HW1_Q1.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all

[32]

```python
import sys
print(sys.version)
```

```
3.11.13 (main, Jun  4 2025, 08:57:29) [GCC 11.4.0]
```

In my Spark pipeline, Spark is leveraged for distributed processing of the social network data. It uses RDD transformations like flatMap and reduceByKey to efficiently identify all pairs of mutual friends. Subsequently, Spark performs joins and filters to exclude existing friendships, retaining only potential recommendations. Finally, it groups these recommendations by user and sorts them to present the top suggestions.

[33]

```python
# !pip install pyspark==3.3.2
# !pip install pyspark
!pip install -U -q PyDrive
!pip uninstall -y PyDrive
!pip install -U PyDrive2
!apt install openjdk-8-jdk-headless -qq
import os
# os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
```

```
  Preparing metadata (setup.py) ... done
  Building wheel for PyDrive (setup.py) ... done
Found existing installation: PyDrive 1.3.1
Uninstalling PyDrive-1.3.1:
  Successfully uninstalled PyDrive-1.3.1
Requirement already satisfied: PyDrive2 in /usr/local/lib/python3.11/dist-packages (1.21.3)
Requirement already satisfied: google-api-python-client>=1.12.5 in /usr/local/lib/python3.11/dist-packages (from PyDrive2) (2.176.0)
Requirement already satisfied: oauth2client>=4.0.0 in /usr/local/lib/python3.11/dist-packages (from PyDrive2) (4.1.3)
Requirement already satisfied: PyYAML>=3.0 in /usr/local/lib/python3.11/dist-packages (from PyDrive2) (6.0.2)
Requirement already satisfied: cryptography<44 in /usr/local/lib/python3.11/dist-packages (from PyDrive2) (43.0.3)
Requirement already satisfied: pyOpenSSL<=24.2.1,>=19.1.0 in /usr/local/lib/python3.11/dist-packages (from PyDrive2) (24.2.1)
Requirement already satisfied: cffi>=1.12 in /usr/local/lib/python3.11/dist-packages (from cryptography<44->PyDrive2) (1.17.1)
Requirement already satisfied: httplib2<1.0.0,>=0.19.0 in /usr/local/lib/python3.11/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (0.22.0)
Requirement already satisfied: google-auth!=2.24.0,!=2.25.0,<3.0.0,>=1.32.0 in /usr/local/lib/python3.11/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (2.38.0)
Requirement already satisfied: google-auth-httplib2<1.0.0,>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (0.2.0)
Requirement already satisfied: google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5 in /usr/local/lib/python3.11/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (2.25.1)
Requirement already satisfied: uritemplate<5,>=3.0.1 in /usr/local/lib/python3.11/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (4.2.0)
Requirement already satisfied: pyasn1>=0.1.7 in /usr/local/lib/python3.11/dist-packages (from oauth2client>=4.0.0->PyDrive2) (0.6.1)
Requirement already satisfied: pyasn1-modules>=0.0.5 in /usr/local/lib/python3.11/dist-packages (from oauth2client>=4.0.0->PyDrive2) (0.4.2)
Requirement already satisfied: rsa>=3.1.4 in /usr/local/lib/python3.11/dist-packages (from oauth2client>=4.0.0->PyDrive2) (4.9.1)
