The following data is build on this dataset from Kaggle:
https://www.kaggle.com/datasets/joanirudh/resumecorpus-cleaned/data

It has to be downloaded, unzipped and inserted into the /Data folder. It is too large to upload to github.

In [1]:
```
import pandas as pd
import numpy as np
```

In [ ]:
```
# Loading of the data
def loadData(file_path):
    """
    Load data from a CSV file.
    """
    try:
        data = pd.read_csv(file_path)
        return data
    except FileNotFoundError:
        print(f"File {file_path} not found.")
        return None


df = loadData("Data/finale.csv")
```

In [ ]:
```
# We print the head and the columns to check what kind of data we have
print(df.head())
print(df.columns)
```

```
   Unnamed: 0                                               Text  \
0           0  Database AdministratorDatabase AdministratorDa...
1           1  Database AdministratorSQL Microsoft PowerPoint...
2           2  Oracle Database AdministratorOracle Database A...
3           3  Amazon Redshift Administrator ETL Developer Bu...
4           4  Scrum MasterOracle Database Administrator Scru...

                        Label  Software_Developer  Database_Administrator  \
0  [b'Database_Administrator']                   0                       1
1  [b'Database_Administrator']                   0                       1
2  [b'Database_Administrator']                   0                       1
3  [b'Database_Administrator']                   0                       1
4  [b'Database_Administrator']                   0                       1

   Systems_Administrator  Project_manager  Web_Developer  \
0                      0                0              0
1                      0                0              0
2                      0                0              0
3                      0                0              0
4                      0                0              0

   Network_Administrator  Security_Analyst  Python_Developer  Java_Developer  \
0                      0                 0                 0               0
1                      0                 0                 0               0
2                      0                 0                 0               0
3                      0                 0                 0               0
4                      0                 0                 0               0

   Front_End_Developer  TextLen  \
0                    0     6511
1                    0     2027
2                    0     3115
3                    0     3328
4                    0     4105

                                              Ents  \
0  [SQL, SSIS, OLTP, Backing, Generating, Log Shi...
1  [Microsoft, XP Microsoft, Assembly Language Mi...
2  [Carrier Objective, Experienced Creating Users...
3  [Amazon Redshift Administrator, yearsPostgresS...
4  [Scrum Master, Scrum Master, October  April  R...

                                               new
0  sql ssis oltp backing generating log shipping ...
1  microsoft xp microsoft assembly language micro...
2  carrier objective experienced creating users r...
3  amazon redshift administrator redshift adminis...
4  scrum masterscrum master scrum master scrum ma...
Index(['Unnamed: 0', 'Text', 'Label', 'Software_Developer',
       'Database_Administrator', 'Systems_Administrator', 'Project_manager',
       'Web_Developer', 'Network_Administrator', 'Security_Analyst',
       'Python_Developer', 'Java_Developer', 'Front_End_Developer', 'TextLen',
       'Ents', 'new'],
      dtype='object')
```

Data inspection

From the first look, we can tell that This dataset contains onehot encoding for each of the job categories. This is useful for training our models later on.

Then we have the resume column, which holds the resume text.

In [ ]:
```
print(df['Text'].head())

print(df['Label'].nunique())
```

```
0    Database AdministratorDatabase AdministratorDa...
1    Database AdministratorSQL Microsoft PowerPoint...
2    Oracle Database AdministratorOracle Database A...
3    Amazon Redshift Administrator ETL Developer Bu...
4    Scrum MasterOracle Database Administrator Scru...
Name: Text, dtype: object
552
```

The text column holds the resumes and the Label column holds the different Categories of the jobs. We can drop the label later and just use the onehot encoded columns for classifying the resumes.

In [5]:
```
df.dtypes
```

Out[5]:
```
Unnamed: 0                 int64
Text                      object
Label                     object
Software_Developer         int64
Database_Administrator     int64
Systems_Administrator      int64
Project_manager            int64
Web_Developer              int64
Network_Administrator      int64
Security_Analyst           int64
Python_Developer           int64
Java_Developer             int64
Front_End_Developer        int64
TextLen                    int64
Ents                      object
new                       object
dtype: object
```

All the different categories are int64 (One hot encoded, which means they are either 0 or 1)

Lets check the textlen, ents and new column.

