

DEI Jobs Step 2 - Job Description Sample

November 17, 2025

1 Intro

This code processes job descriptions that belongs to chosen job posts

```
[3]: import re
import os
import pandas as pd
import numpy as np
from tqdm import tqdm
```

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

2 Hashs to keep

```
[5]: # Run the code twice for public firms and private firms
# job_record_chosen = pd.read_pickle("Processed_Data/total_job_posts_chosen.
    ↪.pkl")
job_record_chosen = pd.read_pickle("Processed_Data/private_JP_chosen_final.pkl")
```

```
[6]: job_record_chosen.head()
```

```
[6]:
```

		hash	title \
401454	000b65e99a24466e9968034288559566	Production/Engineering Liaison IV	
401455	000c9e613888cc383bfd457252a95f33	Security Specialist IV	
401457	0012815326cc4154682ac2bb81eb6405	Material Control Coordinator II	
401463	00212e8a87a2f113b7c64686cb166ed6	Project Manager III	
401468	0028c3b3824326164e70447e4ebd61a8	Project Control Analyst III	

	company_id	company_name	city	state	zip	country \
401454	6902	General Atomics	san diego	CA	92140	USA
401455	6902	General Atomics	boulder	CO	80301	USA
401457	6902	General Atomics	kaysville	UT	84037	USA
401463	6902	General Atomics	san diego	CA	92140	USA
401468	6902	General Atomics	san diego	CA	92140	USA

	created	last_checked	last_updated	\
401454	2021-09-22	2021-10-17 01:26:00+00:00		None
401455	2020-07-04	2020-10-27 18:31:00+00:00		None
401457	2020-06-03	2020-06-03 15:46:00+00:00		None
401463	2021-11-16	2021-11-18 12:17:00+00:00		None
401468	2022-01-16	2022-01-27 11:11:00+00:00		None

	delete_date	unmapped_location	onet_occupation_code	\
401454	2021-10-19 02:00:00+00:00		0 17-2112.00	
401455	2020-10-30 07:52:00+00:00		0 15-1212.00	
401457	2020-06-05 15:53:00+00:00		0 43-5061.00	
401463	2021-11-20 13:34:00+00:00		0 11-9199.00	
401468	2022-01-29 13:42:00+00:00		0 13-1111.00	

	url	\
401454	https://sjobs.brassring.com/TGnewUI/Search/home/HomeWithPreLoad?PageType=JobDetails&partnerid=25...	
401455	https://sjobs.brassring.com/TGnewUI/Search/home/HomeWithPreLoad?PageType=JobDetails&partnerid=25...	
401457	https://sjobs.brassring.com/TGnewUI/Search/home/HomeWithPreLoad?PageType=JobDetails&partnerid=25...	
401463	https://sjobs.brassring.com/TGnewUI/Search/home/HomeWithPreLoad?PageType=JobDetails&partnerid=25...	
401468	https://sjobs.brassring.com/TGnewUI/Search/home/HomeWithPreLoad?PageType=JobDetails&partnerid=25...	

	company_id_updated	created_date	\
401454	NaN	2021-09-22 12:32:00+00:00	
401455	NaN	2020-07-04 00:16:00+00:00	
401457	NaN	2020-06-03 15:46:00+00:00	
401463	NaN	2021-11-16 10:43:00+00:00	
401468	NaN	2022-01-16 16:46:00+00:00	

	onet_occupation_code_NEW	Job	Zone	year_month	factset_entity_id	cusip	\
401454	17-2112.00		4.0	2021-09	05M40Q-E	None	
401455	15-1212.00		4.0	2020-07	05M40Q-E	None	
401457	43-5061.00		3.0	2020-06	05M40Q-E	None	
401463	11-9199.00		NaN	2021-11	05M40Q-E	None	
401468	13-1111.00		5.0	2022-01	05M40Q-E	None	

	cusip_6	gvkey	the_key	Identifier	Name	\
401454	None	None	fskey_05M40Q-E	05M40Q-E	General Atomics (California)	
401455	None	None	fskey_05M40Q-E	05M40Q-E	General Atomics (California)	
401457	None	None	fskey_05M40Q-E	05M40Q-E	General Atomics (California)	
401463	None	None	fskey_05M40Q-E	05M40Q-E	General Atomics (California)	
401468	None	None	fskey_05M40Q-E	05M40Q-E	General Atomics (California)	

	Revenue
401454	1200.254359
401455	1200.254359
401457	1200.254359
401463	1200.254359
401468	1200.254359

```
[7]: hash_chosen = list(job_record_chosen["hash"])
```

