentially malfunctioning sensors in r...\nName: title, dtype: objec
 '56594    lifelong (machine) learning: how automation ca...\nName: title, dtype: objec
 'a dynamic duo  inside machine learning  medium',
 'a tensorflow regression model to predict house values',
 'accelerate your workflow with dsx',
 'airbnb data for analytics: mallorca reviews',
 'airbnb data for analytics: vancouver listings',
 'analyze accident reports on amazon emr spark',
 'analyze energy consumption in buildings',
 'analyze facebook data using ibm watson and watson studio',
```

```
'analyze open data sets with pandas dataframes',
'analyze open data sets with spark & pixiedust',
'analyze precipitation data',
'analyzing data by using the sparkling.data library features',
'apache spark lab, part 2: querying data',
'apache spark lab, part 3: machine learning',
'aspiring data scientists! start to learn statistics with these 6 books!',
'automating web analytics through python',
'awesome deep learning papers',
'better together: spss and data science experience',
'brunel 2.0 preview',
'brunel in jupyter',
'brunel interactive visualizations in jupyter notebooks',
'build a python app on the streaming analytics service',
'car performance data',
'challenges in deep learning',
'data science for real-time streaming analytics',
'data science platforms are on the rise and ibm is leading the way',
'data tidying in data science experience',
'data visualization playbook: telling the data story',
'declarative machine learning',
'deep forest: towards an alternative to deep neural networks',
'deep learning achievements over the past year ',
'deep learning from scratch i: computational graphs',
'deep learning with data science experience',
'deploy your python model as a restful api',
'discover hidden facebook usage insights',
'dsx: hybrid mode',
'easy json loading and social sharing in dsx notebooks',
'experience iot with coursera',
'fertility rate by country in total births per woman',
'flightpredict ii: the sequel   ibm watson data lab',
'from scikit-learn model to cloud with wml client',
'from spark ml model to online scoring with scala',
'generalization in deep learning',
'get social with your notebooks in dsx',
'get started with streams designer by following this roadmap',
'gosales transactions for logistic regression model',
'got zip code data? prep it for analytics.  ibm watson data lab  medium',
'graph-based machine learning',
'healthcare python streaming application demo',
'higher-order logistic regression for large datasets',
'how smart catalogs can turn the big data flood into an ocean of opportunity',
'i am not a data scientist  ibm watson data lab',
'improving real-time object detection with yolo',
'insights from new york car accident reports',
'intents & examples for ibm watson conversation',
'learn basics about notebooks and apache spark',
```

```
'learn tensorflow and deep learning together and now!',
'leverage python, scikit, and text classification for behavioral profiling',
'machine learning and the science of choosing',
'machine learning exercises in python, part 1',
'machine learning for the enterprise',
'markdown for jupyter notebooks cheatsheet',
'maximize oil company profits',
'ml algorithm != learning machine',
'ml optimization using cognitive assistant',
'model bike sharing data with spss',
'modeling energy usage in new york city',
'movie recommender system with spark machine learning',
'optimizing a marketing campaign: moving from predictions to actions',
'overlapping co-cluster recommendation algorithm (ocular)',
'perform sentiment analysis with lstms, using tensorflow',
'pixieapp for outlier detection',
'pixiedust 1.0 is here!  ibm watson data lab',
'pixiedust gets its first community-driven feature in 1.0.4',
'predicting churn with the spss random tree algorithm',
'process events from the watson iot platform in a streams python application',
'programmatic evaluation using watson conversation',
'python machine learning: scikit-learn tutorial',
'recent trends in recommender systems',
'recommender systems: approaches & algorithms',
'shaping data with ibm data refinery',
'simple graphing with ipython and\xa0pandas',
'small steps to tensorflow',
'spark 2.1 and job monitoring available in dsx',
'spark-based machine learning tools for capturing word meanings',
'the 3 kinds of context: machine learning and the art of the frame',
'the nurse assignment problem',
'the power of machine learning in spark',
'the unit commitment problem',
'this week in data science (april 18, 2017)',
'this week in data science (april 25, 2017)',
'this week in data science (february 14, 2017)',
'this week in data science (may 2, 2017)',
'this week in data science (may 30, 2017)',
'times world university ranking analysis',
'timeseries data analysis of iot events by using jupyter notebook',
'twelve\xa0ways to color a map of africa using brunel',
'use decision optimization to schedule league games',
'use sql with data in hadoop python',
'using bigdl in dsx for deep learning on spark',
'using brunel in ipython/jupyter notebooks',
'using deep learning with keras to predict customer churn',
'using github for project control in dsx',
'using machine learning to predict parking difficulty',
```

