Author: Vishal Mundlye, https://www.linkedin.com/in/vishal-mundlye
Created for: Machine Learning Course, https://www.oursecondinnings.org
Project Name: Machine learning based personality prediction using (MBTI) Myers-Briggs Personality Type

1. What is MBTI?

Stands for Myers-Briggs Type Indicator, measure of personality type based on the work of psychologist Carl Jung. Isabel Myers developed the MBTI during the Second World War to facilitate better working relationships between health care professionals, particularly nurses. Based on Jung's theory of "individual preference", meaning - different ways in which individuals prefer to use their minds (Ref: https://www.ncbi.nlm.nih.gov/books/NBK554596/).

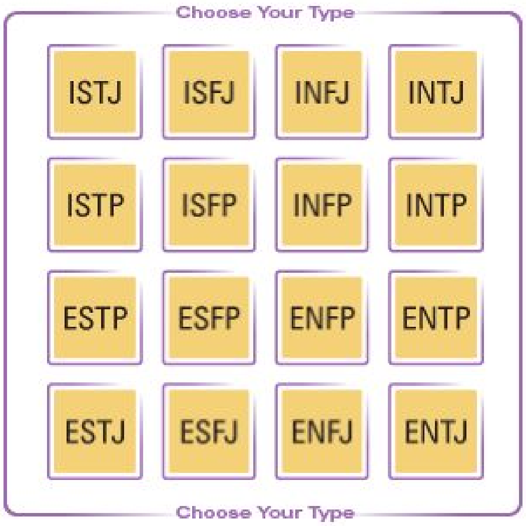
2. Why is MBTI used?

Increase awareness of oneself as well as others. By generating an understanding of one's individual preferences, one can begin analyzing and applying those preferences in work and personal endeavors.

3. How is this indicator calculated?

	Aspects	
Energy	**Extraversion**: direct their attention to external experiences and actions, deriving energy from those around them	**Introversion**: direct their attention towards inner thoughts and ideas, acquiring energy from solitude
Perceiving	**Sensing**: gather information using the five senses, require gathering facts before understanding general ideas and patterns	**Intuition**: rely on instincts and view problems from the "big picture" perspective, realizing general patterns before identifying constituent facts
Judging	**Thinkers**: rely on logic and facts	**Feelers**: rely on seeking harmony in the resolution of an issue
Orientation/Structure	**Judging**: tend towards orderly, decisive, and settled lifestyle	**Perceiving**: prefer a more flexible, unpredictable existence

