Recommendations_with_IBM

August 21, 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

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.

[11]:
```
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()
```

[11]:
```
   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  f01220c46fc92c6e6b1849de11faacd7ccb2
```

[12]:
```
# Show df_content to get an idea of the data
df_content.head()
```

[12]:
```
                                            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  Here's this week's 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
```

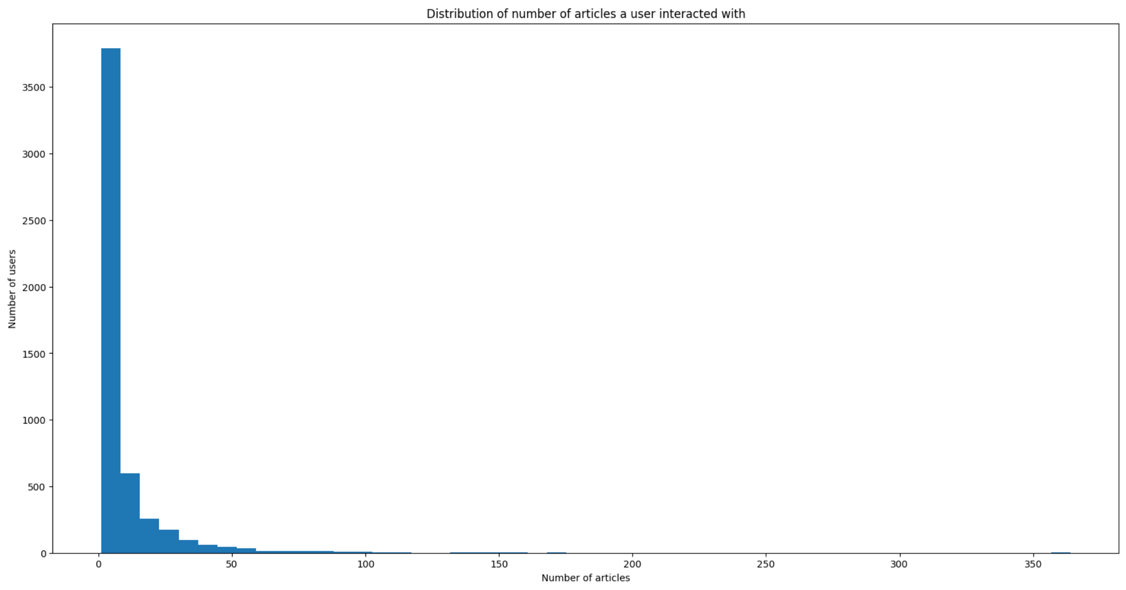
1.1.1 Part I : Exploratory Data Analysis

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.

[13]:
```
# count the number of articles a user has interacted with
num_articles = df.groupby('email')['article_id'].count()

# Plot the distribution in a histogram
fig = plt.figure(figsize=(20, 10))
```

```
plt.hist(num_articles, bins=50)
plt.xlabel('Number of articles')
plt.ylabel('Number of users')
plt.title('Distribution of number of articles a user interacted with')
plt.show()
```



[14]:
```
num_articles.describe()
```

[14]:
```
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
```

[15]:
```
# Fill in the median and maximum number of user_article interactios below

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

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

```
[16]: # Find and explore duplicate articles
      df_content.duplicated(subset='article_id').sum()
```

```
[16]: 5
```

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

3. Use the cells below to find:

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

```
[18]: num_unique_articles_user = df[df['email'].notnull() == True]['article_id'].
       ↪nunique()
      num_unique_articles_user
```

```
[18]: 714
```

```
[19]: num_unique_articles = df_content['article_id'].nunique()
      num_unique_articles
```

```
[19]: 1051
```

```
[20]: num_unique_users = df['email'].nunique()
      num_unique_users
```

```
[20]: 5148
```

```
[21]: # we consider even the interaction with null emails
      df.count()
```

```
[21]: article_id    45993
      title         45993
      email         45976
      dtype: int64
```

