06_tagging_history

February 28, 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 importlib, sys

# Force reload of src modules
mods_to_remove = [k for k in sys.modules if k in ('article_tagger',␣
 ↪'user_profile', 'history_simulator')]
for mod in mods_to_remove:
    del sys.modules[mod]
```

[2]:
```
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

[3]:
```
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
```

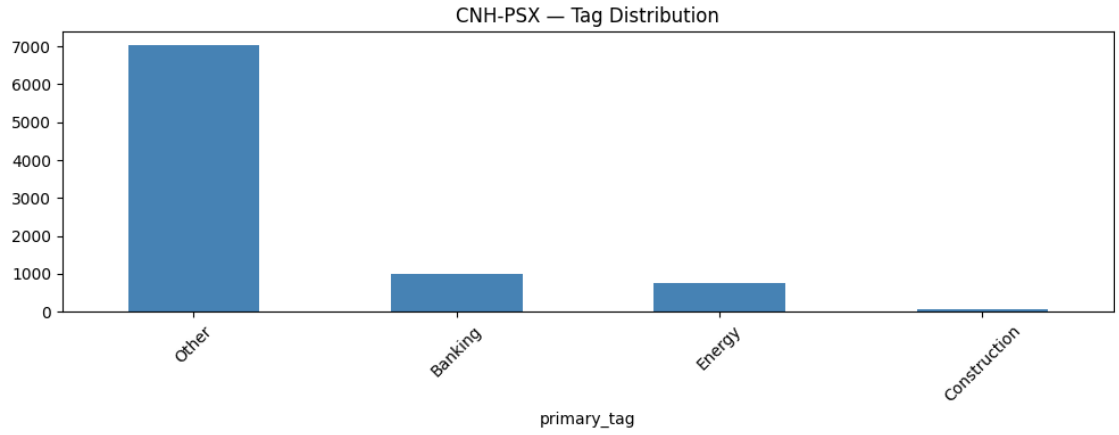
1.2 Step 2 — Tag CNH-PSX Articles

[4]:
```
# 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]
```

[5]:
```
# 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()
```



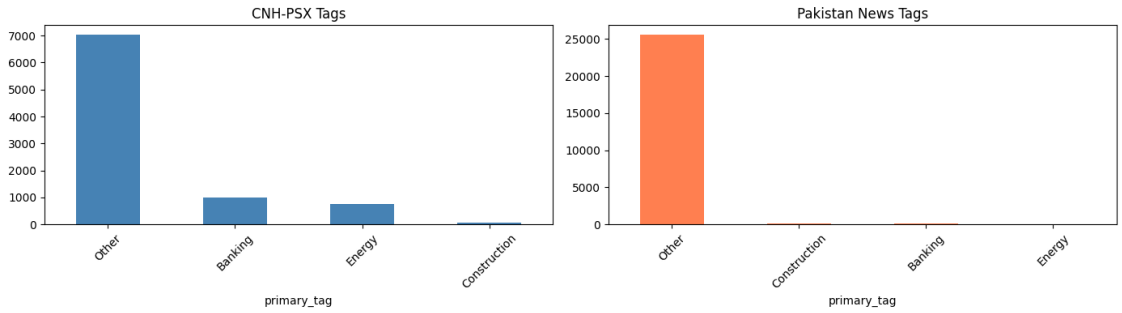
1.3 Step 3 — Tag Pakistan News Articles

[6]:
```
# 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
```

[7]:
```
# 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

[8]:
```
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: 14213 articles
primary_tag
Other           12358
Banking          1016
Energy            773
Construction       66

Pool by source:
source
cnhpsx           8858
pakistan_news    5355
Name: count, dtype: int64

