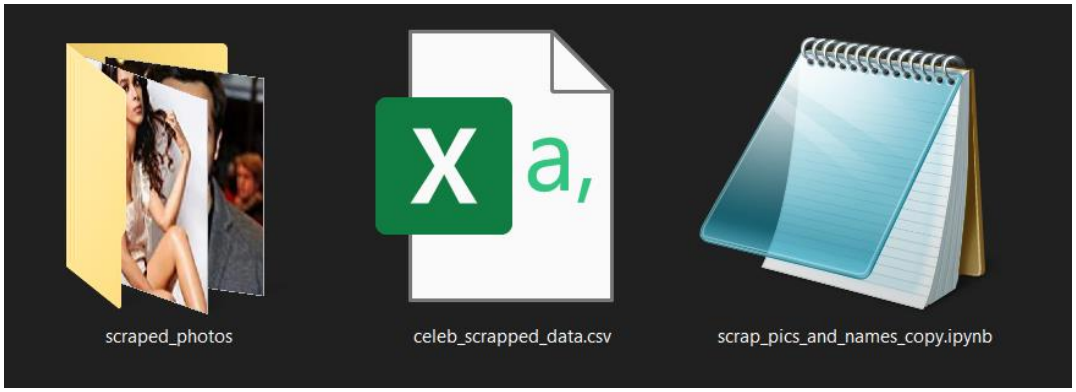
#Note : Coding part is done in Ubuntu, Please test my codes in Ubuntu OS.

I have divided the task given by bipolar factory into 3 parts:

## Part 1 (Folder: 1_crawling_celebrity_data_and_saving_as_csv)

- In this folder, there are 3 things:




1. **scrapped_photos** folder is used to store celebrity images
2. **celeb_scrapped_data.csv** is used to store dataset of celebrities containing
   Image as a numpy array
   Celebrity Name
   Celebrity Information
3. **Scrap_pics_and_names_copy.ipynb** is the script which grabs Image and data of celebrities and generate folder and csv file.

I have used BeautifulSoup library for data scraping.

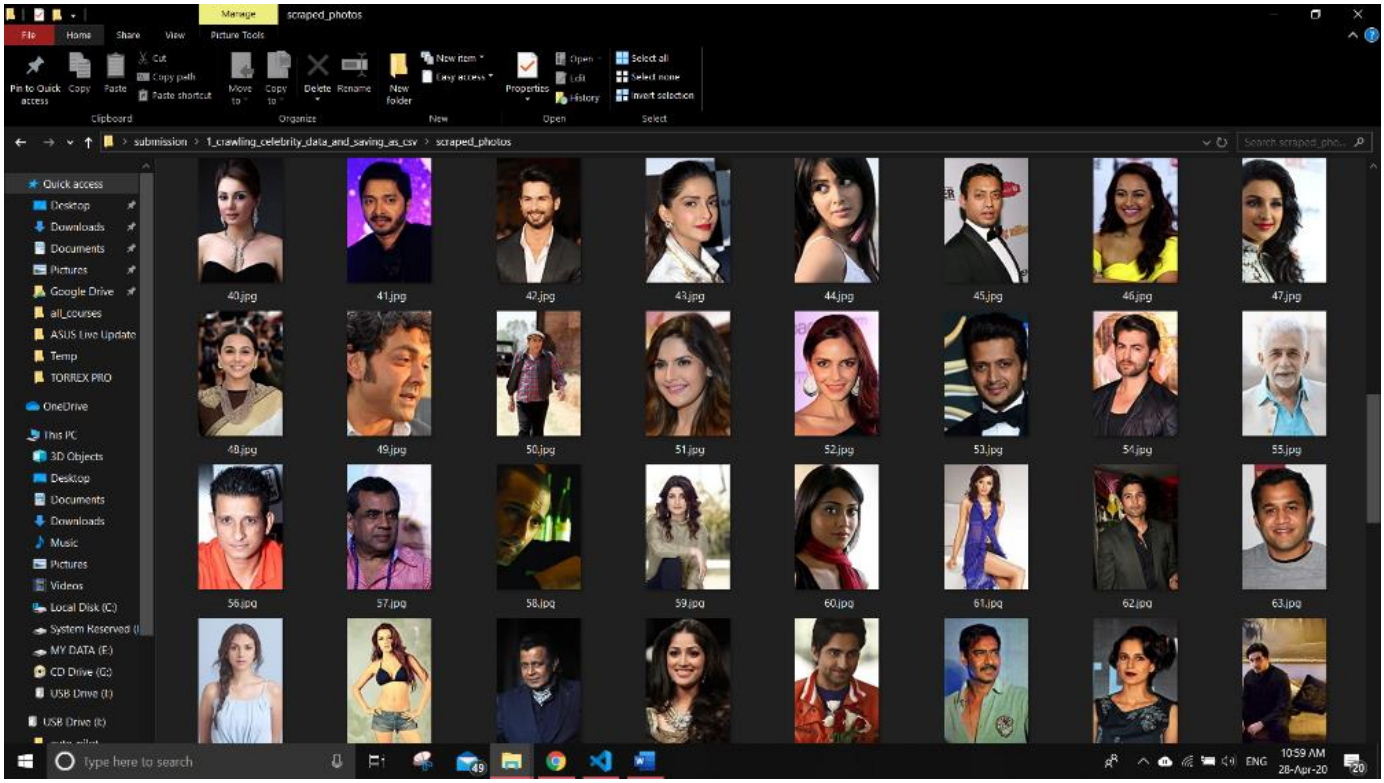
*A glimpse of final data (Image is stored as numpy array)

```
final_data[0:5]
```