4. Mapping of this indicator:



Choose Your Type

ISTJ	ISFJ	INFJ	INTJ
ISTP	ISFP	INFP	INTP
ESTP	ESFP	ENFP	ENTP
ESTJ	ESFJ	ENFJ	ENTJ

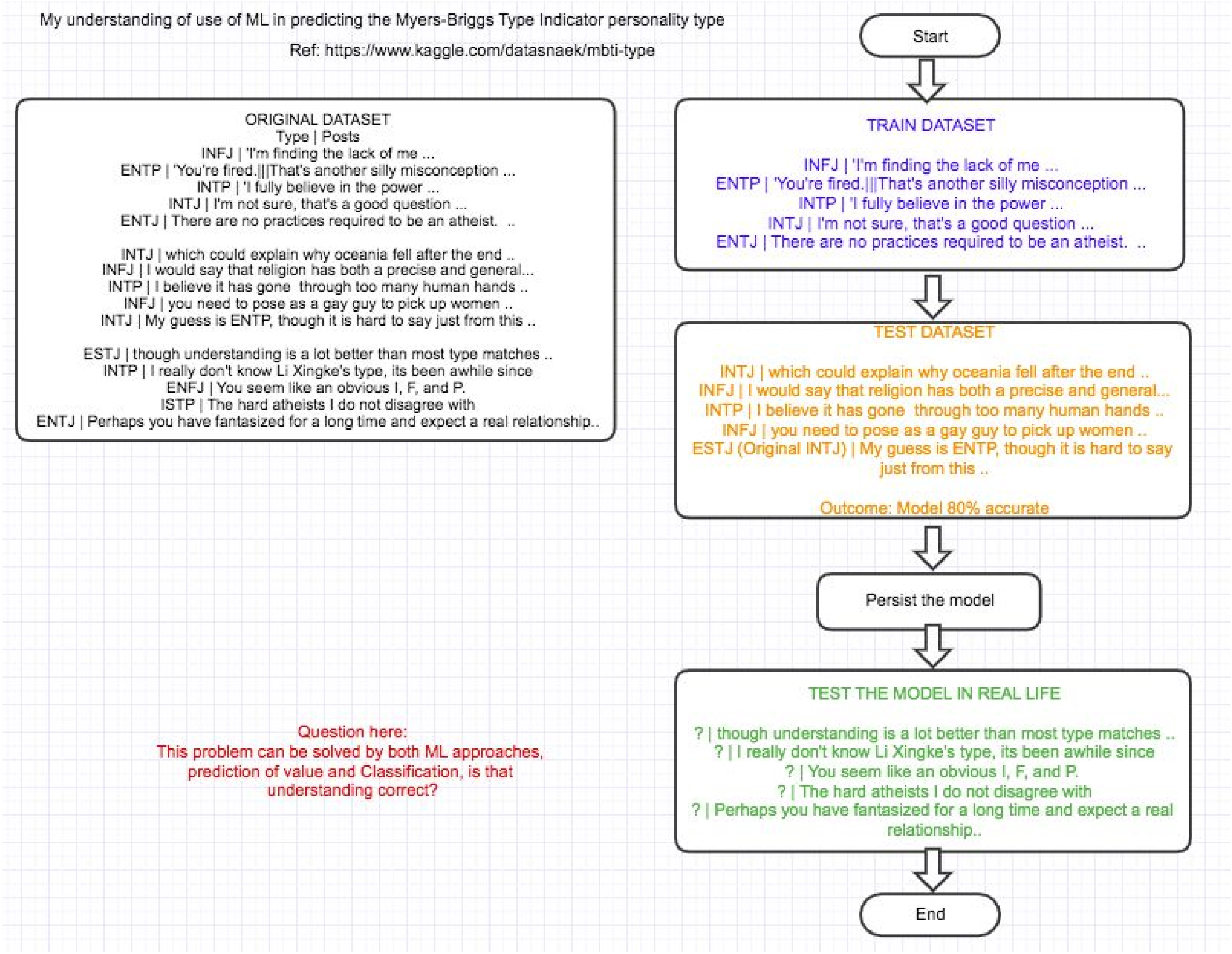
Choose Your Type

Ref: https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/

5. What is the Machine Learning problem statement?

The publicly available dataset hosts personality “**type**” of 8600 users with 50 of their last tweeted “**posts**”. The ***objective is to learn from this data set and predict or classify the personality types*** of new user/person based on similar information available.

6. High level approach to solving the problem:



My understanding of use of ML in predicting the Myers-Briggs Type Indicator personality type

Ref: https://www.kaggle.com/datasnaek/mbti-type

ORIGINAL DATASET
Type | Posts
INFJ | 'I'm finding the lack of me ...
ENTP | 'You're fired.|||That's another silly misconception ...
INTP | 'I fully believe in the power ...
INTJ | I'm not sure, that's a good question ...
ENTJ | There are no practices required to be an atheist. ..

INTJ | which could explain why oceania fell after the end ..
INFJ | I would say that religion has both a precise and general...
INTP | I believe it has gone through too many human hands ..
INFJ | you need to pose as a gay guy to pick up women ..
INTJ | My guess is ENTP, though it is hard to say just from this ..

ESTJ | though understanding is a lot better than most type matches ..
INTP | I really don't know Li Xingke's type, its been awhile since
ENFJ | You seem like an obvious I, F, and P.
ISTP | The hard atheists I do not disagree with
ENTJ | Perhaps you have fantasized for a long time and expect a real relationship..

Start

TRAIN DATASET

INFJ | 'I'm finding the lack of me ...
ENTP | 'You're fired.|||That's another silly misconception ...
INTP | 'I fully believe in the power ...
INTJ | I'm not sure, that's a good question ...
ENTJ | There are no practices required to be an atheist. ..

TEST DATASET

INTJ | which could explain why oceania fell after the end ..
INFJ | I would say that religion has both a precise and general...
INTP | I believe it has gone through too many human hands ..
INFJ | you need to pose as a gay guy to pick up women ..
ESTJ (Original INTJ) | My guess is ENTP, though it is hard to say just from this ..

Outcome: Model 80% accurate

Persist the model

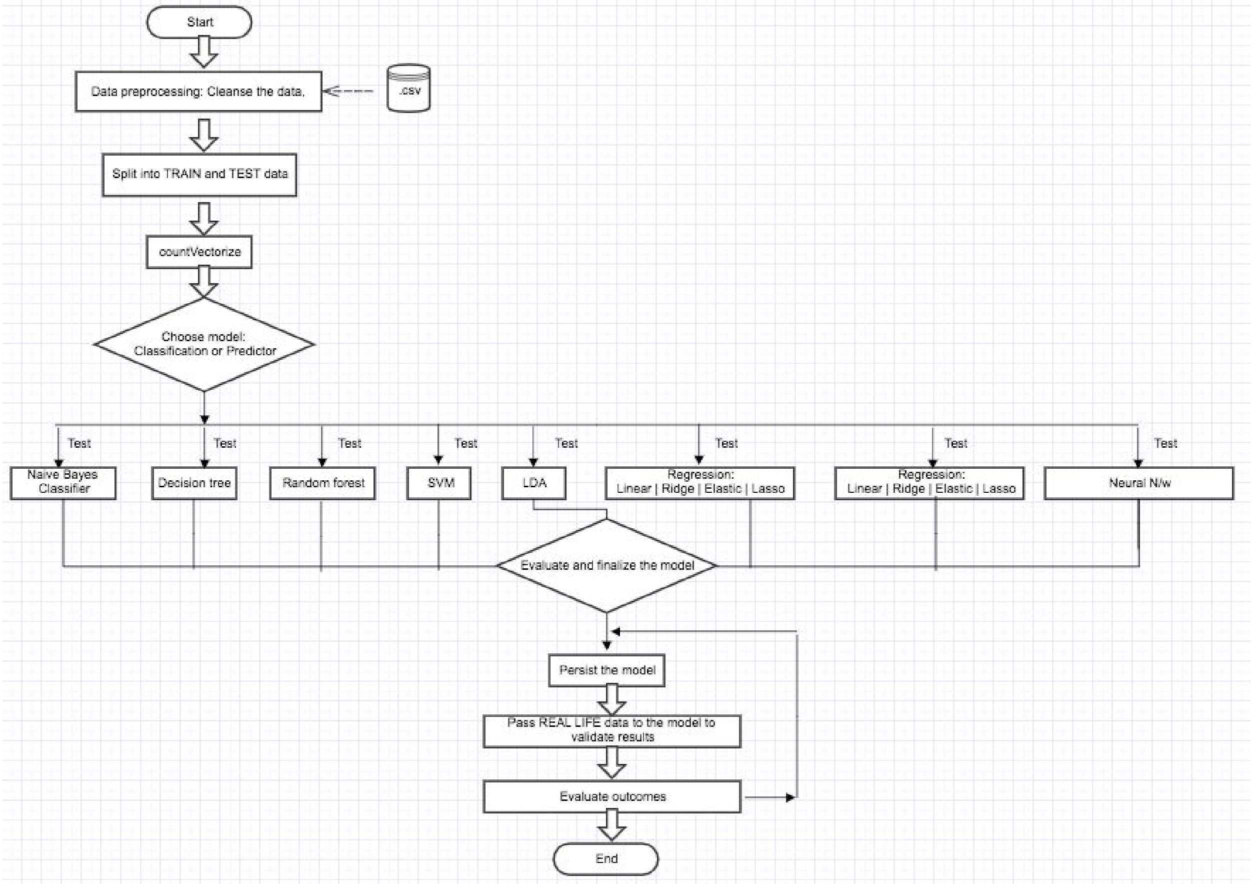
TEST THE MODEL IN REAL LIFE

? | though understanding is a lot better than most type matches ..
? | I really don't know Li Xingke's type, its been awhile since
? | You seem like an obvious I, F, and P.
? | The hard atheists I do not disagree with
? | Perhaps you have fantasized for a long time and expect a real relationship..

