HW1_Q2.ipynb

File Edit View Insert Runtime Tools Help

Share

Commands + Code + Text Run all

Reconnect

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

Spark Pipeline for Frequent Itemset Mining and Association Rule Generation

The notebook demonstrates a Spark pipeline for performing frequent itemset mining using the Apriori-like algorithm and generating association rules. The steps include:

1. **Data Loading and Spark Context Initialization**: Setting up the Spark environment and loading transaction data from a text file.
2. **Frequent Itemset Generation (L1, L2, L3)**: Identifying frequent individual items (L1), frequent item pairs (L2), and frequent item triples (L3) based on a defined support threshold.
3. **Association Rule Generation**: Calculating confidence for 2-item and 3-item association rules from the frequent itemsets.

```
!pip install -U -q PyDrive
!pip uninstall -y PyDrive
!pip install -U PyDrive2
!apt install openjdk-8-jdk-headless -qq
import os
```

```
  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/python3.12/dist-packages (1.21.3)
Requirement already satisfied: google-api-python-client>=1.12.5 in /usr/local/lib/python3.12/dist-packages (from PyDrive2) (2.188.0)
Requirement already satisfied: oauth2client>=4.0.0 in /usr/local/lib/python3.12/dist-packages (from PyDrive2) (4.1.3)
Requirement already satisfied: PyYAML>=3.0 in /usr/local/lib/python3.12/dist-packages (from PyDrive2) (6.0.3)
Requirement already satisfied: cryptography<44 in /usr/local/lib/python3.12/dist-packages (from PyDrive2) (43.0.3)
Requirement already satisfied: pyOpenSSL<=24.2.1,>=19.1.0 in /usr/local/lib/python3.12/dist-packages (from PyDrive2) (24.2.1)
Requirement already satisfied: cffi>=1.12 in /usr/local/lib/python3.12/dist-packages (from cryptography<44->PyDrive2) (2.0.0)
Requirement already satisfied: httplib2<1.0.0,>=0.19.0 in /usr/local/lib/python3.12/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (0.31.2)
Requirement already satisfied: google-auth!=2.24.0,!=2.25.0,<3.0.0,>=1.32.0 in /usr/local/lib/python3.12/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (2.47.0)
Requirement already satisfied: google-auth-httplib2<1.0.0,>=0.2.0 in /usr/local/lib/python3.12/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (0.3.0)
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/python3.12/dist-packages (from google-api-python-client>=1.12.5->PyDrive2) (2.29.0)
Requirement already satisfied: uritemplate<5,>=3.0.1 in /usr/local/lib/python3.12/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/python3.12/dist-packages (from oauth2client>=4.0.0->PyDrive2) (0.6.2)
Requirement already satisfied: pyasn1-modules>=0.0.5 in /usr/local/lib/python3.12/dist-packages (from oauth2client>=4.0.0->PyDrive2) (0.4.2)
Requirement already satisfied: rsa>=3.1.4 in /usr/local/lib/python3.12/dist-packages (from oauth2client>=4.0.0->PyDrive2) (4.9.1)
Requirement already satisfied: six>=1.6.1 in /usr/local/lib/python3.12/dist-packages (from oauth2client>=4.0.0->PyDrive2) (1.17.0)
Requirement already satisfied: pycparser in /usr/local/lib/python3.12/dist-packages (from cffi>=1.12->cryptography<44->PyDrive2) (3.0)
Requirement already satisfied: googleapis-common-protos<2.0.0,>=1.56.2 in /usr/local/lib/python3.12/dist-packages (from google-api-core!=2.0.*,!=2.1.*,!=2.2.*,!=2.3.0,<3.0.0,>=1.31.5->google-api-pyth
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/python3.12/dist-packages (from google-api-core!=2.0.*,
Requirement already satisfied: proto-plus<2.0.0,>=1.22.3 in /usr/local/lib/python3.12/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>=1.1
Requirement already satisfied: requests<3.0.0,>=2.18.0 in /usr/local/lib/python3.12/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.
Requirement already satisfied: pyparsing<4,>=3.1 in /usr/local/lib/python3.12/dist-packages (from httplib2<1.0.0,>=0.19.0->google-api-python-client>=1.12.5->PyDrive2) (3.3.2)
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/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
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/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-
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/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-p
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/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-p
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 2 not upgraded.
Need to get 39.7 MB of archives.
After this operation, 144 MB of additional disk space will be used.
Selecting previously unselected package libxtst6:amd64.
(Reading database ... 117540 files and directories currently installed.)
Preparing to unpack .../libxtst6_2%3a1.2.3-1build4_amd64.deb ...
Unpacking libxtst6:amd64 (2:1.2.3-1build4) ...
Selecting previously unselected package openjdk-8-jre-headless:amd64.
Preparing to unpack .../openjdk-8-jre-headless_8u472-ga-1~22.04_amd64.deb ...
Unpacking openjdk-8-jre-headless:amd64 (8u472-ga-1~22.04) ...
Selecting previously unselected package openjdk-8-jdk-headless:amd64.
Preparing to unpack .../openjdk-8-jdk-headless_8u472-ga-1~22.04_amd64.deb ...
Unpacking openjdk-8-jdk-headless:amd64 (8u472-ga-1~22.04) ...
Setting up libxtst6:amd64 (2:1.2.3-1build4) ...
Setting up openjdk-8-jre-headless:amd64 (8u472-ga-1~22.04) ...
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/jjs to provide /usr/bin/jjs (jjs) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/pack200 to provide /usr/bin/pack200 (pack200) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/rmid to provide /usr/bin/rmid (rmid) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/unpack200 to provide /usr/bin/unpack200 (unpack200) in auto mode
```

```

!java -version
```

```
openjdk version "17.0.17" 2025-10-21
OpenJDK Runtime Environment (build 17.0.17+10-Ubuntu-122.04)
OpenJDK 64-Bit Server VM (build 17.0.17+10-Ubuntu-122.04, mixed mode, sharing)
```