| | Images | Name | Biodata |
|---|---|---|---|
| 0 | [[[160, 172, 174], [160, 172, 174], [161, 173,... | shah rukh khan | Shahrukh Khan was born on 2 November 1965 in N... |
| 1 | [[[15, 71, 76], [15, 71, 76], [14, 72, 77], [1... | aamir khan | Aamir is no doubt one of the most dedicated ac... |
| 2 | [[[250, 250, 250], [250, 250, 250], [250, 250,... | salman khan | Born on December 27, 1965 Salman Khan is the s... |
| 3 | [[[82, 84, 102], [81, 83, 101], [81, 84, 99], ... | katrina kaif | Katrina Kaif is one of eight siblings, 7 girls... |
| 4 | [[[6, 1, 3], [6, 1, 3], [6, 1, 3], [6, 1, 3], ... | kareena kapoor | Kareena was born to Sindhi-speaking Babita (ne... |

This Data will later be used to predict different personality trait classes

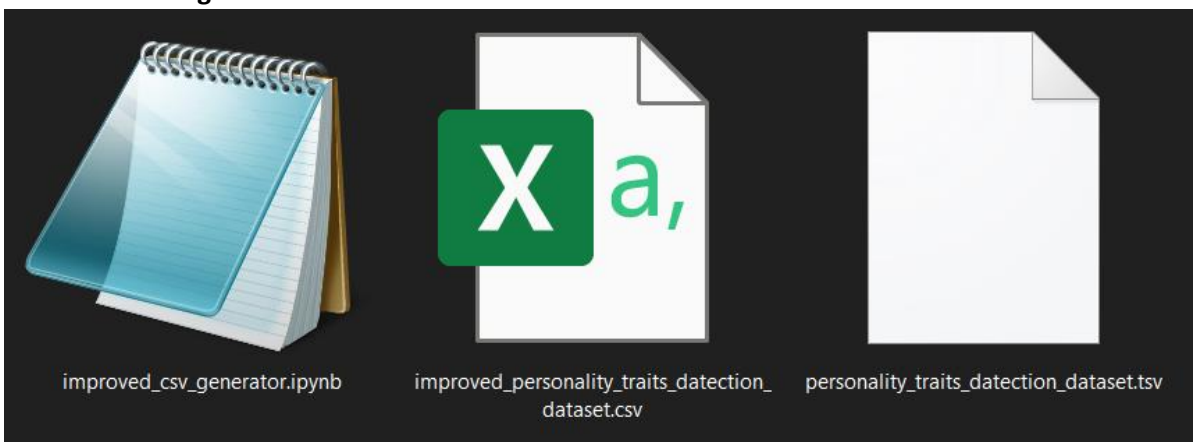
*A glimpse of scraped images



## Part 2 (Folder: 2_generating_labelled_data_for_training_and_saving_as_csv)

This module consists of work related to loading a dataset, cleaning it properly and then saving cleaned dataset back to the same folder

- **There are 3 things in this folder:**




1. **Improved_csv_generator.ipynb** contains code which loads a uncleaned dataset (personality_traits_datection_dataset.tsv) and then converts it into a cleaned dataset saved as (improved_personality_traits_datection_dataset.csv) in the same folder.
2. **personality_traits_datection_dataset.tsv** is the uncleaned dataset
3. **improved_personality_traits_datection_dataset.csv** is cleaned dataset

*A snap of cleaned dataset:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | Review | Extroversion | Neuroticism | Agreeableness | Conscientiousness | Openness |
| 2 | 0 | "Well, right now I just woke up from a mi | 0 | 1 | 1 | 0 | 1 |
| 3 | 1 | Well, here we go with the stream of cons | 0 | 0 | 1 | 0 | 0 |
| 4 | 2 | "An open keyboard and buttons to push. | 0 | 1 | 0 | 1 | 1 |