End

Question here:
This problem can be solved by both ML approaches, prediction of value and Classification, is that understanding correct?

7. Detailed level approach - using aspects of Machine Learning learned in the class:


Start
Data preprocessing: Cleanse the data.
.csv
Split into TRAIN and TEST data
countVectorize
Choose model:
Classification or Predictor
Test
Test
Test
Test
Test
Test
Test
Test
Naive Bayes Classifier
Decision tree
Random forest
SVM
LDA
Regression:
Linear | Ridge | Elastic | Lasso
Regression:
Linear | Ridge | Elastic | Lasso
Neural N/w
Evaluate and finalize the model
Persist the model
Pass REAL LIFE data to the model to validate results
Evaluate outcomes
End

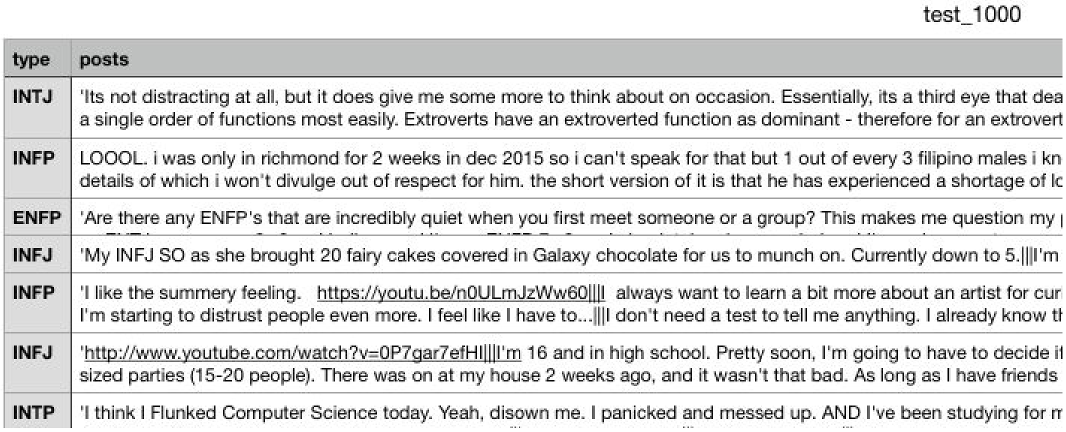
8. Environment and data set up:
Environment: MacOS Sierra 10.12.06 | 8 Gigs RAM | 125 Gigs hard disk
Python: 3.8.5
Editor: VSCODE

Full data set: 8600 records, .csv file, 62 MB,
(link: https://www.kaggle.com/datasnaek/mbti-type)

Scaled down version for demo purpose:
Training dataset: 1000 records
Testing dataset: 10 records

Reason for scaling; certain models take hours to train and predict on an 8 gigs RAM and you don’t want to jeopardize your demo.

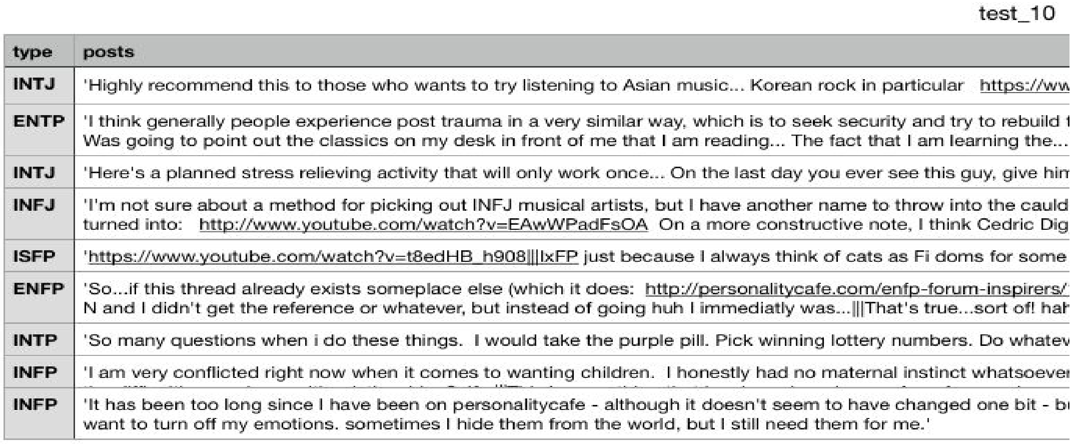
“1000” training dataset:



test_1000

type	posts
INTJ	'Its not distracting at all, but it does give me some more to think about on occasion. Essentially, its a third eye that dea a single order of functions most easily. Extroverts have an extroverted function as dominant - therefore for an extrovert
INFP	LOOOL. i was only in richmond for 2 weeks in dec 2015 so i can't speak for that but 1 out of every 3 filipino males i kn details of which i won't divulge out of respect for him. the short version of it is that he has experienced a shortage of lc
ENFP	'Are there any ENFP's that are incredibly quiet when you first meet someone or a group? This makes me question my
INFJ	'My INFJ SO as she brought 20 fairy cakes covered in Galaxy chocolate for us to munch on. Currently down to 5.|||I'm
INFP	'I like the summery feeling. https://youtu.be/n0ULmJzWw60|||I always want to learn a bit more about an artist for cur I'm starting to distrust people even more. I feel like I have to...|||I don't need a test to tell me anything. I already know tl
INFJ	'http://www.youtube.com/watch?v=0P7gar7efHI|||I'm 16 and in high school. Pretty soon, I'm going to have to decide if sized parties (15-20 people). There was on at my house 2 weeks ago, and it wasn't that bad. As long as I have friends
INTP	'I think I Flunked Computer Science today. Yeah, disown me. I panicked and messed up. AND I've been studying for m