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


auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
```

```
file_id = '1IW4w8IPu7v_D-2kL3pWJ5NOCw-xbvKRf'


downloaded = drive.CreateFile({'id': file_id})
downloaded.GetContentFile('browsing.txt')
```

```
SUPPORT = 100
sc = SparkContext.getOrCreate()

```

```
from itertools import combinations


transactions = (
    sc.textFile("browsing.txt")
      .map(lambda line: set(line.strip().split()))
      .cache()
)

transactions.take(3)


```

```
[{'ELE17451', 'ELE89019', 'FRO11987', 'GRO99222', 'SNA90258'},
 {'ELE17451',
```

```
 'ELE26917',
 'ELE52966',
 'ELE91550',
 'FRO12685',
 'FRO84225',
 'FRO90334',
 'GRO12298',
 'GRO99222',
 'SNA11465',
 'SNA30755',
 'SNA80192'},
{'DAI22896', 'ELE17451', 'FRO86643', 'GRO73461', 'SNA99873'}]
```

```
[ ]
1
2 L1 = (
3     transactions
4     .flatMap(lambda basket: [(item, 1) for item in basket])
5     .reduceByKey(lambda a, b: a + b)
6     .filter(lambda x: x[1] >= SUPPORT)
7     .cache()
8 )
9
10 L1.take(4)
11
12
13
14
```

```
[('SNA90258', 550), ('FRO11987', 104), ('SNA11465', 142), ('SNA80192', 258)]
```

```
[ ]
1 L1_items = set(L1.map(lambda x: x[0]).collect())
2 print("Number of frequent items (L1):", len(L1_items))
```

```
Number of frequent items (L1): 647
```

```
[ ]
1 L1_broadcast = sc.broadcast(L1_items)
```

```
[ ]
1 L2 = (
2     transactions
3     .map(lambda basket: sorted([item for item in basket if item in L1_broadcast.value]))
4     .flatMap(lambda items: [((i, j), 1) for i, j in combinations(items, 2)])
5     .reduceByKey(lambda a, b: a + b)
6     .filter(lambda x: x[1] >= SUPPORT)
7     .cache()
8 )
9
10 print("Number of frequent pairs (L2):", L2.count())
11
```

```
Number of frequent pairs (L2): 1334
```

```
[ ]
1
2
3 pair_supports = dict(L2.collect())
4 pair_supports_broadcast = sc.broadcast(pair_supports)
5 pair_supports
```

