06_tagging_history

February 26, 2026

1 06 — Article Tagging & Reading History

Steps covered: 1. Display synthetic users and interest dictionary 2. Tag CNH-PSX articles using keyword matching 3. Tag Pakistan News articles 4. Combine into a single article pool 5. Simulate reading history for each user → save reading_log.csv

[1]:
```
import sys
sys.path.append('../src')

import pandas as pd
import matplotlib.pyplot as plt

from user_profile import SYNTHETIC_USERS, SECTOR_KEYWORDS
from article_tagger import tag_articles, combine_tagged_datasets
from history_simulator import simulate_reading_history, get_reading_log

df_cnhpsx = pd.read_csv('../data/processed/cnhpsx_clean.csv')
df_news   = pd.read_csv('../data/processed/pakistan_news_clean.csv')

print('CNH-PSX:', df_cnhpsx.shape)
print('Pakistan News:', df_news.shape)
```

```
CNH-PSX: (8858, 5)
Pakistan News: (25912, 5)
```

1.1 Step 1 — Synthetic Users & Interest Dictionary

[2]:
```
print('Synthetic Users ')
for user in SYNTHETIC_USERS:
    print(f'  {user}')

print('\nInterest Dictionary (10 keywords per sector)')
for sector, keywords in SECTOR_KEYWORDS.items():
    print(f'\n{sector}:')
    for kw in keywords:
        print(f'  - {kw}')
```

```
Synthetic Users
  SyntheticUser(id=User_1, sector=Construction, sub_focus=cement-heavy,
```

```
clicks=0)
  SyntheticUser(id=User_2, sector=Construction, sub_focus=infrastructure-heavy,
clicks=0)
  SyntheticUser(id=User_3, sector=Banking, sub_focus=None, clicks=0)
  SyntheticUser(id=User_4, sector=Energy, sub_focus=None, clicks=0)

Interest Dictionary (10 keywords per sector)

Construction:
  - luck
  - dgkc
  - fccl
  - maple leaf cement
  - cement
  - concrete
  - construction
  - infrastructure
  - housing scheme
  - psdp
  - capacity expansion

Banking:
  - mebl
  - habib bank
  - hbl
  - united bank
  - ubl
  - national bank
  - nbp
  - state bank
  - sbp
  - policy rate
  - interest rate
  - net interest margin
  - deposits
  - dividend payout

Energy:
  - ogdc
  - oil and gas development
  - ppl
  - pakistan petroleum
  - pso
  - pakistan state oil
  - crude oil
  - exploration
  - refinery
  - gas production
```

```
  - fuel prices
  - circular debt
```