“10” testing dataset: tally the final prediction of the model (at the end of presentation) with this data

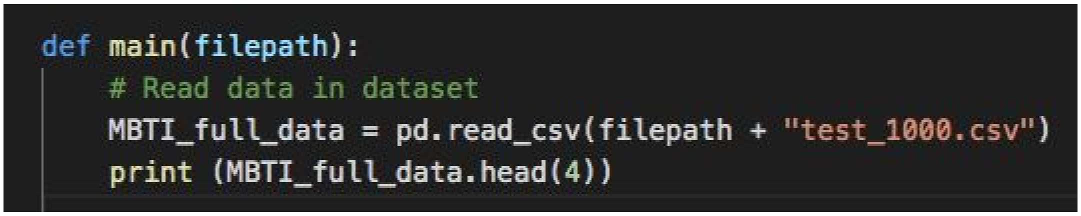


test_10

type	posts
INTJ	'Highly recommend this to those who wants to try listening to Asian music... Korean rock in particular https://ww
ENTP	'I think generally people experience post trauma in a very similar way, which is to seek security and try to rebuild t Was going to point out the classics on my desk in front of me that I am reading... The fact that I am learning the...
INTJ	'Here's a planned stress relieving activity that will only work once... On the last day you ever see this guy, give him
INFJ	'I'm not sure about a method for picking out INFJ musical artists, but I have another name to throw into the cauld turned into: http://www.youtube.com/watch?v=EAwWPadFsOA On a more constructive note, I think Cedric Dig
ISFP	'https://www.youtube.com/watch?v=t8edHB_h908|||IxFP just because I always think of cats as Fi doms for some
ENFP	'So...if this thread already exists someplace else (which it does: http://personalitycafe.com/enfp-forum-inspirers/ N and I didn't get the reference or whatever, but instead of going huh I immediatly was...|||That's true...sort of! hah
INTP	'So many questions when i do these things. I would take the purple pill. Pick winning lottery numbers. Do whatev
INFP	'I am very conflicted right now when it comes to wanting children. I honestly had no maternal instinct whatsoever
INFP	'It has been too long since I have been on personalitycafe - although it doesn't seem to have changed one bit - bi want to turn off my emotions. sometimes I hide them from the world, but I still need them for me.'

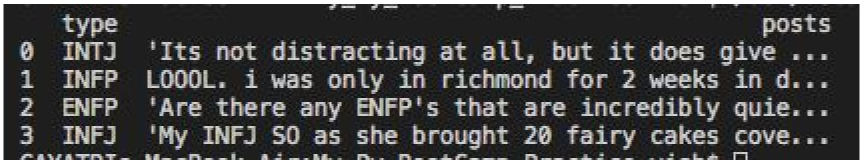
9. Code walk-through:

A. Loading the data set and cleaning the data:



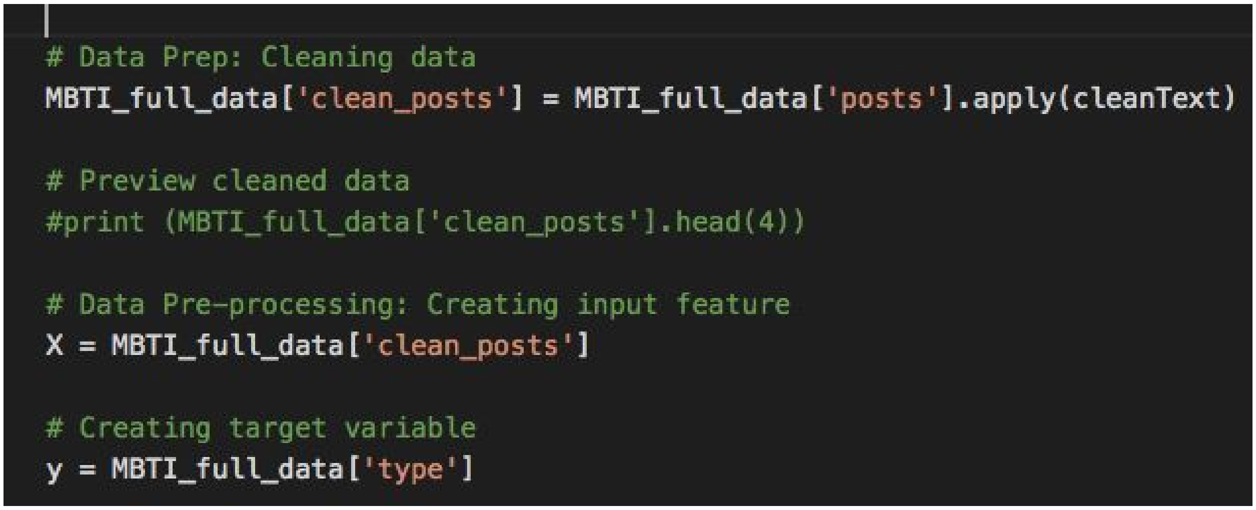
```
def main(filepath):
    # Read data in dataset
    MBTI_full_data = pd.read_csv(filepath + "test_1000.csv")
    print (MBTI_full_data.head(4))
```

Observe whether the data set has loaded as expected:



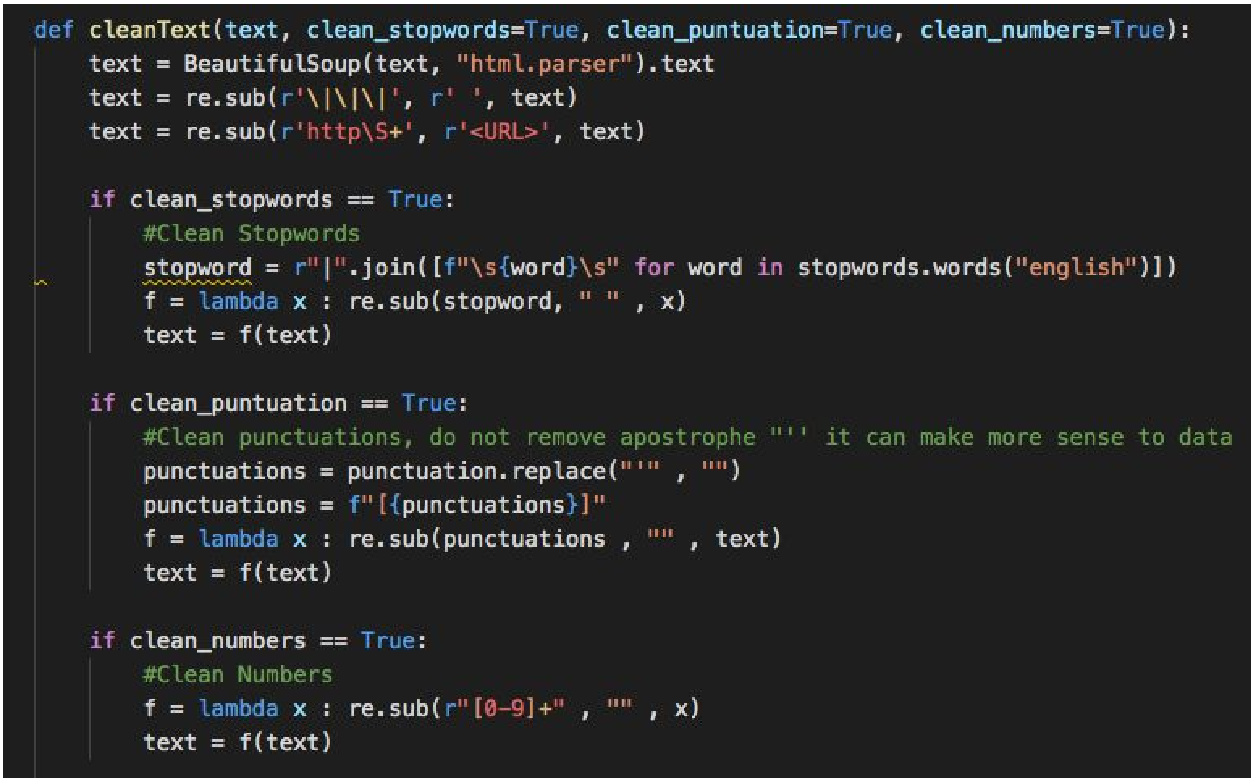
```
   type                                              posts
0  INTJ  'Its not distracting at all, but it does give ...
1  INFP  LOOOL. i was only in richmond for 2 weeks in d...
2  ENFP  'Are there any ENFP's that are incredibly quie...
3  INFJ  'My INFJ SO as she brought 20 fairy cakes cove...
```

B. Cleaning the data:



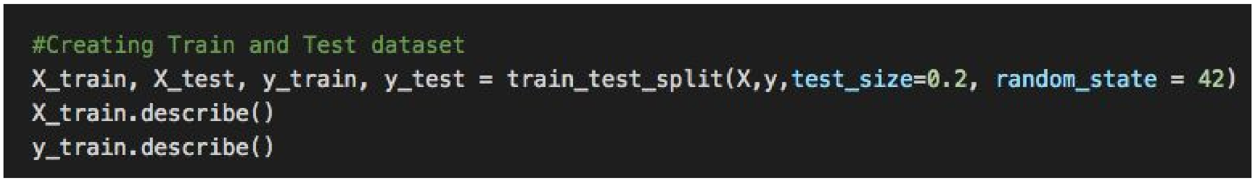
```
# Data Prep: Cleaning data
MBTI_full_data['clean_posts'] = MBTI_full_data['posts'].apply(cleanText)

# Preview cleaned data
#print (MBTI_full_data['clean_posts'].head(4))

# Data Pre-processing: Creating input feature
X = MBTI_full_data['clean_posts']

# Creating target variable
y = MBTI_full_data['type']
```

- There are various approaches and methods to clean data, ignoring stopwords, removing punctuation marks, putting in a min/max tweet count filter across the data, etc. I have made use of the ones that I understood well and found relevant.



```python
def cleanText(text, clean_stopwords=True, clean_puntuation=True, clean_numbers=True):
    text = BeautifulSoup(text, "html.parser").text
    text = re.sub(r'\|\|\|', r' ', text)
    text = re.sub(r'http\S+', r'<URL>', text)

    if clean_stopwords == True:
        #Clean Stopwords
        stopword = r"|".join([f"\s{word}\s" for word in stopwords.words("english")])
        f = lambda x : re.sub(stopword, " " , x)
        text = f(text)

    if clean_puntuation == True:
        #Clean punctuations, do not remove apostrophe "'' it can make more sense to data
        punctuations = punctuation.replace("'" , "")
        punctuations = f"[{punctuations}]"
        f = lambda x : re.sub(punctuations , "" , text)
        text = f(text)

    if clean_numbers == True:
        #Clean Numbers
        f = lambda x : re.sub(r"[0-9]+" , "" , x)
        text = f(text)
```