```
[22]: unique_articles = 714 # The number of unique articles that have at least one␣
       ↪interaction
      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).

```
[23]: most_viewed_article = df['article_id'].value_counts().idxmax()
      most_viewed_article_count = df['article_id'].value_counts().max()
      print('Most viewed article id:', most_viewed_article)
      print('Number of times:', most_viewed_article_count)
```

```
Most viewed article id: 1429.0
Number of times: 937
```

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

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

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

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

              email_encoded.append(coded_dict[val])
          return email_encoded

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

      # show header
      df.head()
```

```
[25]:    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
```

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

```
sol_1_dict = {
    '`50% of individuals have _____ or fewer interactions.`': median_val,
    '`The total number of user-article interactions in the dataset is ______.`':
 ↪ user_article_interactions,
    '`The maximum number of user-article interactions by any 1 user is ______.
 ↪`': max_views_by_user,
    '`The most viewed article in the dataset was viewed _____ times.`':␣
 ↪max_views,
    '`The article_id of the most viewed article is ______.`':␣
 ↪most_viewed_article_id,
    '`The number of unique articles that have at least 1 rating ______.`':␣
 ↪unique_articles,
    '`The number of unique users in the dataset is ______`': unique_users,
    '`The number of unique articles on the IBM platform`': total_articles
}

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

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

1.1.2 Part II: Rank-Based Recommendations

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

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

[27]:
```
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_articles = list(df['title'].value_counts().iloc[:n].index)
    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

    '''
    top_articles = list(df['article_id'].value_counts().iloc[:n].index)
    top_articles = [str(x) for x in top_articles]

    return top_articles # Return the top article ids
```

[28]:
```
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']
```

[29]:
```
# 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.

[30]:
```
# create the user-article matrix with 1's and 0's

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

    OUTPUT:
    user_item - user item matrix

    Description:
    Return a matrix with user ids as rows and article ids on the columns with 1␣
 ↪values where a user interacted with
    an article and a 0 otherwise
    '''
    # Fill in the function here
    user_item = df.drop_duplicates()
    user_item = user_item.set_index('user_id')
    user_item = pd.get_dummies(user_item['article_id']).groupby('user_id').
 ↪sum().clip(upper=1)
    user_item = user_item.astype(int)
    return user_item # return the user_item matrix

user_item = create_user_item_matrix(df)
```

[31]:
```
## Tests: You should just need to run this cell.  Don't change the code.
assert user_item.shape[0] == 5149, "Oops!  The number of users in the␣
 ↪user-article matrix doesn't look right."
assert user_item.shape[1] == 714, "Oops!  The number of articles in the␣
 ↪user-article matrix doesn't look right."
assert user_item.sum(axis=1)[1] == 36, "Oops!  The number of articles seen by␣
 ↪user 1 doesn't look right."
print("You have passed our quick tests!  Please proceed!")
```

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

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

Use the tests to test your function.

[32]:
```python
def find_similar_users(user_id, user_item=user_item):
    '''
    INPUT:
    user_id - (int) a user_id
    user_item - (pandas dataframe) matrix of users by articles:
                1's when a user has interacted with an article, 0 otherwise

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

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

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

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

[33]:
```python
# Do a spot check of your function
print("The 10 most similar users to user 1 are: {}".
 ↪format(find_similar_users(1)[:10]))
print("The 5 most similar users to user 3933 are: {}".
 ↪format(find_similar_users(3933)[:5]))
print("The 3 most similar users to user 46 are: {}".
 ↪format(find_similar_users(46)[:3]))
```

```
The 10 most similar users to user 1 are: [3933, 23, 3782, 203, 4459, 3870, 131,
46, 4201, 395]
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.

[34]:
```python
df.copy()
```

```
[34]:        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
      …             …                                                   …      …
      45988      1324.0                  ibm watson facebook posts for 2015     5146
      45989       142.0  neural networks for beginners: popular types a…     5146
      45990       233.0     bayesian nonparametric models – stats and bots     5147
      45991      1160.0       analyze accident reports on amazon emr spark     5148
      45992        16.0  higher-order logistic regression for large dat…     5149