| 5 | 3 | I can't believe it! It's really happening! M | 1 | 0 | 1 | 1 | 0 |
| 6 | 4 | "Well, here I go with the good old stream | 1 | 0 | 1 | 0 | 1 |
| 7 | 5 | Today. Had to turn the music down. Tod | 1 | 0 | 1 | 0 | 1 |
| 8 | 6 | Stream of consciousness. What should I | 0 | 0 | 1 | 0 | 0 |
| 9 | 7 | "The RTF305 Usenet site is a piece of gar | 0 | 0 | 0 | 1 | 1 |
| 10 | 8 | Today was a tough day for me. I can't be | 1 | 1 | 1 | 1 | 0 |
| 11 | 9 | Well, I am sitting in the library right now, | 1 | 1 | 1 | 1 | 1 |
| 12 | 10 | I have done this assignment three times i | 0 | 0 | 0 | 0 | 0 |
| 13 | 11 | well I am just sitting here thinking about I | 0 | 1 | 0 | 0 | 0 |
| 14 | 12 | "Ok I've put this off long enough and you | 0 | 1 | 1 | 0 | 0 |
| 15 | 13 | sitting here just writing stuff down on pa | 0 | 0 | 1 | 1 | 0 |
| 16 | 14 | always a problem. My hair is really wet a | 1 | 0 | 1 | 0 | 0 |
| 17 | 15 | "Psychologists. Always trying to understa | 0 | 0 | 0 | 0 | 1 |
| 18 | 16 | 1 Freestyle- trying to write down though | 0 | 0 | 1 | 0 | 1 |
| 19 | 17 | Well, I feel good about the fact that I | 0 | 0 | 1 | 1 | 1 |
| 20 | 18 | "Okay here it goes. I am freezing in this c | 1 | 1 | 1 | 0 | 1 |
| 21 | 19 | I miss the way my life used to be a little b | 0 | 0 | 1 | 0 | 0 |
| 22 | 20 | I don't want to be in ROTC, but I have to | 0 | 1 | 1 | 0 | 1 |
| 23 | 21 | My neighbor from across the hall is lettin | 0 | 1 | 1 | 1 | 0 |
| 24 | 22 | "I'm feeling jealous right now. I got an en | 1 | 0 | 1 | 1 | 0 |
| 25 | 23 | Wow, this day has been hectic. I feel reli | 1 | 0 | 1 | 1 | 1 |
| 26 | 24 | As I sit here in my dorm room, I am thinki | 1 | 0 | 1 | 1 | 0 |
| 27 | 25 | I just got off AOL with my best friend for | 1 | 0 | 1 | 1 | 0 |
| 28 | 26 | "I have been typing friends and family fo | 1 | 1 | 0 | 0 | 1 |
| 29 | 27 | "Okay, Ï¿½m not so sure where to begin. | 1 | 1 | 0 | 1 | 1 |