```
        'using rstudio in ibm data science experience',
        'variational auto-encoder for "frey faces" using keras',
        'visualising data the node.js way',
        'visualize data with the matplotlib library',
        'web picks (week of 4 september 2017)',
        'what is smote in an imbalanced class setting (e.g. fraud detection)?',
        'why even a moths brain is smarter than an ai',
        'working with db2 warehouse on cloud in data science experience']
```

```
In [25]: # 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
         assert set(get_article_names(['1320.0', '232.0', '844.0'])) == set(['housing (2015): un
         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
         assert set(get_user_articles(2)[0]) == set(['1024.0', '1176.0', '1305.0', '1314.0', '14
         assert set(get_user_articles(2)[1]) == set(['using deep learning to reconstruct high-re
         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 [26]: 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
                            num_interactions - the number of articles viewed by the user - if a

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

```
    '''
    # Your code here
     # get similaritydf
    similarity_matrix = user_item.dot(np.transpose(user_item))
    similarity_for_user_id = similarity_matrix.loc[user_id, :]
    similarity = pd.DataFrame(similarity_for_user_id)
    similarity.reset_index(level=0, inplace=True)
    similarity.rename(columns = {user_id: 'similarity'}, inplace=True)

    # get ineractionsdf
    interactions = pd.DataFrame(df.groupby('user_id')['article_id'].count())
    interactions.reset_index(level=0, inplace=True)

    # merge the two df
    neighbors_df = pd.merge(similarity, interactions, how='left', on='user_id')
    neighbors_df = neighbors_df[neighbors_df.user_id != user_id]

    neighbors_df.rename(columns = {'article_id': 'num_interactions'}, inplace=True)
    neighbors_df = neighbors_df.sort_values(by=['similarity', 'num_interactions'], asce

    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 rec
    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
```

```
            seen_articles = get_user_articles(user_id, user_item)[0]

            # retrieve similar user lists
            neighbors_df = get_top_sorted_users(user_id, df, user_item)
            neighbors_list = neighbors_df.user_id.tolist()

            # loop through the users based on the closeness

            recs = np.array([])

            for user in neighbors_list:

                neighbor_seen_articles = get_user_articles(user, user_item)[0]

                # find the list of articles that are not seen by the user
                new_recs = np.setdiff1d(neighbor_seen_articles, seen_articles, assume_unique=Tr

                # Update recs with new recs
                recs = np.unique(np.concatenate([new_recs, recs], axis=0))

                # If we have enough recommendations exit the loop
                if len(recs) >= m:
                    break

            recs = recs[:m].tolist()
            rec_names = get_article_names(recs, df=df)
            return recs, rec_names
```

```
In [27]: # 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:
['1024.0', '1085.0', '109.0', '1150.0', '1151.0', '1152.0', '1153.0', '1154.0', '1157.0', '1160.

The top 10 recommendations for user 20 are the following article names:
['airbnb data for analytics: chicago listings', 'airbnb data for analytics: venice calendar', 'a
```

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.

```
In [28]: ### Tests with a dictionary of results
```

```
user1_most_sim = get_top_sorted_users(1, df, user_item).iloc[0,0] # Find the user that
user131_10th_sim = get_top_sorted_users(131, df, user_item).iloc[9,0] # Find the 10th m
```

In [29]:
```
## 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.

- Due to the lack of information on the new user's interactions with articles, we cannot utilize the user-user based collaborative filtering method which depends on finding similar users based on their article interactions. In such cases, we could employ rank-based recommendation methods, such as recommending the most popular articles based on their interactions with users.
- In situations like this, referred to as the "cold start problem," we could also consider using content-based recommendations that use the article's content to make recommendations.

Provide your response here.

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 [30]:
```
new_user = '0.0'

# What would your recommendations be for this new user '0.0'?  As a new user, they have
# Provide a list of the top 10 article ids you would give to
new_user_recs = get_top_article_ids(10)
```

In [31]:
```
assert set(new_user_recs) == set(['1314.0','1429.0','1293.0','1427.0','1162.0','1364.0'

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.

```
In [32]: 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.

