

Assignment 4

@author : @ruhend

DATA PRE-PROCESSING

Need of pre-processing

Real world data are usually

- Inconsistent: containing discrepancies in codes or names.
- Incomplete: lacking attribute values, lacking certain attributes of --interest, or containing only aggregate data.
- Noisy: containing errors or outliers.

What is data pre-processing

- Preprocessing refers to the transformations applied to the data before feeding it to the machine learning algorithms.
- The data gathered from different sources is collected in raw format which is not feasible for the analysis.
- Data Preprocessing technique is used to convert the raw data into a clean data set.

```
In [21]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
In [22]: dataset = pd.read_csv("company_data.csv")
print ('dataset: %s'%(str(dataset.shape)))
```

```
dataset: (11654, 19)
```

In [23]: dataset

Out[23]:

	Unnamed: 0	name	headline	location	followers	connections	about	time_spe
0	0	Dell Technologies	Information Technology & Services	Round Rock, Texas	3,597,845 followers	194,140	Dell Technologies is a unique family of busine...	1 week a
1	1	Dell Technologies	Information Technology & Services	Round Rock, Texas	3,597,845 followers	194,140	Dell Technologies is a unique family of busine...	20 hou a
2	2	Dell Technologies	Information Technology & Services	Round Rock, Texas	3,597,845 followers	194,140	Dell Technologies is a unique family of busine...	20 hou a
3	3	Dell Technologies	Information Technology & Services	Round Rock, Texas	3,597,845 followers	194,140	Dell Technologies is a unique family of busine...	22 hou a
4	4	Dell Technologies	Information Technology & Services	Round Rock, Texas	3,597,845 followers	194,140	Dell Technologies is a unique family of busine...	2 days a
...
11649	11649	John Deere	Machinery	Moline, IL	849,594 followers	39,682	John Deere is a world leader in providing adva...	8 montl a
11650	11650	John Deere	Machinery	Moline, IL	849,594 followers	39,682	John Deere is a world leader in providing adva...	1 year a
11651	11651	John Deere	Machinery	Moline, IL	849,594 followers	39,682	John Deere is a world leader in providing adva...	1 year a
11652	11652	John Deere	Machinery	Moline, IL	849,594 followers	39,682	John Deere is a world leader in providing adva...	1 year a
11653	11653	John Deere	Machinery	Moline, IL	849,594 followers	39,682	John Deere is a world leader in providing adva...	1 year a

11654 rows × 19 columns

```
In [24]: dataset.isnull().sum()
```

```
Out[24]: Unnamed: 0          0
         name              0
         headline          30
         location          30
         followers         30
         connections        0
         about              0
         time_spent         0
         content           283
         content_links       0
         media_type         640
         media_urls          0
         num_hashtags        0
         hashtag_followers   0
         hashtags            0
         reactions           0
         comments            0
         views              11654
         votes               11607
         dtype: int64
```

```
In [25]: dataset.isnull().sum().sum()
```

```
Out[25]: 24274
```

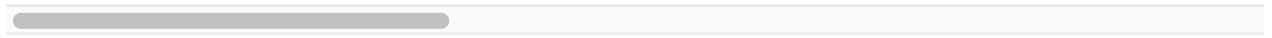
Dropping the Unwanted data columns

```
In [26]: dataset=dataset.drop(['votes', 'views', 'media_type', 'content', 'followers'],
dataset
```

Out[26]:

	Unnamed: 0	name	headline	location	connections	about	time_spent	
0	0	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	1 week ago	[[http
1	1	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[[https:
2	2	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[[http
3	3	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	22 hours ago	
4	4	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	2 days ago	[[http
...	
11649	11649	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	8 months ago	
11650	11650	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[[https:/
11651	11651	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11652	11652	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11653	11653	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[[https

11654 rows × 14 columns



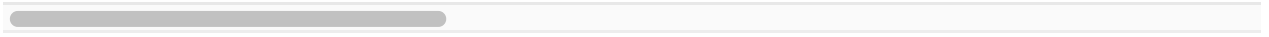
Replacing missing values using Python functions

```
In [27]: # forward fill
dataset.ffmpeg(inplace = True)
dataset
```

Out[27]:

	Unnamed: 0	name	headline	location	connections	about	time_spent	
0	0	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	1 week ago	[['http
1	1	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[['https:
2	2	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[['http
3	3	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	22 hours ago	
4	4	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	2 days ago	[['http
...	
11649	11649	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	8 months ago	
11650	11650	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[['https:/
11651	11651	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11652	11652	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11653	11653	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[['https