      [45993 rows x 3 columns]
```

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

          OUTPUT:
          article_names - (list) a list of article names associated with the list of␣
       ↪article ids
                          (this is identified by the title column)
          '''
          article_names = df.copy().drop_duplicates(subset=['article_id']).
       ↪set_index('article_id')
          article_names = [article_names.loc[float(a_id)]['title'] for a_id in␣
       ↪article_ids]
          return article_names # Return the article names associated with list of␣
       ↪article ids


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

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

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


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

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

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

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

    For the user where the number of recommended articles starts below m
    and ends exceeding m, the last items are chosen arbitrarily

    '''
    user_articles, _ = get_user_articles(user_id)
    user_articles = set(user_articles)
    similar_users = find_similar_users(user_id)
    recs = set()
    for user in similar_users:
            # get article ids for selected similar user
            rec_user_articles, _  = get_user_articles(user)
            # get only the articles not seen articles
            new_rec = list(set(rec_user_articles) - user_articles - recs)
            empty_slots = m - len(recs)
            if empty_slots > 0:
                recs.update(list(new_rec)[:empty_slots])
            else:
                break

    recs = list(recs)
    return recs # return your recommendations for this user_id
```

[36]:
```
# Check Results
get_article_names(user_user_recs(1, 10)) # Return 10 recommendations for user 1
```

[36]:
```
['a dynamic duo - inside machine learning - medium',
 'model bike sharing data with spss',
 'intents & examples for ibm watson conversation',
 'get started with streams designer by following this roadmap',
 'from spark ml model to online scoring with scala',
 'insights from new york car accident reports',
 'challenges in deep learning',
 'dsx: hybrid mode',
 'generalization in deep learning',
 'overlapping co-cluster recommendation algorithm (ocular)']
```

[37]:
```
# Test your functions here - No need to change this code - just run this cell
assert set(get_article_names(['1024.0', '1176.0', '1305.0', '1314.0', '1422.0',␣
 ↪'1427.0'])) == set(['using deep learning to reconstruct high-resolution␣
 ↪audio', 'build a python app on the streaming analytics service', 'gosales␣
 ↪transactions for naive bayes model', 'healthcare python streaming␣
 ↪application demo', 'use r dataframes & ibm watson natural language␣
 ↪understanding', 'use xgboost, scikit-learn & ibm watson machine learning␣
 ↪apis']), "Oops! Your the get_article_names function doesn't work quite how␣
 ↪we expect."
assert set(get_article_names(['1320.0', '232.0', '844.0'])) == set(['housing␣
 ↪(2015): united states demographic measures','self-service data preparation␣
 ↪with ibm data refinery','use the cloudant-spark connector in python␣
 ↪notebook']), "Oops! Your the get_article_names function doesn't work quite␣
 ↪how we expect."
assert set(get_user_articles(20)[0]) == set(['1320.0', '232.0', '844.0'])
assert set(get_user_articles(20)[1]) == set(['housing (2015): united states␣
 ↪demographic measures', 'self-service data preparation with ibm data␣
 ↪refinery','use the cloudant-spark connector in python notebook'])
assert set(get_user_articles(2)[0]) == set(['1024.0', '1176.0', '1305.0', '1314.
 ↪0', '1422.0', '1427.0'])
assert set(get_user_articles(2)[1]) == set(['using deep learning to reconstruct␣
 ↪high-resolution audio', 'build a python app on the streaming analytics␣
 ↪service', 'gosales transactions for naive bayes model', 'healthcare python␣
 ↪streaming application demo', 'use r dataframes & ibm watson natural language␣
 ↪understanding', 'use xgboost, scikit-learn & ibm watson machine learning␣
 ↪apis'])
print("If this is all you see, you passed all of our tests!  Nice job!")
```

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

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

- Instead of arbitrarily choosing when we obtain users who are all the same closeness to a given user - choose the users that have the most total article interactions before choosing those with fewer article interactions.