Requirement already satisfied: six>=1.6.1 in /usr/local/lib/python3.11/dist-packages (from oauth2client>=4.0.0->PyDrive2) (1.17.0)
Requirement already satisfied: pycparser in /usr/local/lib/python3.11/dist-packages (from cffi>=1.12->cryptography<44->PyDrive2) (2.22)
Requirement already satisfied: googleapis-common-protos<2.0.0,>=1.56.2 in /usr/local/lib/python3.11/dist-packages (from google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-python-
Requirement already satisfied: protobuf!=3.20.0,!=3.20.1,!=4.21.0,!=4.21.1,!=4.21.2,!=4.21.3,!=4.21.4,!=4.21.5,<7.0.0,>=3.19.5 in /usr/local/lib/python3.11/dist-packages (from google-api-core!=2.0.*,!=2
Requirement already satisfied: proto-plus<2.0.0,>=1.22.3 in /usr/local/lib/python3.11/dist-packages (from google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-python-client>=1.12.5
Requirement already satisfied: requests<3.0.0,>=2.18.0 in /usr/local/lib/python3.11/dist-packages (from google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-python-client>=1.12.5->
Requirement already satisfied: cachetools<6.0,>=2.0.0 in /usr/local/lib/python3.11/dist-packages (from google-auth!=2.24.0,!=2.25.0,<3.0.0,>=1.32.0->google-api-python-client>=1.12.5->PyDrive2) (5.5.2)
Requirement already satisfied: pyparsing!=3.0.0,!=3.0.1,!=3.0.2,!=3.0.3,<4,>=2.4.2 in /usr/local/lib/python3.11/dist-packages (from httplib2<1.0.0,>=0.19.0->google-api-python-client>=1.12.5->PyDrive2) (
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.18.0->google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-a
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.18.0->google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-python-cl
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.18.0->google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-pyt
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.18.0->google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-pyt
The following additional packages will be installed:
  libxtst6 openjdk-8-jre-headless
Suggested packages:
  openjdk-8-demo openjdk-8-source libnss-mdns fonts-dejavu-extra fonts-nanum
  fonts-ipafont-gothic fonts-ipafont-mincho fonts-wqy-microhei
  fonts-wqy-zenhei fonts-indic
The following NEW packages will be installed:
  libxtst6 openjdk-8-jdk-headless openjdk-8-jre-headless
0 upgraded, 3 newly installed, 0 to remove and 35 not upgraded.
Need to get 39.7 MB/39.7 MB of archives.
After this operation, 144 MB of additional disk space will be used.
E: Failed to fetch http://security.ubuntu.com/ubuntu/pool/universe/o/openjdk-8/openjdk-8-jre-headless_8u452-ga%7eus1-0ubuntu1%7e22.04_amd64.deb  404  Not Found [IP: 91.189.92.23 80]
E: Failed to fetch http://security.ubuntu.com/ubuntu/pool/universe/o/openjdk-8/openjdk-8-jdk-headless_8u452-ga%7eus1-0ubuntu1%7e22.04_amd64.deb  404  Not Found [IP: 91.189.92.23 80]
E: Unable to fetch some archives, maybe run apt-get update or try with --fix-missing?
```