**We can see that there are 5 classes in cleaned dataset** (Personality traits classes)

1) Extroversion
2) Neuroticism
3) Agreeableness
4) Conscientiousness
5) Openness

**Challenge faced by me:**

To predict Personality traits of various celebrities, we need a dataset with labelled Personality classes, because making dataset of thousands of people and labelling them manually is a very tedious work

So, I found this dataset of around 2500 rows having Personality trait labels: Extroversion, Neuroticism. Agreeableness, Conscientiousness and Openness on GitHub (https://github.com/SenticNet/personality-detection)

This page also contains implementation that can predict directly using terminal commands.
They are using a **pre-trained-Word2Vec-Embeddings** from Google. (Transfer learning)

I'm not using their pretrained model and my approach and code is different.

I'm only using 'essays.csv' file that contains uncleaned labelled data for 2500 people (because taking data individually and labelling them from thousands of people is very time consuming) and then I converted them into embeddings.

My dataset is very small and I have created my own word embeddings instead of using transfer learning to demonstrate each section of code.

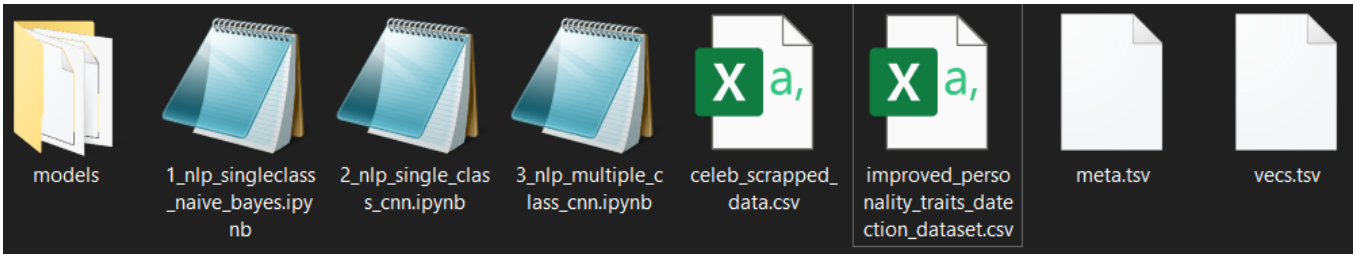
## Part 3

## (Folder: 3_model_training_on_labelled_data_and_predicting_personality_traits_on_celebrity_data)

I have used 2 approach in this module to predict Personality Traits of Celebrities, namely

1) Naïve Bayes
2) Neural networks

- There are 8 things in this folder




1. **Celeb_scrapped_data.csv** is a file that we have extracted from module 1's code, this file will be used to predict Personality trait classes of celebrities.
2. **improved_personality_traits_datection_dataset.csv** is a file that we have extracted from module 2's code, this file contains labelled data used in training process.
3. **1_nlp_singleclass_naive_bayes.ipynb** contains code for training using Naïve Bayes- Bag of words method (without dimensionality reduction) to predict a single class of personality trait , i.e. ('Extrovertness')
4. **2_nlp_single_class_cnn.ipynb** contains code for training using Neural network on self-generated embeddings to predict a single class of personality trait , i.e. ('Extrovertness')
   **In this file I have trained models using 5 different approaches**
   1) Using GRU Layer (For sequential learning)
   2) Using LSTM layer (For Sequential learning)
   3) Using CNN and Global average pooling layer
5. **3_nlp_multiple_class_cnn.ipynb** contains code for training Neural Network on self-generated embeddings to predict all 5 Personality trait classes (Extroversion, Neuroticism. Agreeableness, Conscientiousness and Openness)