```
In [33]: # 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.

```
In [34]: # Load the matrix here
         user_item_matrix = pd.read_pickle('user_item_matrix.p')

In [35]: # quick look at the matrix
         user_item_matrix.head()

Out[35]: 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.

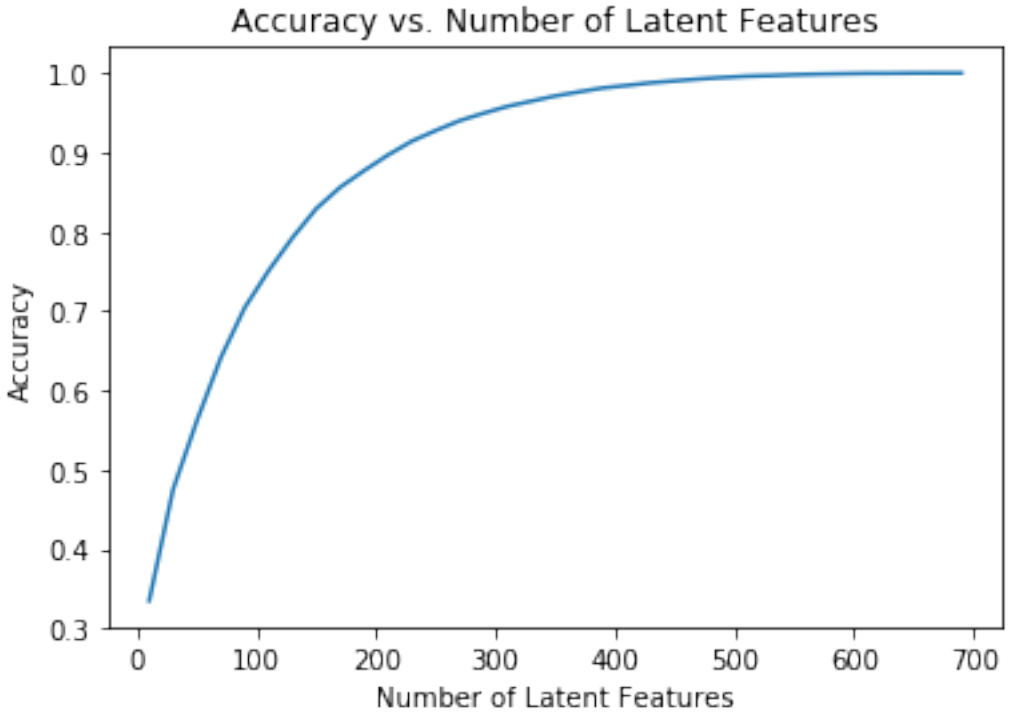
```
In [36]: # Perform SVD on the User-Item Matrix Here

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

In [37]: u.shape, s.shape, vt.shape

Out[37]: ((5149, 5149), (714,), (714, 714))
```

Provide your response here.

- NumPy's implementation of SVD cannot handle matrices with missing values, as was the case with the user-movie matrix used in the lesson. However, our current user_item_matrix does not have any missing values, and therefore, we can use NumPy's Singular Value Decomposition for our purpose.

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 [38]: 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?

```
In [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
                 # make user_item matrix
             user_item_train = create_user_item_matrix(df_train)
             user_item_test = create_user_item_matrix(df_test)

             # Revise user_item_test to only include users and articles that are also found in u
             # List for train & test data - rows and columns values
             # The rows to be used for test data
             train_idx = set(user_item_train.index)
             test_idx = set(user_item_test.index)
             common_rows = train_idx.intersection(test_idx)

             # The columns to be used for test data
             train_arts = set(user_item_train.columns)
```

```
    test_arts = set(user_item_test.columns)
    common_cols = train_arts.intersection(test_arts)

    user_item_test = user_item_test.loc[common_rows, common_cols]
    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(
```

In [41]:
```
user_item_train.shape, user_item_test.shape, len(test_idx), len(test_arts)
```

Out[41]:
```
((4487, 714), (20, 574), 682, 574)
```

In [43]:
```
# 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
    'How many articles can we make predictions for in the test set?': b,
    'How many articles in the test set are we not able to make predictions for because
}