- Instead of arbitrarily choosing articles from the user where the number of recommended articles starts below m and ends exceeding m, choose articles with the articles with the most total interactions before choosing those with fewer total interactions. This ranking should be what would be obtained from the **top_articles** function you wrote earlier.

[38]:
```python
def get_top_sorted_users(user_id, df=df, user_item=user_item):
    '''
    INPUT:
    user_id - (int)
    df - (pandas dataframe) df as defined at the top of the notebook
    user_item - (pandas dataframe) matrix of users by articles:
            1's when a user has interacted with an article, 0 otherwise


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

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

    '''
    most_similar_users = user_item.drop(user_id).dot(user_item.loc[user_id])
    most_similar_users = most_similar_users.to_frame(name='similarity')
    interactions = df.groupby('user_id')['article_id'].count().
 ↪to_frame(name='num_interactions')
    neighbors_df = most_similar_users.merge(interactions, on='user_id')
    neighbors_df = neighbors_df.
 ↪sort_values(by=['similarity','num_interactions'], ascending=[False, False])
    neighbors_df.reset_index(inplace=True)
    neighbors_df.rename(columns={'user_id': 'neighbor_id'}, inplace=True)

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

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

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

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

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

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

    '''
    # get articles the user interacted with
    user_articles, _ = get_user_articles(user_id)
    user_articles = set(user_articles)

    # get top most similar users
    top_similar_users = get_top_sorted_users(user_id)

    # get top articles in terms of interactions
    top_articles = get_top_article_ids(len(df))

    recs = []

    for user in top_similar_users['neighbor_id']:
        # get article ids for selected similar user
        rec_user_articles, _  = get_user_articles(user)
        # get only the articles not seen articles
        new_rec = list(set(rec_user_articles) - user_articles - set(recs))
        # sort according to total article interactions
        new_rec_sorted = [art for art in top_articles if art in new_rec]
        empty_slots = m - len(recs)
        if empty_slots > 0:
            recs.extend(new_rec_sorted[:empty_slots])
        else:
            break


    rec_names = get_article_names(recs)

    return recs, rec_names
```

[39]:
```
# 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:
['1330.0', '1427.0', '1364.0', '1170.0', '1162.0', '1304.0', '1351.0', '1160.0',
'1354.0', '1368.0']

The top 10 recommendations for user 20 are the following article names:
['insights from new york car accident reports', 'use xgboost, scikit-learn & ibm
watson machine learning apis', 'predicting churn with the spss random tree
algorithm', 'apache spark lab, part 1: basic concepts', 'analyze energy
consumption in buildings', 'gosales transactions for logistic regression model',
'model bike sharing data with spss', 'analyze accident reports on amazon emr
spark', 'movie recommender system with spark machine learning', 'putting a human
face on machine learning']
```

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.

[40]:
```
### Tests with a dictionary of results

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

[41]:
```
## 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 we are dealing with a new user, there is no recorded activity, and therefore the article interaction data for the user is empty. Hence, we cannot rely on collaborative filtering to provide article recommendations to the user as it relies upon the similarity between users' data (article interactions). So we naturally shift to rank-based recommendations, based on global article popularity amongst users. Furthermore, if the platforms conduct an onboarding to the user to collect their preferences, we can exploit content-based recommendation methods, which try to recommend content similar to the user's recorded preferences.

`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.

[42]:
```
new_user = '0.0'

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

[43]:
```
assert set(new_user_recs) == set(['1314.0','1429.0','1293.0','1427.0','1162.
↪0','1364.0','1304.0','1170.0','1431.0','1330.0']), "Oops!  It makes sense␣
↪that in this case we would want to recommend the most popular articles,␣
↪because we don't know anything about these users."

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

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

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

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

`1.` Use the function body below to create a content based recommender. Since there isn't one right answer for this recommendation tactic, no test functions are provided. Feel free to change the function inputs if you decide you want to try a method that requires more input values. The input values are currently set with one idea in mind that you may use to make content based recommendations. One additional idea is that you might want to choose the most popular recommendations that meet your 'content criteria', but again, there is a lot of flexibility in how you might make these recommendations.