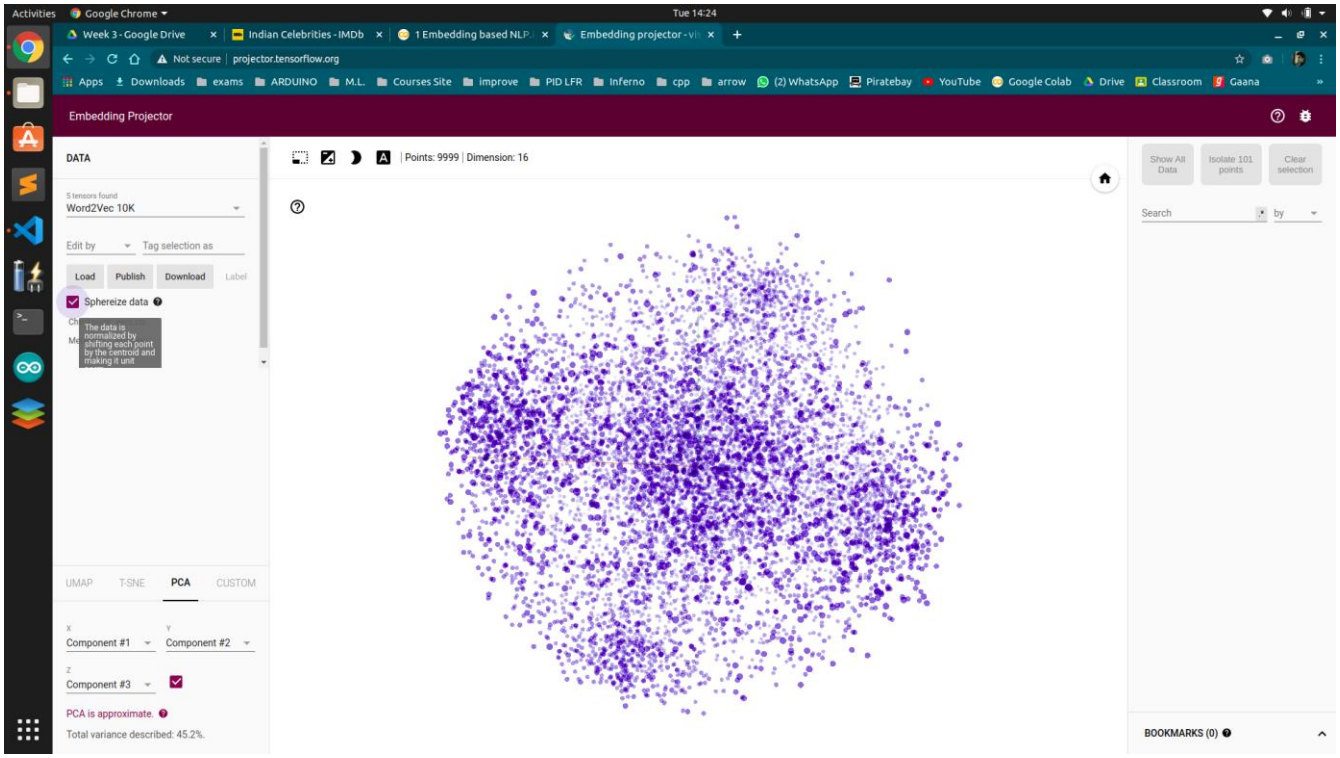
   Here we can't simply use multiclass method to predict personalities because a celebrity can have more than one type of personality traits. But multiclass' output is a single value (Single type of personality trait)
   Making a classifier with more than 1 output is a complex task, so to fix this problem
   I trained 5 different Sequential learning Neural network models corresponding to each personality trait class where each of them having one output node.
   You can see this in the next page:

```
model_Extroversion = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),
    tf.keras.layers.Bidirectional(tf.keras.layers.GRU(32)),
    tf.keras.layers.Dense(6, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model_Extroversion.compile(loss='binary_crossentropy',optimizer='adam',metrics=
['accuracy'])



model_Neuroticism = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),
    tf.keras.layers.Bidirectional(tf.keras.layers.GRU(32)),
    tf.keras.layers.Dense(6, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model_Neuroticism.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy']
)



model_Agreeableness = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),
    tf.keras.layers.Bidirectional(tf.keras.layers.GRU(32)),
    tf.keras.layers.Dense(6, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model_Agreeableness.compile(loss='binary_crossentropy',optimizer='adam',metrics=
['accuracy'])



model_Conscientiousness = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),
    tf.keras.layers.Bidirectional(tf.keras.layers.GRU(32)),
    tf.keras.layers.Dense(6, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model_Conscientiousness.compile(loss='binary_crossentropy',optimizer='adam',metrics=
['accuracy'])



model_Openness = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),
    tf.keras.layers.Bidirectional(tf.keras.layers.GRU(32)),
    tf.keras.layers.Dense(6, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model_Openness.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
```