t.sol_4_test(sol_4_dict)
```

```
Awesome job!  That's right!  All of the test articles are in the training data, but there are on
```

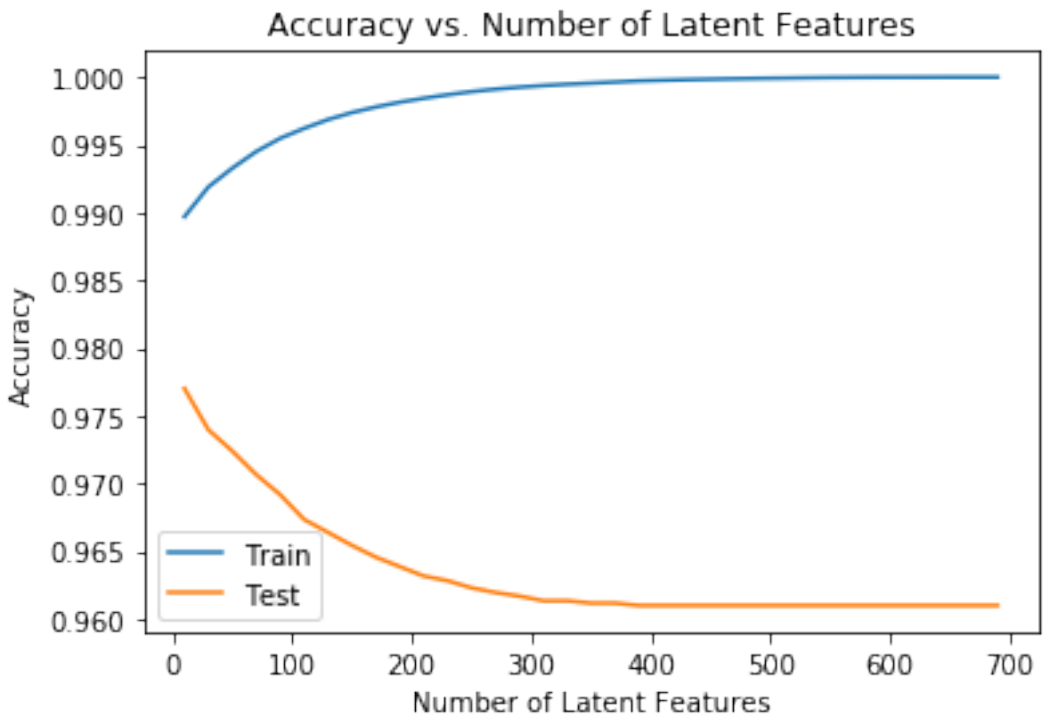
5. Now use the **user_item_train** dataset from above to find U, S, and V transpose using SVD. Then find the subset of rows in the **user_item_test** dataset that you can predict using this matrix decomposition with different numbers of latent features to see how many features makes sense to keep based 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 [44]:
```
# fit SVD on the user_item_train matrix
u_train, s_train, vt_train = np.linalg.svd(user_item_train)
```

In [45]:
```
# Use these cells to see how well you can use the training
# decomposition to predict on test data
row_idx = user_item_train.index.isin(test_idx)
col_idx = user_item_train.columns.isin(test_arts)

u_test = u_train[row_idx, :]
vt_test = vt_train[:, col_idx]
```

```
In [46]: u_test.shape, vt_test.shape

Out[46]: ((20, 4487), (714, 574))

In [50]: 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_t
             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_n
             user_item_test_ets = np.around(np.dot(np.dot(u_test_new, s_train_new), vt_test_new)

             # Calculate the error of each prediction with the true value
             diffs_train = np.subtract(user_item_train, user_item_train_est)
             diffs_test = np.subtract(user_item_test, user_item_test_ets)

             # Total Error
             err_train = np.sum(np.sum(np.abs(diffs_train)))
             err_test = np.sum(np.sum(np.abs(diffs_test)))

             sum_errs_train.append(err_train)
             sum_errs_test.append(err_test)

In [51]: plt.plot(num_latent_feats, 1 - np.array(sum_errs_train)/(user_item_train.shape[0]*user_
         plt.plot(num_latent_feats, 1 - np.array(sum_errs_test)/(user_item_test.shape[0]*user_it
         plt.xlabel('Number of Latent Features');
         plt.ylabel('Accuracy');
         plt.title('Accuracy vs. Number of Latent Features');
         plt.legend();
```



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?

Your response here.

- Although the prediction accuracy for the train data improves with an increase in the number of latent features, the accuracy for the test data decreases. This could be attributed to the imbalanced dataset, where only 20 out of over 4000 users in the train set are also in the test set. The latent features that capture relationships between users and activities from the train dataset may not be useful for predicting activities from users in the test set.
- To address this issue, we could consider combining content-based recommendation with other recommendation methods. One way to do this is by utilizing information from df_content to define categories for each article and combining it with rank-based recommendation, using it as a filter for user-user based collaborative filtering.
- For new users, displaying the most popular articles can definitely be an improvement.
- Matrix factorization is an offline testing method that can be used to evaluate the effectiveness of our model. Additionally, we could carry out A/B testing (online testing) or use group analysis to obtain feedback on the performance of the recommendation engine.

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 certainly capable of taking these tasks on to improve upon your work here!

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!

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