11654 rows × 14 columns



```
In [28]: dataset.isnull().sum().sum()
```

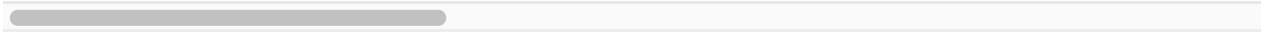
```
Out[28]: 0
```

```
In [29]: # Backward fill
dataset.bfill(inplace = True)
dataset
```

Out[29]:

	Unnamed: 0	name	headline	location	connections	about	time_spent	
0	0	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	1 week ago	[['http
1	1	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[['https:
2	2	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[['http
3	3	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	22 hours ago	
4	4	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	2 days ago	[['http
...	
11649	11649	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	8 months ago	
11650	11650	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[['https:/
11651	11651	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11652	11652	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11653	11653	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[['https

11654 rows x 14 columns



```
In [30]: dataset.isnull().sum().sum()
```

```
Out[30]: 0
```

Replacing Missing Values with Mean Values of the columns

```
In [31]: dataset.fillna(np.mean(dataset["reactions"]), inplace = True) #Using python
dataset
```

Out[31]:

	Unnamed: 0	name	headline	location	connections	about	time_spent	
0	0	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	1 week ago	[[http
1	1	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[[https:
2	2	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[[http
3	3	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	22 hours ago	
4	4	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	2 days ago	[[http
...	
11649	11649	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	8 months ago	
11650	11650	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[[https:/
11651	11651	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11652	11652	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11653	11653	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[[https

11654 rows × 14 columns

```
In [32]: dataset.isnull().sum().sum()
```

```
Out[32]: 0
```

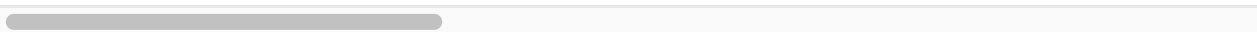
Replacing missing values with a global constant

```
In [33]: dataset.fillna(0.00000,inplace = True)
dataset
```

Out[33]:

	Unnamed: 0	name	headline	location	connections	about	time_spent	
0	0	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	1 week ago	[['http
1	1	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[['https:
2	2	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[['http
3	3	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	22 hours ago	
4	4	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	2 days ago	[['http
...	
11649	11649	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	8 months ago	
11650	11650	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[['https:/
11651	11651	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11652	11652	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11653	11653	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[['https

11654 rows x 14 columns



```
In [34]: dataset.isnull().sum().sum()
```

```
Out[34]: 0
```

Eliminating the missing value rows

```
In [35]: dataset.dropna(axis=0,inplace=True) #using dropna function
dataset
```

Out[35]:

	Unnamed: 0	name	headline	location	connections	about	time_spent	
0	0	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	1 week ago	[['http
1	1	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[['https:
2	2	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[['http
3	3	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	22 hours ago	
4	4	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	2 days ago	[['http
...	
11649	11649	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	8 months ago	
11650	11650	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[['https:/
11651	11651	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11652	11652	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11653	11653	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[['https

11654 rows × 14 columns



```
In [36]: dataset.isnull().sum().sum()
```

```
Out[36]: 0
```

Handling the noisy data

```
In [37]: min_value = dataset['reactions'].min()  
max_value = dataset['reactions'].max()
```

```
In [38]: bins = np.linspace(min_value,max_value,25)  
labels = bins[1:]
```

```
In [39]: dataset['reactions_Bin'] = pd.cut(dataset['reactions'], bins=bins, labels=1
dataset
```

Out[39]:

	Unnamed: 0	name	headline	location	connections	about	time_spent	
0	0	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	1 week ago	[[http
1	1	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[[https:
2	2	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	20 hours ago	[[http
3	3	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	22 hours ago	
4	4	Dell Technologies	Information Technology & Services	Round Rock, Texas	194,140	Dell Technologies is a unique family of busine...	2 days ago	[[http
...	
11649	11649	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	8 months ago	
11650	11650	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[[https:/
11651	11651	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11652	11652	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	
11653	11653	John Deere	Machinery	Moline, IL	39,682	John Deere is a world leader in providing adva...	1 year ago	[[https

11654 rows × 15 columns