6. **models** folder contains all saved models(Architecture and weights) that are created after my model training . this models can be directly to save training time

7. **meta.tsv** and **vecs.tsv** are two files generated by code that can be loaded to TensorFlow website to plot 16-D embeddings in a 3D  spherical space using PCA.

   The results are here:




We can see that embeddings are not that much distinct because we have used very small training set due to lack of resources.
I have also provided a video of this plot in the main folder

## Final Result (Prediction)

[34]

```
dataset
```

| | Name | Biodata | Extroversion | Neuroticism | Agreeableness | Conscientiousness | Openness |
|---|---|---|---|---|---|---|---|
| 0 | shah rukh khan | Shahrukh Khan was born on 2 November 1965 in N... | 65.174576 | 0.334661 | 100.000000 | 76.878670 | 99.999199 |
| 1 | aamir khan | Aamir is no doubt one of the most dedicated ac... | 6.999949 | 99.976631 | 100.000000 | 76.878670 | 23.500147 |
| 2 | salman khan | Born on December 27, 1965 Salman Khan is the s... | 73.561378 | 0.023484 | 100.000000 | 0.929405 | 99.337097 |
| 3 | katrina kaif | Katrina Kaif is one of eight siblings, 7 girls... | 0.019341 | 99.865402 | 99.999580 | 57.087261 | 80.293938 |
| 4 | kareena kapoor | Kareena was born to Sindhi-speaking Babita (ne... | 80.860443 | 99.957886 | 99.996651 | 0.005028 | 79.838287 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 95 | emraan hashmi | Emraan Hashmi was born 24 March 1979) is an In... | 74.184242 | 99.915466 | 99.988113 | 25.738808 | 79.208809 |
| 96 | mallika sherawat | Mallika Sherawat was born on October 24, 1976 ... | 1.340078 | 55.362026 | 99.999985 | 76.878670 | 81.229927 |
| 97 | nawazuddin siddiqui | Nawazuddin Siddiqui (born 1974) also known as ... | 53.767990 | 86.853424 | 99.997040 | 0.001022 | 96.537247 |
| 98 | aditya roy kapoor | Aditya Roy Kapoor was born on November 16, 198... | 2.040210 | 4.592927 | 99.993683 | 0.000343 | 23.500147 |
| 99 | sunny leone | Born in Ontario, Canada, Sunny Leone grew up i... | 72.562492 | 7.890504 | 100.000000 | 76.878670 | 23.500147 |

100 rows × 7 columns