Internship Report On ‘Data Analytics, Machine Learning and AI with Python’

At

KUSHAGRA MITTAL

B.TECH, I.K.G. PUNJAB TECHNICAL UNIVERSITY

JUNE-JULY, 2019

CONTENT

SR. NO.	TOPIC	PAGE NO.
1.	ACKNOWLEDGEMENT	2
2.	COMPANY PROFILE	3
3.	INTRODUCTION	4-6
4.	DETAILS OF PROJECT	7-10
5.	DETAILS OF STUDY	11-17
6.	ANNEXURE A	18-21

ACKNOWLEDGEMENT

“Gratitude is not a thing of expression; it is more matters of feeling.”

There is always a sense of gratitude which one express towards others for their help and supervision in achieving the goals.

I would like to express my deep gratitude to Mr. Bipul Shahi , my training mentor for their constantco-operation. He was always there with his competent guidance and valuable suggestions throughout the pursuance of this research project.

COMPANY PROFILE

Diginique TechLabs, an IIT Roorkee alumni venture, is one of India’s leading organizations in the field of modern age technologies. Team Diginique has been working day and night towards the betterment of students in the field of technology and management and provides a steady source of highly skilled talent to the nation as well as overseas.

We constantly evolve our teaching methods, and provide quality workshops, excellent trainings and smart services to our students/clients, whom we see as unique individuals with different interests and aspirations. We keep the quality of our course content and trainers unparaDiginique TechLabs, an IIT Roorkee alumni venture, is one of India’s leading organizations in the field of modern age technologies. Team Diginique has been working day and night towards the betterment of students in the field of technology and management and provides a steady source of highly skilled talent to the nation as well as overseas.

INTRODUCTION

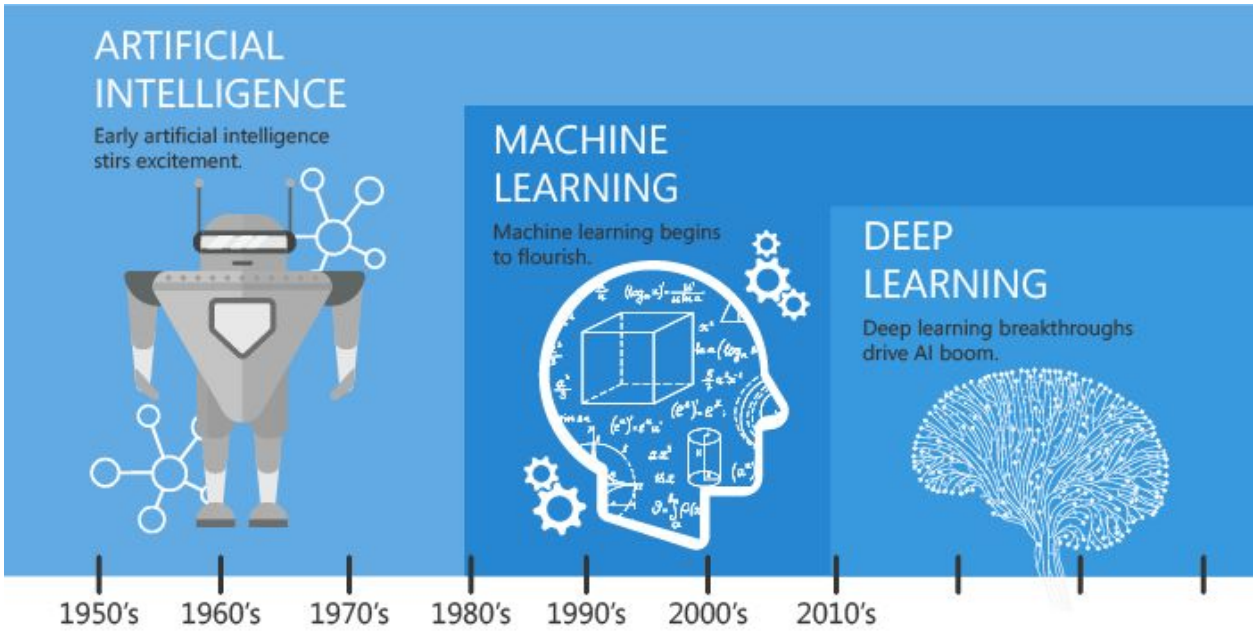
DATA SCIENCE: Data science is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

Big Data Analytics or Data Science is a very common term in IT industry because everyone knows this is some fancy term which is gonna help us to deal with this huge amount of data we are generating these days.