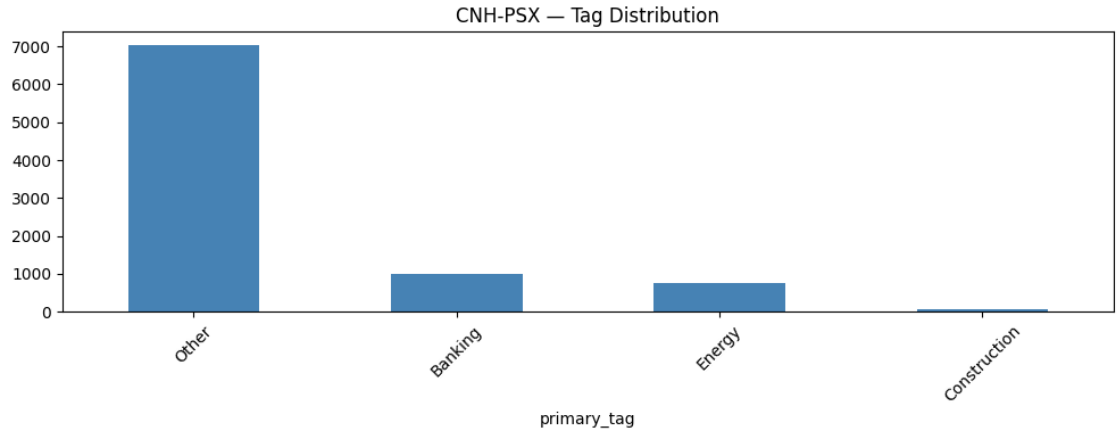
1.2 Step 2 — Tag CNH-PSX Articles

[3]:
```
# tag using raw headline - more keywords visible before cleaning
df_cnhpsx_tagged = tag_articles(df_cnhpsx, text_col='headline')

print('\nSample tagged articles:')
print(df_cnhpsx_tagged[df_cnhpsx_tagged['primary_tag'] != 'Other'][['headline',␣
 ↪'primary_tag', 'tags']].head(10))
```

```
[Tagger] 1826/8858 articles tagged (20.6%)
  Construction: 51
  Banking: 1011
  Energy: 764
  Other: 7032

Sample tagged articles:
                                                headline   primary_tag  \
57   ['Energy, cement shares help KSE gain 94 points']  Construction
92      ['Buying in cement averts big losses at KSE']  Construction
269     ['KSE further up on buying in cement stocks']  Construction
284  ['PSO sell off delay fuels bearish drive at KSE']        Energy
285  ['Cement, banking, oil scrips push KSE up by 1…  Construction
334                      ['KSE approves HBL listing']       Banking
377  ['KSE above 14,000 points on buying in cement,…  Construction
510  ['KSE gains 138 points on buying in banking, c…  Construction
552    ['KSE falls 861 points on interest rate hike']       Banking
589  ['KSE breaks decade low turnover record in abs…       Banking

              tags
57   [Construction]
92   [Construction]
269  [Construction]
284        [Energy]
285  [Construction]
334       [Banking]
377  [Construction]
510  [Construction]
552       [Banking]
589       [Banking]
```

[4]:
```
# Visualize tag distribution for CNH-PSX
df_cnhpsx_tagged['primary_tag'].value_counts().plot(
    kind='bar', figsize=(10, 4),
    title='CNH-PSX - Tag Distribution',
    color='steelblue'
```

```
)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

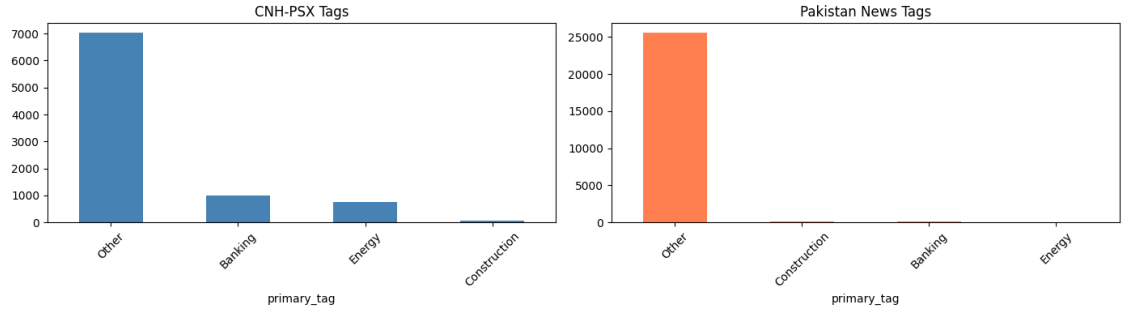


1.3 Step 3 — Tag Pakistan News Articles

[5]:
```
# Tag using full text (heading + excerpt) for better coverage
df_news_tagged = tag_articles(df_news, text_col='text_combined')

print('\nSample tagged articles:')
print(df_news_tagged[df_news_tagged['primary_tag'] != 'Other'][['heading',␣
 ↪'primary_tag']].head(10))
```

```
[Tagger] 325/25912 articles tagged (1.3%)
  Construction: 134
  Banking: 133
  Energy: 58
  Other: 25587

Sample tagged articles:
                                                heading   primary_tag
161          CCP launches probe into 62 housing schemes  Construction
181   30 fishing boats destroyed by sea current in G…  Construction
438                 4pc growth target to be missed: SBP       Banking
500   PSO dues, Wapda loans and Tapi on second ECC m…        Energy
611        Gas shortfall set to almost double next year        Energy
958   Housing society approved for PAF martyrs defra…  Construction
996     Huge sinkhole swallows bus, kills nine in China  Construction
1033  Thousands of litres of diesel leaks in Karachi…        Energy
1079         Govt raises Rs274.7bn from T-bill auction       Banking
1321         Ahsan aims to cement spot in Pakistan team  Construction
```

[6]:
```
# Side-by-side comparison
fig, axes = plt.subplots(1, 2, figsize=(14, 4))

df_cnhpsx_tagged['primary_tag'].value_counts().plot(
    kind='bar', ax=axes[0], title='CNH-PSX Tags', color='steelblue')
axes[0].tick_params(axis='x', rotation=45)

df_news_tagged['primary_tag'].value_counts().head(8).plot(
    kind='bar', ax=axes[1], title='Pakistan News Tags', color='coral')
axes[1].tick_params(axis='x', rotation=45)

plt.tight_layout()
plt.show()
```



1.4 Step 4 — Combine into Article Pool

[7]:
```
df_pool = combine_tagged_datasets(df_cnhpsx_tagged, df_news_tagged)

print('\nPool by source:')
print(df_pool['source'].value_counts())
print('\nPool by tag:')
print(df_pool['primary_tag'].value_counts())
```

```
[Pool] Combined pool: 34770 articles
primary_tag
Other           32619
Banking          1144
Energy            822
Construction      185

Pool by source:
source
pakistan_news    25912
cnhpsx            8858
Name: count, dtype: int64
```