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

[44]:
```
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.

[45]:
```
# 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.

[46]:
```
# Load the matrix here
user_item_matrix = pd.read_pickle('user_item_matrix.p')
```

[47]:
```
# quick look at the matrix
user_item_matrix.head()
```

[47]:
```
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.

[48]:
```python
# 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
```

[49]:
```python
user_item_matrix.isna().sum().sum()
```

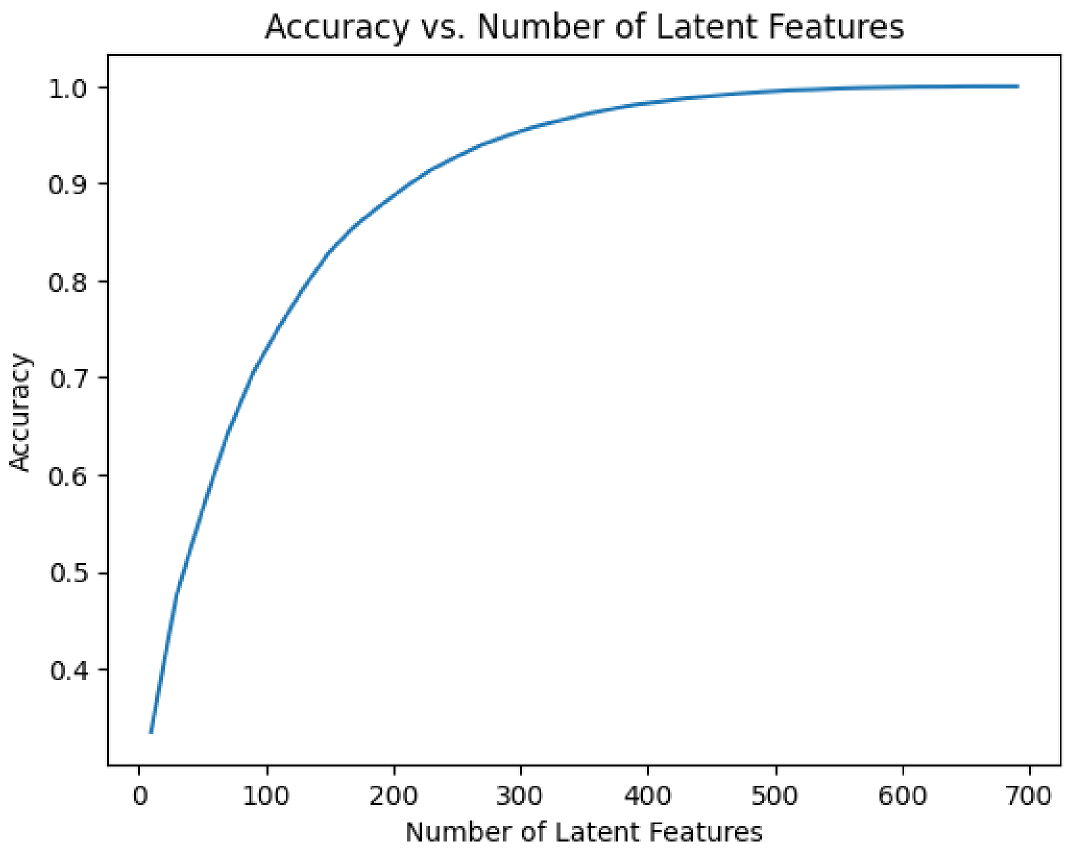
[49]:
```
0
```

Since there are no NaN values in our dataframe, we can use the built-in numpy svd without the need to use FunkSVD. This is advantageous, because we avoid the incurred computation complexity burden, avoid possible overfitting, and retain interpretability.

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.

[50]:
```python
import matplotlib.pyplot as plt
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?