MACHINE LEARNING: **Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that which makes it more similar to humans: ***The ability to learn***. Machine learning is actively being used today, perhaps in many more places than one would expect.

- Prediction — Machine learning can also be used in the prediction systems. Considering the loan example, to compute the probability of a fault, the system will need to classify the available data in groups.
- Image recognition — Machine learning can be used for face detection in an image as well. There is a separate category for each person in a database of several people.

- Speech Recognition — It is the translation of spoken words into the text. It is used in voice searches and more. Voice user interfaces include voice dialing, call routing, and appliance control. It can also be used a simple data entry and the preparation of structured documents.
- Medical diagnoses — ML is trained to recognize cancerous tissues.
- Financial industry and trading — companies use ML in fraud investigations and credit checks.



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

ARTIFICIAL INTELLIGENCE: Artificial Intelligence is no longer a term one should treat to be non-pervasive in its impacts. To some extent, many including myself have stayed aloof of the whole concept of a world

takes over by automation and simulations. The reality is it will happen and what is more alarming is that it is happening as we speak.

There are predictions and theories that suggest that by 2050, circa 50% of human jobs in developed countries would be taken over by robots. Technological singularity, which is defined as the point when AI surpasses human intelligence will in fact occur for real and our generation will actually get to see how technology has disrupted the world around us.

But beyond the general euphoria around AI, it is important to understand what it actually means. Artificial Intelligence is when machines perform things for you. As machines these are not necessarily physical robots but computational codes that resemble the way human brain operates. Similar to how our brain

AI is a broad term for describing anything that is done by computers that humans could do. As a starter, data storage (floppys, storage buses etc) back in the days was one of the first versions of AI. I take the analogy of Science: Algorithms = AI: Machine learning. Machine learning is an approach to achieving Artificial Intelligence.

DETAILS OF PROJECT

Sentiment Analysis is the process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral. It’s also known as opinion mining, deriving the opinion or attitude of a speaker. Sentiment analysis is a type of data mining that measures the inclination of people’s opinions through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze subjective information from the Web - mostly social media and similar sources. The analyzed data quantifies the general public's sentiments or reactions toward certain products, people or ideas and reveal the contextual polarity of the information.Sentiment analysis is also known as opinion mining.

In this project, we will analyse each tweet of particular twitter handle. With the help of Twitter API , we will able to get tweets . First of all, we will setup Twitter API:

- We should have Twitter account for API then, login to the Developer Twitter .
- After login, click on the Create an App .
- By this, it will generate consumer key, secret , access key and secret.
- For it, we have to mail them all the details i.e. what is the purpose of it, etc to twitter team.

For this project, we mainly use ‘tweepy’ library for twitter API , ‘TextBlob’ library for sentiment analysis , ‘re’ library for regular expression like removing links , hashtags or username from tweets.

TextBlob: Simplified Text Processing :*TextBlob* is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Features

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Classification (Naive Bayes, Decision Tree)
- Language translation and detection powered by Google Translate
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- **n**-grams
- Word inflection (pluralization and singularization) and lemmatization
- Spelling correction
- Add new models or languages through extensions

- WordNet integrat

REGULAR EXPRESSION: Module Regular Expressions(RE) specifies a set of strings(pattern) that matches it.To understand the RE analogy, MetaCharacters are useful, important and will be used in functions of module re.

There are a total of 14 metacharacters and will be discussed as they follow into functions:

- \ Used to drop the special meaning of characters following it (discussed below)
- [] Represent a character class
- ^ Matches the beginning
- $ Matches the end
- . Matches any character except newline
- ? Matches zero or one occurrence.
- | Means OR (Matches with any of the characters separated by it.
- * Any number of occurrences (including 0 occurrences)
- + One more more occurrences
- {} Indicate number of occurrences of a preceding RE to match.
- () Enclose a group of REs

For code refer to Annexure A.