[34]

```python
# !update-alternatives --set java /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/java
!java -version
```

```
openjdk version "11.0.27" 2025-04-15
OpenJDK Runtime Environment (build 11.0.27+6-post-Ubuntu-0ubuntu122.04)
OpenJDK 64-Bit Server VM (build 11.0.27+6-post-Ubuntu-0ubuntu122.04, mixed mode, sharing)
```

[35]

```python
from pydrive2.auth import GoogleAuth
from pydrive2.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials

# Authenticate and create the PyDrive client
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

[36]

```python
file_id = '1heo2wadgXyS54zh-XNukAyq8saDzWS6H'

# Download
downloaded = drive.CreateFile({'id': file_id})
downloaded.GetContentFile('soc-LiveJournal1Adj.txt')
```

[37]

```python
from pyspark import SparkContext
from itertools import combinations

from pyspark.sql import SparkSession

spark = (
    SparkSession.builder
        .appName("ColabSpark")
        .master("local[*]")
        .config("spark.ui.showConsoleProgress", "true")
        .config("spark.driver.memory", "4g")
        .getOrCreate()
)

spark

```

SparkSession - in-memory

SparkContext

Spark UI

Version
v3.5.1
Master
local[*]
AppName
pyspark-shell

[38]

```python
from pyspark import SparkConf, SparkContext

conf = SparkConf().setAppName("PeopleYouMightKnow")
```

```
conf = SparkConf().setAppName("PeopleYouMightKnow")

sc = SparkContext.getOrCreate()

lines = sc.textFile("soc-LiveJournal1Adj.txt")
```

[39] 3m

```
# sc.parallelize([1, 2, 3]).map(lambda x: x * 10).collect()

lines.take(5)
```

```
['0\t1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69

'1\t0,5,20,135,2409,8715,8932,10623,12347,12846,13840,13845,14005,20075,21556,22939,23520,28193,29724,29791,29826,30691,31232,31435,32317,32489,34394,35589,35605,35606,35613,35633,35648,35678,38737,4344
 '2\t0,117,135,1220,2755,12453,24539,24714,41456,45046,49927,6893,13795,16659,32828,41878',
 '3\t0,12,41,55,1532,12636,13185,27552,38737',
 '4\t0,8,14,15,18,27,72,80,15326,19068,19079,24596,42697,46126,74,77,33269,38792,38822']
```

[40] 0s

```
users_friends = lines.map(
    lambda line: (
        int(line.split('\t')[0]),
        [int(f) for f in line.split('\t')[1].split(',') if f.strip() != ""]
        if len(line.split('\t')) > 1 else []
    )
)
```

[41] 0s

```
print("USERS_FRIENDS:", users_friends.take(5))
```

```
USERS_FRIENDS: [(0, [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48
```

[42] 0s

```
users_friends.take(5)
```

```
[(0,
  [1,
   2,
   3,
   4,
   5,
   6,
   7,
   8,
   9,
   10,
   11,
   12,
   13,
   14,
   15,
   16,
   17,
   18,
   19,
   20,
   21,
   22,
   23,
   24,
   25,
   26,
   27,
   28,
   29,
   30,
   31,
   32,
   33,
   34,
   35,
   36,
   37,
   38,
   39,
   40,
   41,
   42,
   43,
   44,
   45,
   46,
   47,
   48,
   49,
   50,
   51,
   52,
   53,
   54,
   55,
   56,
   57,
```

[43] 0s

```
for u, friends in users_friends.take(5):
    print(f"User {u} has friends: {friends}")
```

```
User 0 has friends: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48
User 1 has friends: [0, 5, 20, 135, 2409, 8715, 8932, 10623, 12347, 12846, 13840, 13845, 14005, 20075, 21556, 22939, 23520, 28193, 29724, 29791, 29826, 30691, 31232, 31435, 32317, 32489, 34394, 35589, 3
User 2 has friends: [0, 117, 135, 1220, 2755, 12453, 24539, 24714, 41456, 45046, 49927, 6893, 13795, 16659, 32828, 41878]
User 3 has friends: [0, 12, 41, 55, 1532, 12636, 13185, 27552, 38737]
User 4 has friends: [0, 8, 14, 15, 18, 27, 72, 80, 15326, 19068, 19079, 24596, 42697, 46126, 74, 77, 33269, 38792, 38822]
```

[44] 0s

```
def mutual_pairs(user, friends):
    out = []
    for f1, f2 in combinations(friends, 2):
        out.append(((f1, f2), 1))
        out.append(((f2, f1), 1))
    return out
```

[45] 0s

```
pairs = users_friends.flatMap(lambda x: mutual_pairs(x[0], x[1]))
```

[46] 0s

```
pairs.take(50)
```

```
[((1, 2), 1),
 ((2, 1), 1),
 ((1, 3), 1),
 ((3, 1), 1),
 ((1, 4), 1),
 ((4, 1), 1),
 ((1, 5), 1),
 ((5, 1), 1),
 ((1, 6), 1),
 ((6, 1), 1),
 ((1, 7), 1),
 ((7, 1), 1),
 ((1, 8), 1),
 ((8, 1), 1),
 ((1, 9), 1),
 ((9, 1), 1),
 ((1, 10), 1),
```