```
{('ELE17451', 'GRO99222'): 148,
 ('ELE17451', 'SNA30755'): 111,
 ('DAI22896', 'ELE17451'): 193,
 ('ELE17451', 'SNA99873'): 270,
 ('DAI22177', 'ELE17451'): 203,
 ('ELE17451', 'ELE59935'): 181,
 ('DAI22177', 'ELE66810'): 105,
 ('DAI46755', 'FRO81176'): 148,
 ('ELE17451', 'ELE66810'): 154,
 ('ELE17451', 'GRO94758'): 227,
 ('ELE17451', 'SNA55952'): 123,
 ('ELE26917', 'GRO73461'): 255,
 ('GRO36567', 'GRO73461'): 117,
 ('DAI48891', 'GRO36567'): 128,
 ('FRO78087', 'GRO73461'): 192,
 ('ELE17451', 'FRO92261'): 127,
 ('ELE11111', 'ELE17451'): 121,
 ('DAI95741', 'ELE17451'): 102,
 ('DAI22896', 'GRO30386'): 102,
 ('ELE17451', 'GRO30386'): 468,
 ('FRO16142', 'GRO73461'): 197,
 ('FRO24098', 'GRO73461'): 112,
 ('DAI35347', 'ELE26917'): 111,
 ('DAI22896', 'FRO31317'): 167,
 ('DAI22896', 'SNA72163'): 227,
 ('DAI55911', 'ELE26917'): 113,
 ('DAI55911', 'GRO73461'): 116,
 ('ELE17451', 'FRO31317'): 359,
 ('ELE17451', 'SNA59903'): 351,
 ('ELE17451', 'SNA72163'): 272,
 ('SNA59903', 'SNA72163'): 310,
 ('DAI22177', 'FRO31317'): 160,
 ('ELE26917', 'GRO15017'): 164,
 ('ELE26917', 'GRO59710'): 132,
 ('GRO15017', 'GRO59710'): 112,
 ('GRO15017', 'GRO73461'): 288,
 ('GRO59710', 'GRO73461'): 305,
 ('DAI22177', 'SNA99873'): 148,
 ('ELE17451', 'ELE66600'): 168,
 ('DAI63921', 'DAI91290'): 120,
 ('DAI62779', 'ELE14480'): 230,
 ('DAI62779', 'FRO78087'): 482,
 ('DAI35347', 'DAI62779'): 226,
 ('DAI62779', 'DAI63921'): 353,
 ('DAI62779', 'ELE26917'): 650,
 ('DAI62779', 'SNA55762'): 593,
 ('DAI63921', 'ELE26917'): 120,
 ('DAI63921', 'SNA55762'): 120,
 ('ELE26917', 'SNA55762'): 138,
 ('FRO31317', 'SNA99873'): 277,
 ('ELE91337', 'SNA99873'): 125,
 ('DAI62779', 'SNA43319'): 132,
 ('ELE17451', 'ELE68605'): 115,
 ('DAI62779', 'FRO32293'): 299,
 ('DAI62779', 'FRO78994'): 201,
 ('DAI62779', 'GRO56989'): 155,
 ('DAI62779', 'SNA93730'): 154,
 ('ELE17451', 'SNA45677'): 304,
```

```
[ ]
1 L3 = (
2     transactions
3     .map(lambda basket: sorted([item for item in basket if item in L1_broadcast.value]))
4     .flatMap(lambda items: [((i, j, k), 1) for i, j, k in combinations(items, 3)])
5     .reduceByKey(lambda a, b: a + b)
6     .filter(lambda x: x[1] >= SUPPORT)
7 )
8
9 print("Number of frequent triples (L3):", L3.count())
```

```
Number of frequent triples (L3): 233
```

```
[ ]
1
2
3
4 item_supports = dict(L1.collect())
5 pair_supports = dict(L2.collect())
6
```

```
pair_rules = []

for (x, y), supp_xy in pair_supports.items():
    conf_x_y = supp_xy / item_supports[x]
    conf_y_x = supp_xy / item_supports[y]

    pair_rules.append((x, y, conf_x_y))
    pair_rules.append((y, x, conf_y_x))


pair_rules_sorted = sorted(
    pair_rules,
    key=lambda r: (-r[2], r[0])
)

print("\nTop 5 rules for 2(d):")
for r in pair_rules_sorted[:5]:
    print(f"{r[0]} -> {r[1]} : {r[2]:.4f}")

```

```
Top 5 rules for 2(d):
DAI93865 -> FRO40251 : 1.0000
GRO85051 -> FRO40251 : 0.9992
GRO38636 -> FRO40251 : 0.9907
ELE12951 -> FRO40251 : 0.9906
DAI88079 -> FRO40251 : 0.9867
```

[ ]

```


triple_supports = dict(L3.collect())

triple_rules = []

for (x, y, z), supp_xyz in triple_supports.items():
    supp_xy = pair_supports[(x, y)]
    supp_xz = pair_supports[(x, z)]
    supp_yz = pair_supports[(y, z)]

    triple_rules.append(((x, y), z, supp_xyz / supp_xy))
    triple_rules.append(((x, z), y, supp_xyz / supp_xz))
    triple_rules.append(((y, z), x, supp_xyz / supp_yz))


triple_rules_sorted = sorted(
    triple_rules,
    key=lambda r: (-r[2], r[0][0], r[0][1], r[1])
)

print("\nTop 5 rules for 2(e):")
for r in triple_rules_sorted[:5]:
    lhs = ",".join(r[0])
    print(f"{lhs} -> {r[1]} : {r[2]:.4f}")

```

```
Top 5 rules for 2(e):
DAI23334,ELE92920 -> DAI62779 : 1.0000
DAI31081,GRO85051 -> FRO40251 : 1.0000
DAI55911,GRO85051 -> FRO40251 : 1.0000
DAI62779,DAI88079 -> FRO40251 : 1.0000
DAI75645,GRO85051 -> FRO40251 : 1.0000
```

[ ]

```
Start coding or generate with AI.
```