BERT

Importing the data

In [1]:

```python
# from google.colab import drive
# drive.mount('/content/drive', force_remount=True)
```

In [2]:

```python
import pandas as pd
import seaborn as sns
```

In [3]:

```python
path = 'Before_BERT(Preprocessed).csv'
df = pd.read_csv(path, index_col = 0, keep_default_na=False)
```

In [4]:

```python
df.head()
```

Out[4]:

	Job Title	Company Name	Industry	Company Location	New Job (90 Days)	Year Started	Profile Headline	Profile Su
0	Battery Designer	Rivian	Mechanical or Industrial Engineering	Dublin, Ohio, United States	False	2020.0	Mechanical Design Engineer, System Integration...	In the ever- technologic w
1	Digital DevOps Engineer	HSBC	Information Technology and Services	New York City Metropolitan Area	False	2018.0	Digital DevOps Engineer at HSBC	AWS Certifie Engineer h AW
2	Product Designer	Two Point Conversions, Inc.	Information Technology and Services	Chicago, Illinois, United States	False	2018.0	Leading Product + UX at Remedy (Two Point Conv...	http://aroonmatl
3	Product Designer	udaan.com	Information Technology and Services	Bangalore Urban, Karnataka, India	False	2018.0	Product Designer at udaan	
4	Digital Technology Intern	GE	Information Technology and Services	Jaipur, Rajasthan, India	True	2021.0	Digital Technology Intern at General Electric ...	

In [5]:

```
df.info()
```

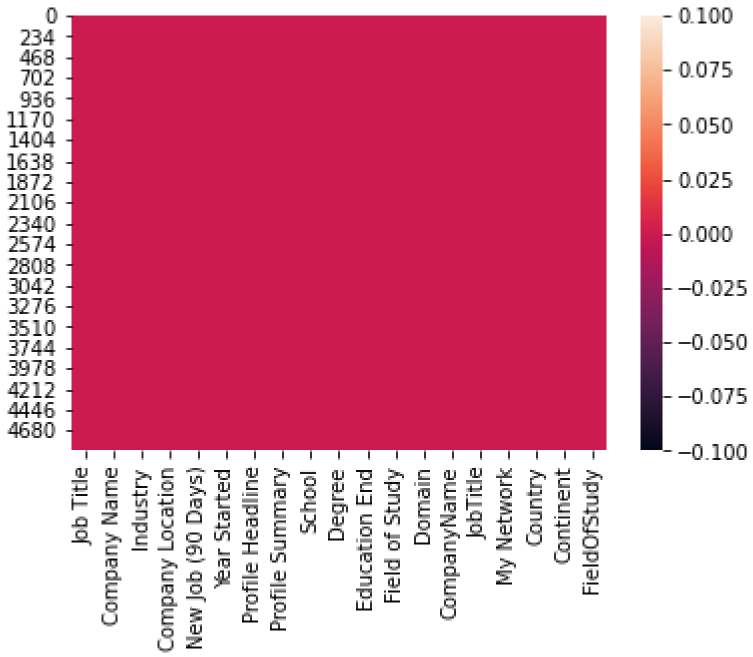
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4910 entries, 0 to 4909
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Job Title           4910 non-null   object 
 1   Company Name        4910 non-null   object 
 2   Industry            4910 non-null   object 
 3   Company Location    4910 non-null   object 
 4   New Job (90 Days)   4910 non-null   bool   
 5   Year Started        4910 non-null   float64
 6   Profile Headline    4910 non-null   object 
 7   Profile Summary     4910 non-null   object 
 8   School              4910 non-null   object 
 9   Degree              4910 non-null   object 
 10  Education End       4910 non-null   object 
 11  Field of Study      4910 non-null   object 
 12  Domain              4910 non-null   object 
 13  CompanyName         4910 non-null   object 
 14  JobTitle            4910 non-null   object 
 15  My Network          4910 non-null   object 
 16  Country             4910 non-null   object 
 17  Continent           4910 non-null   object 
 18  FieldOfStudy        4910 non-null   object 
dtypes: bool(1), float64(1), object(17)
memory usage: 733.6+ KB
```

In [6]:

```
sns.heatmap(df.isnull())
```

Out[6]:

```
<AxesSubplot:>
```



BERT

In [7]:

```
! pip install sentence-transformers
```

```
Requirement already satisfied: sentence-transformers in c:\users\hp\anaconda
3\lib\site-packages (1.1.1)
Requirement already satisfied: transformers<5.0.0,>=3.1.0 in c:\users\hp\ana
conda3\lib\site-packages (from sentence-transformers) (3.3.1)
Requirement already satisfied: scipy in c:\users\hp\anaconda3\lib\site-packa
ges (from sentence-transformers) (1.6.1)
Requirement already satisfied: tqdm in c:\users\hp\anaconda3\lib\site-packag
es (from sentence-transformers) (4.58.0)
Requirement already satisfied: scikit-learn in c:\users\hp\anaconda3\lib\sit
e-packages (from sentence-transformers) (0.23.2)
Requirement already satisfied: sentencepiece in c:\users\hp\anaconda3\lib\si
te-packages (from sentence-transformers) (0.1.91)
Requirement already satisfied: nltk in c:\users\hp\anaconda3\lib\site-packag
es (from sentence-transformers) (3.5)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from sentence-transformers) (1.20.3)
Requirement already satisfied: torch>=1.6.0 in c:\users\hp\anaconda3\lib\sit
e-packages (from sentence-transformers) (1.6.0)
Requirement already satisfied: torchvision in c:\users\hp\anaconda3\lib\site
-packages (from sentence-transformers) (0.7.0)
Requirement already satisfied: regex in c:\users\hp\anaconda3\lib\site-packa
ges (from nltk->sentence-transformers) (2020.11.13)
Requirement already satisfied: click in c:\users\hp\anaconda3\lib\site-packa
ges (from nltk->sentence-transformers) (7.1.2)
Requirement already satisfied: joblib in c:\users\hp\anaconda3\lib\site-pack
ages (from nltk->sentence-transformers) (1.0.1)
Requirement already satisfied: tqdm in c:\users\hp\anaconda3\lib\site-packag
es (from sentence-transformers) (4.58.0)
Requirement already satisfied: joblib in c:\users\hp\anaconda3\lib\site-pack
ages (from nltk->sentence-transformers) (1.0.1)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from sentence-transformers) (1.20.3)
Requirement already satisfied: scipy in c:\users\hp\anaconda3\lib\site-packa
ges (from sentence-transformers) (1.6.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\hp\anaconda3
\lib\site-packages (from scikit-learn->sentence-transformers) (2.1.0)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from sentence-transformers) (1.20.3)
Requirement already satisfied: future in c:\users\hp\anaconda3\lib\site-pack
ages (from torch>=1.6.0->sentence-transformers) (0.18.2)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from sentence-transformers) (1.20.3)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from sentence-transformers) (1.20.3)
Requirement already satisfied: torch>=1.6.0 in c:\users\hp\anaconda3\lib\sit
e-packages (from sentence-transformers) (1.6.0)
Requirement already satisfied: pillow>=4.1.1 in c:\users\hp\anaconda3\lib\si
te-packages (from torchvision->sentence-transformers) (8.0.1)
Requirement already satisfied: tokenizers==0.8.1.rc2 in c:\users\hp\anaconda
3\lib\site-packages (from transformers<5.0.0,>=3.1.0->sentence-transformers)
(0.8.1rc2)
Requirement already satisfied: sentencepiece in c:\users\hp\anaconda3\lib\si
te-packages (from sentence-transformers) (0.1.91)
Requirement already satisfied: regex in c:\users\hp\anaconda3\lib\site-packa
ges (from nltk->sentence-transformers) (2020.11.13)
Requirement already satisfied: filelock in c:\users\hp\anaconda3\lib\site-pa
ckages (from transformers<5.0.0,>=3.1.0->sentence-transformers) (3.0.12)
```