In [6]:
```
print(df['TextLen'])
```

```
0         6511
1         2027
2         3115
3         3328
4         4105
         ...
29778    18467
29779     7961
29780    14170
29781      693
29782     4384
Name: TextLen, Length: 29783, dtype: int64
```

In [7]:
```
# So we think that the textlen is the length of the resume text, but lets check it
# Lets check the length of the first resume text
print(len(df['Text'][0]))
print(df['TextLen'][0])
```

```
6511
6511
```

So the textlen is just the length of the resume.

In [8]:
```
print(df['Ents'])
print(df['new'])
```

```
0        [SQL, SSIS, OLTP, Backing, Generating, Log Shi...
1        [Microsoft, XP Microsoft, Assembly Language Mi...
2        [Carrier Objective, Experienced Creating Users...
3        [Amazon Redshift Administrator, yearsPostgresS...
4        [Scrum Master, Scrum Master, October  April  R...
                               ...
29778    [ServiceNow DeveloperServiceNow DeveloperServi...
29779    [DeveloperAndroid, DeveloperJoomla CMS, Samsun...
29780    [UI, WEB, DeveloperVisual, Web Applications De...
29781                             [September  July  , PHP]
29782    [Sr Software, Web DeveloperWeb DeveloperWeb De...
Name: Ents, Length: 29783, dtype: object
0        sql ssis oltp backing generating log shipping ...
1        microsoft xp microsoft assembly language micro...
2        carrier objective experienced creating users r...
3        amazon redshift administrator redshift adminis...
4        scrum masterscrum master scrum master scrum ma...
                               ...
29778    servicenow developerservicenow developerservic...
29779    developerandroid developerjoomla cms samsung w...
29780    ui web developervisual web applications develo...
29781                                                  php
29782    sr software web developerweb developerweb deve...
Name: new, Length: 29783, dtype: object
```

So it looks like that the 'Ents' column are the categories, so we wont need them, as they are already onehot encoded into different columns. The 'new' column just contains the categories in lowercase.

In [9]:
```
# Lets drop the 'new' column, as it is not needed
newDF = df.drop(columns=['new'])

# We drop the 'Ents' column, as it is not needed
newDF = newDF.drop(columns=['Ents'])
# And we also rename the 'Text' column to 'Resume' instead:
newDF = newDF.rename(columns={'Text': 'Resume'})
# We rename the 'Unnamed: 0' column to 'ID' instead:
newDF = newDF.rename(columns={'Unnamed: 0': 'ID'})

print(newDF.columns )
```

```
Index(['ID', 'Resume', 'Label', 'Software_Developer', 'Database_Administrator',
       'Systems_Administrator', 'Project_manager', 'Web_Developer',
       'Network_Administrator', 'Security_Analyst', 'Python_Developer',
       'Java_Developer', 'Front_End_Developer', 'TextLen'],
      dtype='object')
```