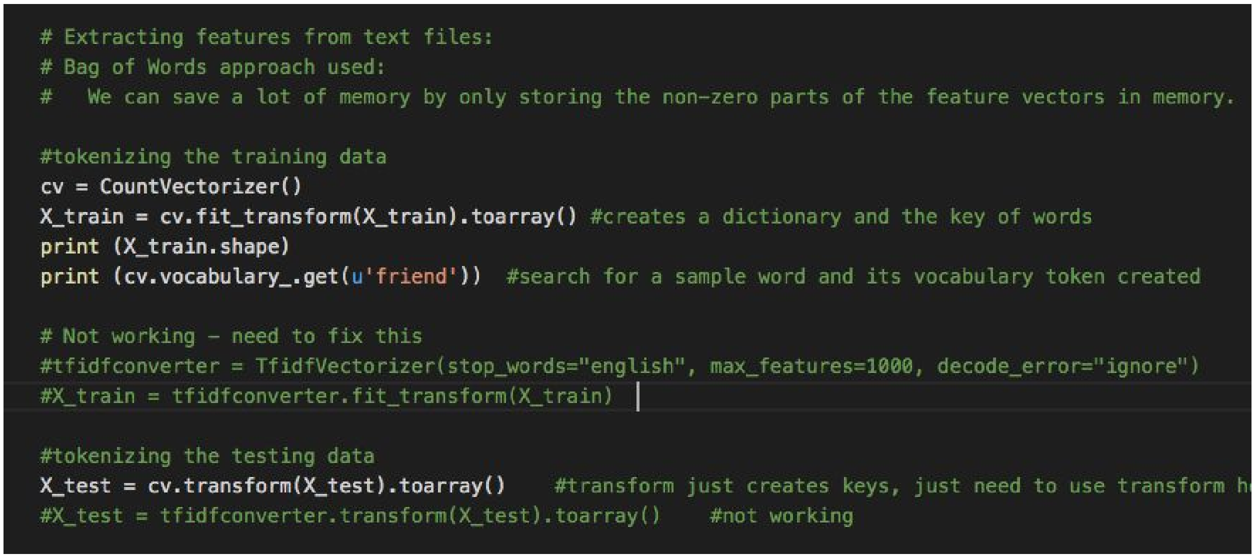
C. Preprocessing the data:



```
#Creating Train and Test dataset
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2, random_state = 42)
X_train.describe()
y_train.describe()
```

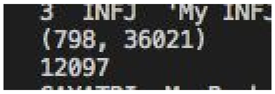
Involves separating training and testing dataset. In most of the ML examples, the testing and training dataset are demonstrated to be within the same file and hence split randomly so as not let the test set leak into training. However, I have demonstrated an example where these datasets reside in separate files.

D. Extracting features from the dataset:



```
# Extracting features from text files:
# Bag of Words approach used:
#   We can save a lot of memory by only storing the non-zero parts of the feature vectors in memory.

#tokenizing the training data
cv = CountVectorizer()
X_train = cv.fit_transform(X_train).toarray() #creates a dictionary and the key of words
print (X_train.shape)
print (cv.vocabulary_.get(u'friend'))  #search for a sample word and its vocabulary token created

# Not working - need to fix this
#tfidfconverter = TfidfVectorizer(stop_words="english", max_features=1000, decode_error="ignore")
#X_train = tfidfconverter.fit_transform(X_train)

#tokenizing the testing data
X_test = cv.transform(X_test).toarray()    #transform just creates keys, just need to use transform h
#X_test = tfidfconverter.transform(X_test).toarray()    #not working
```

Bag of words approach is used, saves memory. Verify the shape of training dataset
Output:

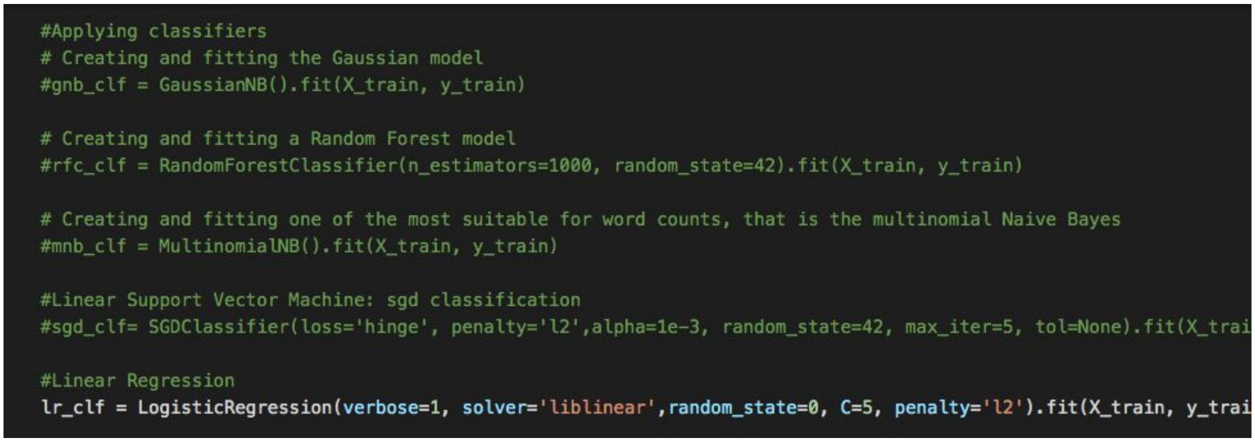


```
3  INFJ  'My INFJ
(798, 36021)
12097
```

Above values highlight ***n_samples***, ***n_features*** (number of words) in the training dataset