[51]:
```python
df_train = df.head(40000)
df_test = df.tail(5993)

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

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

    '''
    # Your code here
    user_item_train = create_user_item_matrix(df_train)
    user_item_test = create_user_item_matrix(df_test)
    test_idx = user_item_test.index.values
    test_arts = user_item_test.columns.values

    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)
```

[52]:
```python
len(list(set(test_idx)))
```

[52]: 682

[53]:
```
user_prediction_test = len(list(set(user_item_train.index.
 ↪values)&set(test_idx)))
print('user_prediction_test',user_prediction_test)
user_noprediction_test = len(list(set(test_idx) - set(user_item_train.index.
 ↪values)))
print('user_no_prediction_test',user_noprediction_test)
article_prediction_test = len(list(set(user_item_train.columns.
 ↪values)&set(test_arts)))
print('article_prediction_test',article_prediction_test)
article_noprediction_test = len(list(set(test_arts)-set(user_item_train.columns.
 ↪values)))
print('article_prediction_test',article_noprediction_test)
```

```
user_prediction_test 20
user_no_prediction_test 662
article_prediction_test 574
article_prediction_test 0
```

[54]:
```
# Replace the values in the dictionary below
a = 662
b = 574
c = 20
d = 0

# had to change articles to movies in order to pass the test
sol_4_dict = {
    'How many users can we make predictions for in the test set?': c,
    'How many users in the test set are we not able to make predictions for␣
 ↪because of the cold start problem?': a,
    'How many movies can we make predictions for in the test set?': b,
    'How many movies in the test set are we not able to make predictions for␣
 ↪because of the cold start problem?': d
}

t.sol_4_test(sol_4_dict)
```

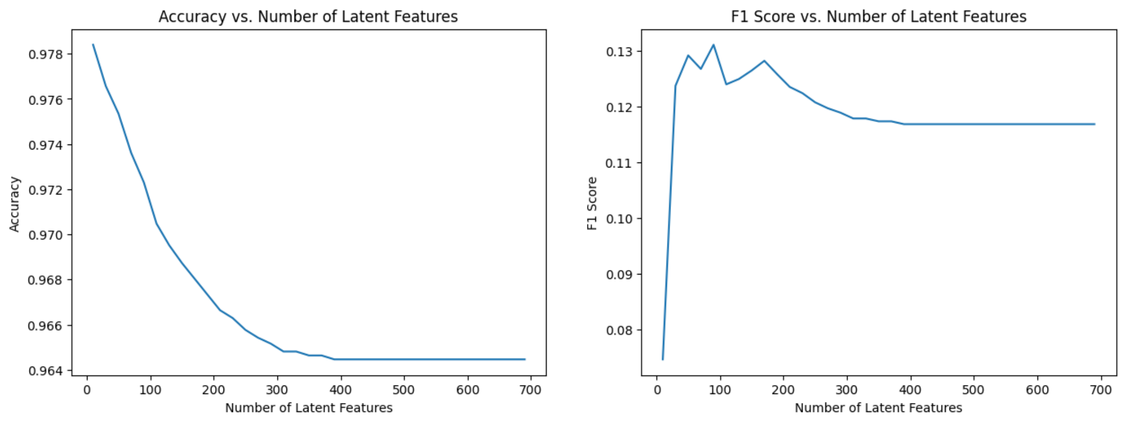
```
Awesome job!  That's right!  All of the test movies are in the training data,
but there are only 20 test users that were also in the training set.  All of the
other users that are in the test set we have no data on.  Therefore, we cannot
make predictions for these users using SVD.
```

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

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

[55]:
```
# fit SVD on the user_item_train matrix
u_train, s_train, vt_train = np.linalg.svd(user_item_train) # fit svd similar␣
 ↪to above then use the cells below
```