Now we can check for null and duplicate values

In [10]:
```
print(newDF.isnull().sum())
# We have no null values in the dataframe, so we can proceed with the next steps

duplicates = newDF[newDF.duplicated(subset=['Resume'], keep=False)]
print(duplicates)
```

```
ID                        0
Resume                    0
Label                     0
Software_Developer        0
Database_Administrator    0
Systems_Administrator     0
Project_manager           0
Web_Developer             0
Network_Administrator     0
Security_Analyst          0
Python_Developer          0
Java_Developer            0
Front_End_Developer       0
TextLen                   0
dtype: int64
          ID                                             Resume  \
55        55  Web DeveloperDatabase AdministratorComputer Sp...
65        65  Database AdministratorDatabase AdministratorNe...
81        81  Database AdministratorDirector Support Systems...
93        93  Database AdministratorContent Operations Proje...
96        96  Network Database AdministratorInterim IT Manag...
...      ...                                                ...
29750  29750  Front End Web DeveloperWeb DeveloperSenior Fro...
29757  29757  Web  developerSOA DeveloperComputer Technician...
29759  29759  Software EngineerWeb Application DeveloperCSSH...
29761  29761  Freelance Web DeveloperJava ProgrammerMany Asp...
29767  29767  Software DeveloperFrontEnd DeveloperWeb Develo...

                                                   Label  Software_Developer  \
55     [b'Web_Developer\n', b'Software_Developer\n', ...                    1
65                              [b'Database_Administrator']                 0
81                              [b'Database_Administrator']                 0
93     [b'Database_Administrator\n', b'Project_manage...                    0
96     [b'Database_Administrator\n', b'Network_Admini...                    0
...                                                  ...                  ...
29750  [b'Web_Developer\n', b'Software_Developer\n', ...                    1
29757        [b'Web_Developer\n', b'Software_Developer']                    1
29759        [b'Software_Developer\n', b'Web_Developer']                    1
29761        [b'Web_Developer\n', b'Software_Developer']                    1
29767  [b'Software_Developer\n', b'Front_End_Develope...                    1

       Database_Administrator  Systems_Administrator  Project_manager  \
55                          1                      0                0
65                          1                      0                0
81                          1                      0                0
93                          1                      0                1
96                          1                      0                1
...                       ...                    ...              ...
29750                       0                      0                0
29757                       0                      0                0
29759                       0                      0                0
29761                       0                      0                0
29767                       0                      0                0

       Web_Developer  Network_Administrator  Security_Analyst  \
55                 1                      0                 0
65                 0                      0                 0
81                 0                      0                 0
93                 0                      1                 0
96                 0                      1                 0
...              ...                    ...               ...
29750              1                      0                 0
29757              1                      0                 0
29759              1                      0                 0
```

```
29761              1                     0                0
29767              1                     0                0

       Python_Developer  Java_Developer  Front_End_Developer  TextLen
55                    0               0                    0     2651
65                    0               0                    0     5690
81                    0               0                    0     5287
93                    0               0                    0     2599
96                    0               0                    0     1401
...                 ...             ...                  ...      ...
29750                 0               0                    1     3006
29757                 0               0                    0     8885
29759                 0               0                    0     1555
29761                 0               0                    0     1369
29767                 0               0                    1     2280

[1668 rows x 14 columns]
```

In [11]:
```
# We have 1668 duplicated resumes, so we will drop them
newDF = newDF.drop_duplicates(subset=['Resume'], keep='first')
print(newDF.isnull().sum())

print(newDF.count())
```

```
ID                         0
Resume                     0
Label                      0
Software_Developer         0
Database_Administrator     0
Systems_Administrator      0
Project_manager            0
Web_Developer              0
Network_Administrator      0
Security_Analyst           0
Python_Developer           0
Java_Developer             0
Front_End_Developer        0
TextLen                    0
dtype: int64
ID                         28947
Resume                     28947
Label                      28947
Software_Developer         28947
Database_Administrator     28947
Systems_Administrator      28947
Project_manager            28947
Web_Developer              28947
Network_Administrator      28947
Security_Analyst           28947
Python_Developer           28947
Java_Developer             28947
Front_End_Developer        28947
TextLen                    28947
dtype: int64
```

We now have our finished dataframe, which has been cleaned. We will have to save it as a pickle file. The pickle files are gonna be too large to upload to github, so we save it in a gitignored folder in /Data/DataFrames

In [ ]:
```
import pickle

with open('Data/Dataframes/newDF.pkl', 'wb') as f:
    pickle.dump(newDF, f)
```

```
with open('Data/Dataframes/newDF.pkl', 'rb') as f:
    newDF = pickle.load(f)
print(newDF.head())
print(newDF.columns)
```

```
   ID                                             Resume  \
0   0  Database AdministratorDatabase AdministratorDa...
1   1  Database AdministratorSQL Microsoft PowerPoint...
2   2  Oracle Database AdministratorOracle Database A...
3   3  Amazon Redshift Administrator ETL Developer Bu...
4   4  Scrum MasterOracle Database Administrator Scru...

                        Label  Software_Developer  Database_Administrator  \
0  [b'Database_Administrator']                   0                       1
1  [b'Database_Administrator']                   0                       1
2  [b'Database_Administrator']                   0                       1
3  [b'Database_Administrator']                   0                       1
4  [b'Database_Administrator']                   0                       1

   Systems_Administrator  Project_manager  Web_Developer  \
0                      0                0              0
1                      0                0              0
2                      0                0              0
3                      0                0              0
4                      0                0              0

   Network_Administrator  Security_Analyst  Python_Developer  Java_Developer  \
0                      0                 0                 0               0
1                      0                 0                 0               0
2                      0                 0                 0               0
3                      0                 0                 0               0
4                      0                 0                 0               0

   Front_End_Developer  TextLen
0                    0     6511
1                    0     2027
2                    0     3115
3                    0     3328
4                    0     4105
```