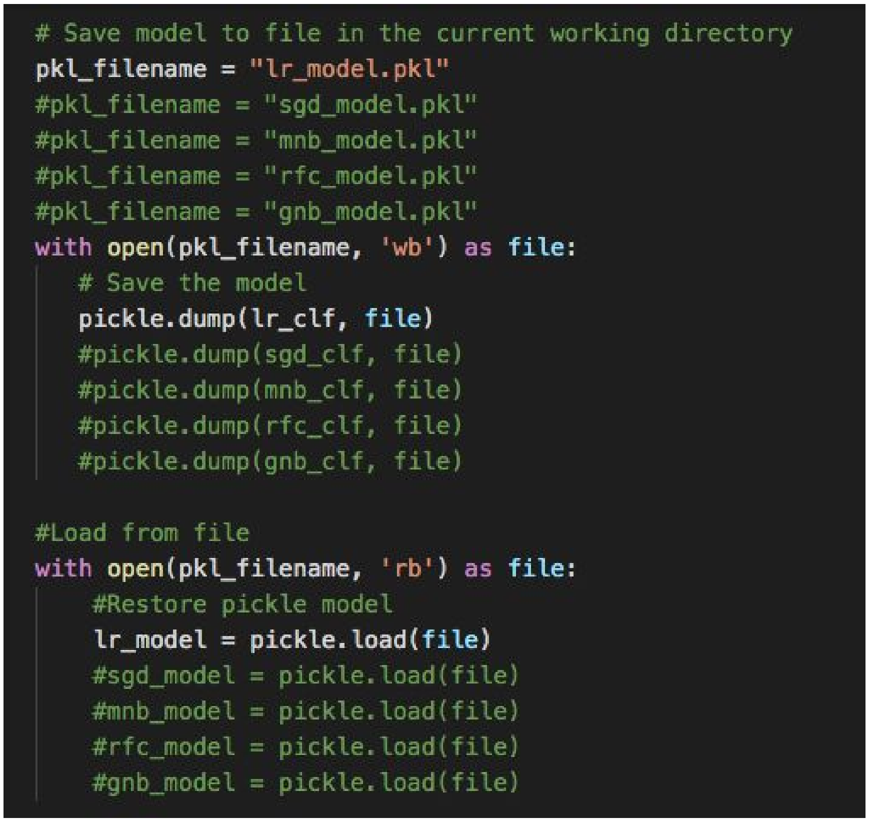
12097 represents the token of the word “friend” created by the vectorizer

E. Applying different predictors and classifiers:



```
#Applying classifiers
# Creating and fitting the Gaussian model
#gnb_clf = GaussianNB().fit(X_train, y_train)

# Creating and fitting a Random Forest model
#rfc_clf = RandomForestClassifier(n_estimators=1000, random_state=42).fit(X_train, y_train)

# Creating and fitting one of the most suitable for word counts, that is the multinomial Naive Bayes
#mnb_clf = MultinomialNB().fit(X_train, y_train)

#Linear Support Vector Machine: sgd classification
#sgd_clf= SGDClassifier(loss='hinge', penalty='l2',alpha=1e-3, random_state=42, max_iter=5, tol=None).fit(X_trai

#Linear Regression
lr_clf = LogisticRegression(verbose=1, solver='liblinear',random_state=0, C=5, penalty='l2').fit(X_train, y_trai
```

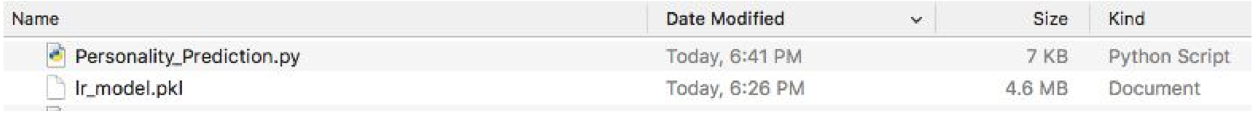
F. Saving the model and loading it:



```
# Save model to file in the current working directory
pkl_filename = "lr_model.pkl"
#pkl_filename = "sgd_model.pkl"
#pkl_filename = "mnb_model.pkl"
#pkl_filename = "rfc_model.pkl"
#pkl_filename = "gnb_model.pkl"
with open(pkl_filename, 'wb') as file:
   # Save the model
   pickle.dump(lr_clf, file)
   #pickle.dump(sgd_clf, file)
   #pickle.dump(mnb_clf, file)
   #pickle.dump(rfc_clf, file)
   #pickle.dump(gnb_clf, file)

#Load from file
with open(pkl_filename, 'rb') as file:
    #Restore pickle model
    lr_model = pickle.load(file)
    #sgd_model = pickle.load(file)
    #mnb_model = pickle.load(file)
    #rfc_model = pickle.load(file)
    #gnb_model = pickle.load(file)
```

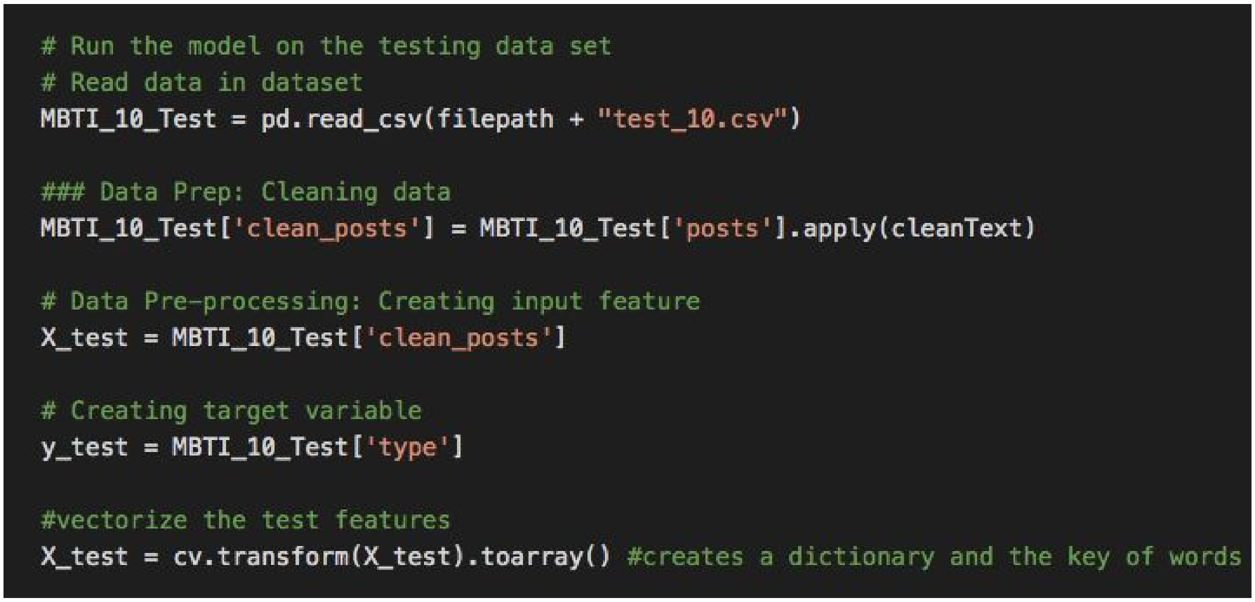
- Done to save processing time. Some of the models take a few hours to train, so we can save this model to a file and load it to predict.
- Pickle also helps you save the training parameters and scores as a tuple. Be watchful of the size of this file if saving as a tuple, it can run in gigs.
- This way, we can also use some of our team members' models shared using version control.