3 Job Descriptions

```
[8]: import re
```

```
[9]: job_record_dir_newones = "Feeds/Raw Full Job Descriptions/"
```

```
[10]: import glob
print(glob.glob(job_record_dir_newones+"*"))
```

```
['Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_0.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_1.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_2.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_3.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_4.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_5.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_6.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_7.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_8.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_9.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_a.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_b.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_c.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_d.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_e.xml',
'Feeds/Raw Full Job Descriptions\\linkup_job_descriptions_2022-04-30_f.xml']
```

```
[11]: names_descriptions = os.listdir(job_record_dir_newones)
names_descriptions.sort(reverse=True)
names_descriptions
```

```
[11]: ['linkup_job_descriptions_2022-04-30_f.xml',
'linkup_job_descriptions_2022-04-30_e.xml',
'linkup_job_descriptions_2022-04-30_d.xml',
'linkup_job_descriptions_2022-04-30_c.xml',
'linkup_job_descriptions_2022-04-30_b.xml',
'linkup_job_descriptions_2022-04-30_a.xml',
'linkup_job_descriptions_2022-04-30_9.xml',
'linkup_job_descriptions_2022-04-30_8.xml',
```

```
'linkup_job_descriptions_2022-04-30_7.xml',
'linkup_job_descriptions_2022-04-30_6.xml',
'linkup_job_descriptions_2022-04-30_5.xml',
'linkup_job_descriptions_2022-04-30_4.xml',
'linkup_job_descriptions_2022-04-30_3.xml',
'linkup_job_descriptions_2022-04-30_2.xml',
'linkup_job_descriptions_2022-04-30_1.xml',
'linkup_job_descriptions_2022-04-30_0.xml']
```

```
[12]: total_job_descriptions = pd.DataFrame()
```

```
[11]: len(names_descriptions)
```

```
[11]: 16
```

```
[13]: # for i in range(0, 6):
# for i in range(6, 11):
# for i in range(11, 16):
for i in range(0, 16):
    print(names_descriptions[i])
```

```
linkup_job_descriptions_2022-04-30_f.xml
linkup_job_descriptions_2022-04-30_e.xml
linkup_job_descriptions_2022-04-30_d.xml
linkup_job_descriptions_2022-04-30_c.xml
linkup_job_descriptions_2022-04-30_b.xml
linkup_job_descriptions_2022-04-30_a.xml
linkup_job_descriptions_2022-04-30_9.xml
linkup_job_descriptions_2022-04-30_8.xml
linkup_job_descriptions_2022-04-30_7.xml
linkup_job_descriptions_2022-04-30_6.xml
linkup_job_descriptions_2022-04-30_5.xml
linkup_job_descriptions_2022-04-30_4.xml
linkup_job_descriptions_2022-04-30_3.xml
linkup_job_descriptions_2022-04-30_2.xml
linkup_job_descriptions_2022-04-30_1.xml
linkup_job_descriptions_2022-04-30_0.xml
```

```
[ ]: total_job_descriptions = pd.DataFrame()
# for i in range(0, 6):
# for i in range(6, 11):
# for i in range(11, 16):
for i in range(0, 16):
    xml_path = job_record_dir_newones + "/" + names_descriptions[i]

    # Load job records file into pandas dataframe
    job_records = pdx.read_xml(xml_path, encoding="utf8")
```

```

job_records2 = pd.DataFrame(job_records['jobs'][0])
job_records2["hash"] = job_records2["job"].apply(lambda x:x["hash"])
job_records2["description"] = job_records2["job"].apply(lambda x:
↪x["description"])
job_records2 = job_records2[["hash", "description"]]

print(xml_path, " is being processed, number of jobs", len(job_records2))

job_records2 = job_records2[job_records2["hash"].isin(hash_chosen)]

#     job_records = job_records[["hash", "description"]]

total_job_descriptions = total_job_descriptions.append(job_records2)

print("total_job_descriptions Length:", len(total_job_descriptions))
print(xml_path, " has been processed, number of chosen jobs:",
↪len(job_records2))

```

```

[ ]: total_job_descriptions.to_pickle("Processed_Data/job_descriptions_chosen_4-0.
↪pkl")

```

```

[ ]: total_job_descriptions.to_pickle("Processed_Data/job_descriptions_chosen_f-a.
↪pkl")

```

```

[ ]: total_job_descriptions.to_pickle("Processed_Data/job_descriptions_chosen_9-5.
↪pkl")

```

3.1 Process batch by batch

```

[14]: total_job_descriptions = pd.read_pickle("Processed_Data/
↪job_descriptions_chosen_4-0.pkl")

```

```

[15]: total_job_descriptions2 = pd.read_pickle("Processed_Data/
↪job_descriptions_chosen_9-5.pkl")

```

```

[16]: total_job_descriptions3 = pd.read_pickle("Processed_Data/
↪job_descriptions_chosen_f-a.pkl")

```

```

[17]: total_job_descriptions = total_job_descriptions.append(total_job_descriptions2)
total_job_descriptions = total_job_descriptions.append(total_job_descriptions3)

```

```

[19]: print(len(total_job_descriptions))
print(len(total_job_descriptions2))
print(len(total_job_descriptions3))

```

33613597

10510587

12603380

```
[20]: total_job_descriptions.to_pickle("Processed_Data/job_descriptions_chosen_full.  
    ↪pk1")
```

```
[ ]:
```