Pool by tag:
primary_tag
Other           12358
Banking          1016
Energy            773
Construction       66
Name: count, dtype: int64
```

1.5 Step 5 — Simulate Reading History

[9]:
```
# 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
```

[10]:
```
# 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        1252
['KSE-100 gains 57 points on buying in cement shares']  Construction
1   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
2   User_1  Construction       14139
World  Construction
3   User_1  Construction       12197
World         Other
4   User_1  Construction         285
['Cement, banking, oil scrips push KSE up by 147 points', 'Range bound activity
witnessed at KSE']  Construction
5   User_1  Construction          92
['Buying in cement averts big losses at KSE']  Construction
6   User_1  Construction        1044
['Cement-led rally pushes KSE as CCI approves Diamer-Bhasha Dam']  Construction
7   User_1  Construction        7684
['State Bank revises infrastructure financing rules', 'Silkbank rights shares
subscription to be completed by March 2011']  Construction
8   User_1  Construction        7054
['Lucky Cement plans 660MW coal power plant in Karachi']  Construction
9   User_1  Construction        7113
['KCCI urges K-Electric to improve infrastructure']  Construction
10  User_2  Construction        2210
['KSE-100 ends flat cement remains upbeat']  Construction
11  User_2  Construction       12716
World  Construction
12  User_2  Construction        1668
['KSE index rises on back of investment in cement sector']  Construction
13  User_2  Construction        2574  ['China eyes infrastructure boost to
cushion growth as trade war escalates', 'Emerging markets, in trade war
crossfire, face deepening policy conundrum', 'US farm aid flimsy bandage on deep
trade war wound']  Construction
14  User_2  Construction        9816
Housing society approved for PAF martyrs defrauded people of billions, NAB tells
court  Construction
15  User_2  Construction        7765
['State Bank to facilitate construction sector financing', 'Habib Bank growing
steadily in UK', 'NBP initiates awareness campaign']  Construction
16  User_2  Construction       12494
World         Other
17  User_2  Construction         377
['KSE above 14,000 points on buying in cement, E&P stocks']  Construction
18  User_2  Construction        7202
['K-Electric initiates construction of $71 million  grid station at Port Qasim']
Construction
19  User_2  Construction        1909
```

```
['Futures trading in KSE-30 index delayed: officials', 'KSE index up by 41
points on possible increase in cement prices']  Construction
20  User_3       Banking        8215
['State Bank rejects allegations of counterfeit currency']       Banking
21  User_3       Banking       12971
World         Other
22  User_3       Banking        7714
['State Bank seen raising discount rate by 50 basis points']       Banking
23  User_3       Banking        7856
['State Bank amends regulations']       Banking
24  User_3       Banking        7831
['Unions thank NBP president for pay package']       Banking
25  User_3       Banking        7884
['State Bank likely to cut discount rate: experts']       Banking
26  User_3       Banking        8617
['HBL reached settlement with New York State Department']       Banking
27  User_3       Banking        7456
['NBP announces 40pc dividend', ''Habib Bank's IPO, GDR approved'']
Banking
28  User_3       Banking        5030
['Cotton spot rate plunges to Rs6,000 a maund', 'Share of textile exports rises
to 52 percent in 2009/10: SBP']       Banking
29  User_3       Banking        8557
['SBP allows due diligence of Bank Alfalah']       Banking
30  User_4        Energy        4248
['PPL donates Rs9.02mln to LDH']        Energy
31  User_4        Energy        3408
['PR gets breather after payment to PSO for fuel', 'Decision on POL prices
today']        Energy
32  User_4        Energy        2837
['PSO buys up to 150,000T of gas oil']        Energy
33  User_4        Energy        4512
['No incentive for textile in trade policy']         Other
34  User_4        Energy        7315
['China starts work on oil refinery in Niger']        Energy
35  User_4        Energy        4334
['Ogra awaits PSO's compliance report to set LNG price']        Energy
36  User_4        Energy        2787
['PSO declares Rs11 a share dividend']        Energy
37  User_4        Energy        2722
['PPL plans domestic, Yemen exploration']        Energy
38  User_4        Energy        3250
['Govt appoints Asim M Khan as PPL acting MD']        Energy
39  User_4        Energy        3159
['PPL Chairman seeks management's']        Energy

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

[11]:
```
# 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