```
Requirement already satisfied: packaging in c:\users\hp\anaconda3\lib\site-p
ackages (from transformers<5.0.0,>=3.1.0->sentence-transformers) (20.7)
Requirement already satisfied: tqdm in c:\users\hp\anaconda3\lib\site-packag
es (from sentence-transformers) (4.58.0)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from sentence-transformers) (1.20.3)
Requirement already satisfied: requests in c:\users\hp\anaconda3\lib\site-pa
ckages (from transformers<5.0.0,>=3.1.0->sentence-transformers) (2.25.0)
Requirement already satisfied: sacremoses in c:\users\hp\anaconda3\lib\site-
packages (from transformers<5.0.0,>=3.1.0->sentence-transformers) (0.0.43)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\hp\anaconda3\lib
\site-packages (from packaging->transformers<5.0.0,>=3.1.0->sentence-transfo
rmers) (2.4.7)
Requirement already satisfied: idna<3,>=2.5 in c:\users\hp\anaconda3\lib\sit
e-packages (from requests->transformers<5.0.0,>=3.1.0->sentence-transformer
s) (2.10)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\hp\anaconda
3\lib\site-packages (from requests->transformers<5.0.0,>=3.1.0->sentence-tra
nsformers) (1.25.11)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\hp\anaconda3\li
b\site-packages (from requests->transformers<5.0.0,>=3.1.0->sentence-transfo
rmers) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\hp\anaconda3\l
ib\site-packages (from requests->transformers<5.0.0,>=3.1.0->sentence-transf
ormers) (2020.12.5)
Requirement already satisfied: joblib in c:\users\hp\anaconda3\lib\site-pack
ages (from nltk->sentence-transformers) (1.0.1)
Requirement already satisfied: regex in c:\users\hp\anaconda3\lib\site-packa
ges (from nltk->sentence-transformers) (2020.11.13)
Requirement already satisfied: click in c:\users\hp\anaconda3\lib\site-packa
ges (from nltk->sentence-transformers) (7.1.2)
Requirement already satisfied: tqdm in c:\users\hp\anaconda3\lib\site-packag
es (from sentence-transformers) (4.58.0)
Requirement already satisfied: six in c:\users\hp\anaconda3\lib\site-package
s (from sacremoses->transformers<5.0.0,>=3.1.0->sentence-transformers) (1.1
5.0)
```

In [8]:

```
!pip install nlu
!pip install pyspark
```

```
Requirement already satisfied: nlu in c:\users\hp\anaconda3\lib\site-package
s (3.0.1)
Requirement already satisfied: pandas in c:\users\hp\anaconda3\lib\site-pack
ages (from nlu) (1.1.5)
Requirement already satisfied: dataclasses in c:\users\hp\anaconda3\lib\site
-packages (from nlu) (0.6)
Requirement already satisfied: pyarrow>=0.16.0 in c:\users\hp\anaconda3\lib
\site-packages (from nlu) (4.0.0)
Requirement already satisfied: spark-nlp<3.1.0,>=3.0.0 in c:\users\hp\anacon
da3\lib\site-packages (from nlu) (3.0.3)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from nlu) (1.20.3)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from nlu) (1.20.3)
Requirement already satisfied: pytz>=2017.2 in c:\users\hp\anaconda3\lib\sit
e-packages (from pandas->nlu) (2020.4)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\hp\anacond
a3\lib\site-packages (from pandas->nlu) (2.8.1)
Requirement already satisfied: numpy in c:\users\hp\anaconda3\lib\site-packa
ges (from nlu) (1.20.3)
Requirement already satisfied: six>=1.5 in c:\users\hp\anaconda3\lib\site-pa
ckages (from python-dateutil>=2.7.3->pandas->nlu) (1.15.0)
Requirement already satisfied: pyspark in c:\users\hp\anaconda3\lib\site-pac
kages (3.1.1)
Requirement already satisfied: py4j==0.10.9 in c:\users\hp\anaconda3\lib\sit
e-packages (from pyspark) (0.10.9)
```

In [9]:

```
from sentence_transformers import SentenceTransformer
sbert_model = SentenceTransformer('bert-base-nli-mean-tokens')
```

In [10]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4910 entries, 0 to 4909
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Job Title           4910 non-null   object 
 1   Company Name        4910 non-null   object 
 2   Industry            4910 non-null   object 
 3   Company Location    4910 non-null   object 
 4   New Job (90 Days)   4910 non-null   bool   
 5   Year Started        4910 non-null   float64
 6   Profile Headline    4910 non-null   object 
 7   Profile Summary     4910 non-null   object 
 8   School              4910 non-null   object 
 9   Degree              4910 non-null   object 
 10  Education End       4910 non-null   object 
 11  Field of Study      4910 non-null   object 
 12  Domain              4910 non-null   object 
 13  CompanyName         4910 non-null   object 
 14  JobTitle            4910 non-null   object 
 15  My Network          4910 non-null   object 
 16  Country             4910 non-null   object 
 17  Continent           4910 non-null   object 
 18  FieldOfStudy        4910 non-null   object 
dtypes: bool(1), float64(1), object(17)
memory usage: 733.6+ KB
```

In [11]:

```
df1 = df.drop(['Job Title', 'Company Name', 'Field of Study'], axis = 1)
```

In [12]:

```
for column in df1.columns:
    if (column == 'New Job (90 Days)') or (column == 'Year Started'):
        continue
    new_col = column + "_embedding"
    df1[new_col] = 0
    df1[new_col] = sbert_model.encode(df1[column])
```

In [13]:

```
path = 'After_BERT(Embedded_Data).csv'
df1.to_csv(path)
```