

LinkUp Step 1 Preset - Raw JP to US JP

November 17, 2025

0.1 Intro

This code processes LinkUp job record files and filters out non-US job records

```
[1]: import re
import os
import pandas as pd
import numpy as np
# import textdistance
import difflib
from tqdm import tqdm
from fake_useragent import UserAgent
import tarfile
import time
import datetime
import sqlite3
```

```
[2]: pd.set_option('display.max_colwidth', 100)
pd.set_option('display.max_rows', 2000)
pd.set_option('display.max_columns', 2000)
```

1 Job Records

```
[3]: job_record_dir = "E:/LinkUp/Feeds/Raw Full Job Records/"
```

```
[4]: # Due to the large file sizes, we process the data in two batches:
# the first using names1 and the second using names2.

names1 = ['raw_job_archive_2022-05-31_0.csv',
'raw_job_archive_2022-05-31_1.csv',
'raw_job_archive_2022-05-31_2.csv',
'raw_job_archive_2022-05-31_3.csv',
'raw_job_archive_2022-05-31_4.csv',
'raw_job_archive_2022-05-31_5.csv',
'raw_job_archive_2022-05-31_6.csv',
'raw_job_archive_2022-05-31_7.csv']
names2 = ['raw_job_archive_2022-05-31_8.csv',
'raw_job_archive_2022-05-31_9.csv',
```

```
'raw_job_archive_2022-05-31_a.csv',
'raw_job_archive_2022-05-31_b.csv',
'raw_job_archive_2022-05-31_c.csv',
'raw_job_archive_2022-05-31_d.csv',
'raw_job_archive_2022-05-31_e.csv',
'raw_job_archive_2022-05-31_f.csv']
```

```
[5]: # For the second run, change names1 to names2
total_job_posts = pd.DataFrame()
names_focal = names1
```

```
[6]: # Read in files and filter out non-US job postings
for i in range(0, len(names_focal)):
    csv_path = job_record_dir + names_focal[i]

    start = time.time()
    print(csv_path, " is being processed.")

    # Load job records file into pandas dataframe
    job_records = pd.read_csv(csv_path,
                              parse_dates = [
    ↪['created', 'delete_date', 'last_checked', 'last_updated'],
                              low_memory = False)

    print(" Length Full ", len(job_records), " ")
    job_records_filter = job_records[job_records["country"]=="USA"]

    print(" Length US Jobs ", len(job_records_filter), " ")

    total_job_posts = total_job_posts.append(job_records_filter)
    print("total_job_posts Length:", len(total_job_posts), " ")

    end = time.time()
    print(csv_path, " has been processed, time used:", end - start)
```

E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_0.csv is being processed.

Length Full 13165148

Length US Jobs 10204053

total_job_posts Length: 10204053

E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_0.csv has been processed, time used: 177.34268832206726

E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_1.csv is being processed.

Length Full 13150327

Length US Jobs 10186521

total_job_posts Length: 20390574
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_1.csv has been
 processed, time used: 178.43142867088318
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_2.csv is being
 processed.
 Length Full 13167215
 Length US Jobs 10201561
 total_job_posts Length: 30592135
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_2.csv has been
 processed, time used: 179.1702697277069
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_3.csv is being
 processed.
 Length Full 13184071
 Length US Jobs 10218212
 total_job_posts Length: 40810347
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_3.csv has been
 processed, time used: 180.9726481437683
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_4.csv is being
 processed.
 Length Full 13183464
 Length US Jobs 10217246
 total_job_posts Length: 51027593
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_4.csv has been
 processed, time used: 181.21062922477722
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_5.csv is being
 processed.
 Length Full 13186211
 Length US Jobs 10219563
 total_job_posts Length: 61247156
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_5.csv has been
 processed, time used: 184.5922622680664
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_6.csv is being
 processed.
 Length Full 13174879
 Length US Jobs 10210176
 total_job_posts Length: 71457332
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_6.csv has been
 processed, time used: 180.72780752182007
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_7.csv is being
 processed.
 Length Full 13182592
 Length US Jobs 10214729
 total_job_posts Length: 81672061
 E:/LinkUp/Feeds/Raw Full Job Records/raw_job_archive_2022-05-31_7.csv has been
 processed, time used: 185.7789750099182

```
[7]: # Output for files in name1
total_job_posts.to_pickle("E:/LinkUp/Processed_Data/
↳20220607US_Job_Record_batch1.pkl")
```

```
[ ]: # Output for files in name2
total_job_posts.to_pickle("E:/LinkUp/Processed_Data/
↳20220607US_Job_Record_batch2.pkl")
```

```
[ ]:
```