[67]:
```
from sklearn.metrics import precision_recall_fscore_support

num_latent_feats = np.arange(10,700+10,20)
sum_errs = []
precision_scores = []
recall_scores = []
f1_scores = []

joint_users_train_test = list(set(user_item_train.index)&set(test_idx))
joint_articles_train_test = list(set(user_item_train.columns)&set(test_arts))

joint_user_item_matrix = user_item_test.
 ↪loc[joint_users_train_test,joint_articles_train_test]
non_zero_idx = np.nonzero(joint_user_item_matrix.values)
actual_values = joint_user_item_matrix.values[non_zero_idx]

user_indices = user_item_train.index.get_indexer(joint_users_train_test)
article_indices = user_item_train.columns.get_indexer(joint_articles_train_test)

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

    # get the only the relevant values for u, s, vt
    u_joint = u_train_new[user_indices, :]
    vt_joint = vt_train_new[:, article_indices]
    s_joint = s_train_new

    # take dot product
    user_item_est_test = np.around(np.dot(np.dot(u_joint, s_joint), vt_joint))


    # compute error for each prediction to actual value
    diffs = np.subtract(joint_user_item_matrix.values, user_item_est_test)

    # compute precision and recall
    y_true = np.array(joint_user_item_matrix.values).flatten()
    y_pred = np.array(user_item_est_test).flatten()
    precision, recall, f1, _ = precision_recall_fscore_support(y_true, y_pred,␣
↪average='binary')

    precision_scores.append(precision)
    recall_scores.append(recall)
    f1_scores.append(f1)

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


# create two subplots one for error and one for the f1 score
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15, 5))

# plot the error
ax1.plot(num_latent_feats,1 - np.array(sum_errs)/joint_user_item_matrix.size)
ax1.set_title('Accuracy vs. Number of Latent Features')
ax1.set_xlabel('Number of Latent Features')
ax1.set_ylabel('Accuracy')

# plot the f1 score
ax2.plot(num_latent_feats, f1_scores, label='F1 Score');
ax2.set_title('F1 Score vs. Number of Latent Features');
ax2.set_xlabel('Number of Latent Features');
ax2.set_ylabel('F1 Score');
```



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 figures above, we can see a peak in the F1 score of 0.131 for 90 latent features. The F1 score is relatively low due to poor recall from the matrix factorization method. We chose to adopt the F1 score as it helps to strike a balance between precision and recall during our performance assessment, as it is the harmonic mean between precision and recall. This performance could be due to the sparsity in the user-item matrix, as the users only interact with a smaller subset of articles.

From the accuracy figure, we are starting with a 97.84% accuracy at 10 latent features, the model's accuracy drops and plateaus at 96.46% by 350 latent features. This trend supports the idea of a simpler, more streamlined model for balanced performance.

From both figures, we can see that while increasing the number of latent features generally improves performance, diminishing returns are noticeable after a certain number of latent features. This suggests overfitting and the need for caution when adding more features.

Yet, evaluating the recommendation system solely based on F1 score and accuracy could be misleading. Three key challenges in our evaluation approach are:

1. The Cold Start Problem: Using offline evaluation metrics may be inaccurate due to the lack of sufficient interactions between new users and articles.
2. Data Drift: As user preferences and trends evolve, the pre-computed user-item matrix only represents a snapshot in time of the user's preferences.
3. Complex Real User Behaviour: Offline evaluation metrics may not fully capture complex user behavior, which can be influenced by a wide range of endogenous and exogenous factors, such as the time of day or the user's mood.

Therefore, we need to test our recommendation systems using real-world context. We can use A/B testing, which divides our user base into two distinct groups. The first group interacts with the new recommendation system (experimental group), while the second group is left unchanged (control group). Then, we regularly acquire feedback from both groups, either directly (surveys) or indirectly (click-through rate, conversion rate, etc). Finally, we measure the statistical significance (statistical tests) and practical significance (if there is enough added value to incur the cost of operating the new recommendation system).

### 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!

```
[47]: from subprocess import call
      call(['python', '-m', 'nbconvert', '--to', 'pdf', 'Recommendations_with_IBM.
       ↪ipynb'])
```

```
[NbConvertApp] Converting notebook Recommendations_with_IBM.ipynb to pdf
[NbConvertApp] Support files will be in Recommendations_with_IBM_files/
[NbConvertApp] Making directory ./Recommendations_with_IBM_files
[NbConvertApp] Writing 120208 bytes to notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', 'notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 213266 bytes to Recommendations_with_IBM.pdf
```

```
[47]: 0
```

```
[ ]:
```