```
Pool by tag:
primary_tag
Other           32619
Banking          1144
Energy            822
Construction      185
Name: count, dtype: int64
```

1.5 Step 5 — Simulate Reading History

[8]:
```
# 10 clicks per user, 10% noise from other sectors
users = simulate_reading_history(
    users=SYNTHETIC_USERS,
    df_tagged=df_pool,
    n_clicks=10,
    noise_ratio=0.1,
    seed=42
)
```

```
[History] User_1 (Construction / cement-heavy): 9 sector + 1 noise clicks
[History] User_2 (Construction / infrastructure-heavy): 9 sector + 1 noise
clicks
[History] User_3 (Banking / None): 9 sector + 1 noise clicks
[History] User_4 (Energy / None): 9 sector + 1 noise clicks
```

[9]:
```
# export and inspect reading log
reading_log = get_reading_log(users, df_pool)
print(reading_log.to_string())
reading_log.to_csv('../data/processed/reading_log.csv', index=False)
print('\nSaved to data/processed/reading_log.csv')
```

```
   user_id        sector  article_id
headline   primary_tag
0   User_1  Construction       13791
Pakistan  Construction
1   User_1  Construction       10547
WB-funded plan to improve city's municipal infrastructure launched  Construction
2   User_1  Construction          92
['Buying in cement averts big losses at KSE']  Construction
3   User_1  Construction       12703
Pakistan  Construction
4   User_1  Construction        2210
['KSE-100 ends flat cement remains upbeat']  Construction
5   User_1  Construction       32054
Pakistan  Construction
6   User_1  Construction        2164
['KSE index up 74 points amid recovery in cement, oil stocks']  Construction
7   User_1  Construction       29129
```

```
Pakistan  Construction
8   User_1  Construction        4023  ['Qatar says too early to exit OPEC oil
cuts as investment still low', 'Pso & Faw Motors China working to modernise Pso
fleet to support Cpec infrastructure']  Construction
9   User_1  Construction        5733
['Cotton trading surges']         Other
10  User_2  Construction       24996
Pakistan  Construction
11  User_2  Construction       18319
Pakistan         Other
12  User_2  Construction       29376
Newspaper  Construction
13  User_2  Construction       33497
Newspaper  Construction
14  User_2  Construction        7162
['Lucky Cement to award power plant contract next month']  Construction
15  User_2  Construction        9854
Huge sinkhole swallows bus, kills nine in China  Construction
16  User_2  Construction       21139
Pakistan  Construction
17  User_2  Construction       20687
Pakistan  Construction
18  User_2  Construction       34185
Newspaper  Construction
19  User_2  Construction       25350
Business  Construction
20  User_3       Banking       26729
Newspaper       Banking
21  User_3       Banking        7587
['NBP announces 20pc bonus, 65pc cash dividend']       Banking
22  User_3       Banking        8401
['NBP president, sugar mills body chief appear before NAB']       Banking
23  User_3       Banking        7599
['NBP to target agriculture, SMEs']       Banking
24  User_3       Banking        8739
['NBP disburses Rs6.8bln youth loans']       Banking
25  User_3       Banking        8101
['Major Companies Declare Results HBL earns profit of Rs7.30 billion']
Banking
26  User_3       Banking        9191
Polio cases for 2019 still surfacing, tally rises  to 128         Other
27  User_3       Banking        8356
['HBL profit falls to Rs34.206bln']       Banking
28  User_3       Banking        8461
['NBP, UBL collaborate']       Banking
29  User_3       Banking        7443
['SBP okays Standard Chartered, Union Bank merger']       Banking
30  User_4        Energy        4334
```

```
['Ogra awaits PSO's compliance report to set LNG price']        Energy
31  User_4        Energy        7315
['China starts work on oil refinery in Niger']        Energy
32  User_4        Energy        3265
['Circular debt haunts PSO', 'Govt to sell 21m PPL shares']        Energy
33  User_4        Energy        7374
['Byco Petroleum posts profit of Rs414.54mln']         Other
34  User_4        Energy        2787
['PSO declares Rs11 a share dividend']        Energy
35  User_4        Energy        2722
['PPL plans domestic, Yemen exploration']        Energy
36  User_4        Energy       21710
Newspaper        Energy
37  User_4        Energy        2837
['PSO buys up to 150,000T of gas oil']        Energy
38  User_4        Energy        3250
['Govt appoints Asim M Khan as PPL acting MD']        Energy
39  User_4        Energy        3408
['PR gets breather after payment to PSO for fuel', 'Decision on POL prices
today']        Energy

Saved to data/processed/reading_log.csv
```

[10]:
```
# save pool for next notebook
df_pool.to_csv('../data/processed/article_pool.csv', index=False)
print('Article pool saved to data/processed/article_pool.csv')
```

```
Article pool saved to data/processed/article_pool.csv
```

1.6 Conclusions

- ☐ CNH-PSX articles tagged: … / 8858
- ☐ Pakistan News articles tagged: … / 25912
- ☐ Combined pool size: …
- ☐ Reading log looks correct: yes/no
- ☐ Ready for 07_profile_vectors: yes/no