Output:



Name	Date Modified	Size	Kind
Personality_Prediction.py	Today, 6:41 PM	7 KB	Python Script
lr_model.pkl	Today, 6:26 PM	4.6 MB	Document

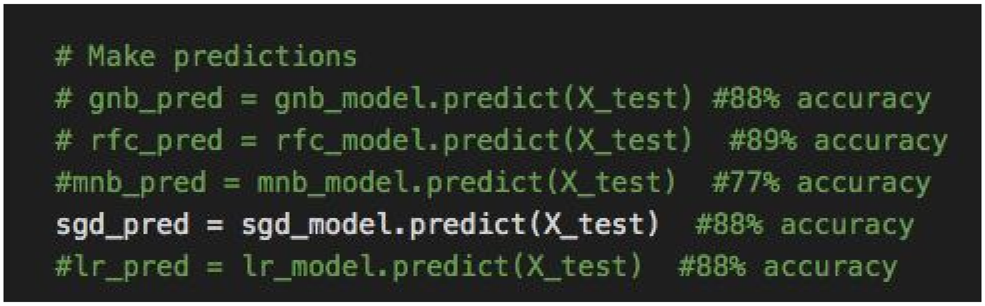
G. Running the model on the testing dataset:



```
# Run the model on the testing data set
# Read data in dataset
MBTI_10_Test = pd.read_csv(filepath + "test_10.csv")

### Data Prep: Cleaning data
MBTI_10_Test['clean_posts'] = MBTI_10_Test['posts'].apply(cleanText)

# Data Pre-processing: Creating input feature
X_test = MBTI_10_Test['clean_posts']

# Creating target variable
y_test = MBTI_10_Test['type']

#vectorize the test features
X_test = cv.transform(X_test).toarray() #creates a dictionary and the key of words
```

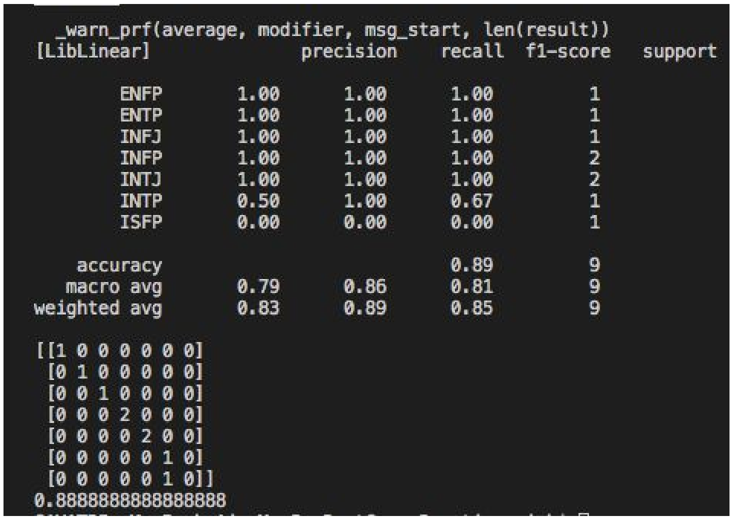
Remember: you need to ONLY transform the testing features, since the model is already stored

H. Make predictions



```
# Make predictions
# gnb_pred = gnb_model.predict(X_test) #88% accuracy
# rfc_pred = rfc_model.predict(X_test)  #89% accuracy
#mnb_pred = mnb_model.predict(X_test)  #77% accuracy
sgd_pred = sgd_model.predict(X_test)  #88% accuracy
#lr_pred = lr_model.predict(X_test)  #88% accuracy
```

I. Evaluate model performance:



```
  _warn_prf(average, modifier, msg_start, len(result))
[LibLinear]              precision    recall  f1-score   support

        ENFP       1.00      1.00      1.00         1
        ENTP       1.00      1.00      1.00         1
        INFJ       1.00      1.00      1.00         1
        INFP       1.00      1.00      1.00         2
        INTJ       1.00      1.00      1.00         2
        INTP       0.50      1.00      0.67         1
        ISFP       0.00      0.00      0.00         1

    accuracy                           0.89         9
   macro avg       0.79      0.86      0.81         9
weighted avg       0.83      0.89      0.85         9

[[1 0 0 0 0 0 0]
 [0 1 0 0 0 0 0]
 [0 0 1 0 0 0 0]
 [0 0 0 2 0 0 0]
 [0 0 0 0 2 0 0]
 [0 0 0 0 0 1 0]
 [0 0 0 0 0 1 0]]
0.8888888888888888
```

---End---