```
((10, 1), 1),
((1, 11), 1),
((11, 1), 1),
((1, 12), 1),
((12, 1), 1),
((1, 13), 1),
((13, 1), 1),
((1, 14), 1),
((14, 1), 1),
((1, 15), 1),
((15, 1), 1),
((1, 16), 1),
((16, 1), 1),
((1, 17), 1),
((17, 1), 1),
((1, 18), 1),
((18, 1), 1),
((1, 19), 1),
((19, 1), 1),
((1, 20), 1),
((20, 1), 1),
((1, 21), 1),
((21, 1), 1),
((1, 22), 1),
((22, 1), 1),
((1, 23), 1),
((23, 1), 1),
((1, 24), 1),
((24, 1), 1),
((1, 25), 1),
((25, 1), 1),
((1, 26), 1),
((26, 1), 1)]
```

[47] 0s

```
pairs = pairs.filter(lambda x: x[0][0] != x[0][1])
```

[48] 0s

```
print("pairs (filtered):", pairs.take(10))
```

```
pairs (filtered): [((1, 2), 1), ((2, 1), 1), ((1, 3), 1), ((3, 1), 1), ((1, 4), 1), ((4, 1), 1), ((1, 5), 1), ((5, 1), 1), ((1, 6), 1), ((6, 1), 1)]
```

[49] 0s

```
mutual_counts = pairs.reduceByKey(lambda a, b: a + b)
```

Double-click (or enter) to edit

[50] 0s

```
# print("mutual_counts:", mutual_counts.take(2))
```

[51] 0s

```
edges = users_friends.flatMap(
    lambda x: [((x[0], f), 1) for f in x[1]] +
              [((f, x[0]), 1) for f in x[1]]
)
```

[52] 0s

```
# print("edges:", edges.take(3))
```

[53] 0s

```
candidates = mutual_counts.leftOuterJoin(edges)
```

[54] 0s

```
filtered = candidates.filter(lambda x: x[1][1] is None).map(lambda x: (x[0], x[1][0]))
```

[55] 0s

```
# print("filtered:", filtered.take(3))
```

[56] 0s

```
user_recs = filtered.map(
    lambda x: (x[0][0], (x[0][1], x[1]))
).groupByKey()
```

[57] 0s

```
# print("user_recs:", user_recs.take(3))
```

[58] 0s

```

def top_10(recs):
    return sorted(list(recs), key=lambda x: (-x[1], x[0]))[:10]

final_recs = user_recs.mapValues(top_10)

```

[59] 0s

```
formatted = final_recs.map(
    lambda x: f"{x[0]}\t{','.join(str(rec[0]) for rec in x[1])}"
)

```

[60] 0s

```
# formatted.take(5)
```

[61] 20m

```
formatted.filter(lambda x: x.startswith("11\t")).take(1)

```

```
['11\t27552,7785,27573,27574,27589,27590,27600,27617,27620,27667']
```

[62] 10s

```
user_ids = [924, 8941, 8942, 9019, 9020, 9021, 9022, 9990, 9992, 9993]


filtered_recommendations = formatted.filter(lambda x: int(x.split('\t')[0]) in user_ids)

# Collected and printed the recommendations
for rec in filtered_recommendations.collect():
    print(rec)
```

```
9020    9021,9016,9017,9022,317,9023
924     439,2409,6995,11860,15416,43748,45881
9992    9987,9989,35667,9991
9021    9020,9016,9017,9022,317,9023
9993    9991,13134,13478,13877,34299,34485,34642,37941
8941    8943,8944,8940
8942    8939,8940,8943,8944
9022    9019,9020,9021,317,9016,9017,9023
9990    13134,13478,13877,34299,34485,34642,37941
9019    9022,317,9023
```

[64] 12s

```
formatted.saveAsTextFile("/content/output")
```

[ ]

Start coding or generate with AI.

Colab paid products - Cancel contracts here