PANDAS: Pandas stands for “Python Data Analysis Library”. According to the Wikipedia page on Pandas, “the name is derived from the term “panel data”, an econometrics term for multidimensional structured data sets.” But I think it’s just a cute name to a super-useful Python library! Panda takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software (think Excel or SPSS for example.

- Convert a Python’s list, dictionary or Numpy array to a Pandas data frame
- Open a local file using Pandas, usually a CSV file, but could also be a delimited text file (like TSV), Excel, etc
- Open a remote file or database like a CSV or a JSONon a website through a URL or read from a SQL table/database

Goals of my project:

- Authorize twitter API client.
- Make a GET request to Twitter API to fetch tweets for a particular query.
- Parse the tweets. Classify each tweet as positive, negative or neutral.

DETAILS OF STUDY

Machine Learning is subcategory or subset of Artificial Intelligence. But

Machine Learning = building blocks/ model of data

We have a lot of Data ; i.e, structured or Unstructured data . But these data will have no meaning until without these models. ML involves building mathematical models to help data. Once these models fit on previous or current data then we can predict and understand aspects of new data observed.

There are two types of machine learning apparently,

i) **Supervised Learning**: It involves some modelling the relationship between features of data and label associated with data. Once model is recognized, it can be apply labels to new data.Models that can predicts labels based on labeled training data.

Supervised learning is divided into two parts :

a) **Classifications**: In which, labels are discrete categories.

b) **Regression**: In this, labels are continuous

ii) **Unsupervised Learning**: It involves modelling the features of dataset without reference to any label, and is often described as 'Letting the dataset speaks for itself.'

a)**Clustering**: It is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data

points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

b) **Dimensionality Reduction**: It is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

Now , we will study some of the libraries -

1. NumPY: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.
It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities.

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data.

Syntax: import numpy as nm

ndarray: It is also known as N-Dimensional Array type, Every item in an ndarray takes the same size of block in the memory. Each element in ndarray is an object of data-type object (called dtype).It have some attributes like :

Ndarray.shape , Ndarray.ndim, Ndarray.arange, Ndarray.linspace,

2.**Matplotlib**: Matplotlib is a plotting library for Python. It is used along with NumPy to provide an environment that is an effective open source alternative for MatLab.

Syntax: from matplotlib import pyplot as plt

plt.plot(x,y)

plt.bar(x, y, align='center', color='g')

plt.scatter(x , y)

plt.hist(y)

plt.show()

3.Seaborn: Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.

Syntax: import seaborn as sns

```
sns.relplot(x=”xlabel”, y=”ylabel”, data=mydata);

sns.countplot(y, data=mydata)
```

Scikit-Learn: It is a free software machine learning library for the Python programming language. It features various classification,regression and clustering algorithms including support vector machines,random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific librariesNumPy and SciPy.

Training And Test Data

```
>>> from sklearn.model_selection import train_test_split
>>> X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=0)
```

Create Your Model

Supervised Learning Estimators:

Linear Regression

```
>>> from sklearn.linear_model import LinearRegression
```

```
>>> lr = LinearRegression(normalize=True)
```

Support Vector Machines (SVM)

```
>>> from sklearn.svm import SVC
>>> svc = SVC(kernel='linear')
```

Naive Bayes

```
>>> from sklearn.naive_bayes import GaussianNB
>>> gnb = GaussianNB()
```

KNN

```
>>> from sklearn import neighbors
>>> knn = neighbors.KNeighborsClassifier(n_neighbors=5)
```

Unsupervised Learning Estimators:

Principal Component Analysis (PCA)

```
>>> from sklearn.decomposition import PCA
>>> pca = PCA(n_components=0.95)
```

K Means

```
>>> from sklearn.cluster import KMeans
>>> k_means = KMeans(n_clusters=3, random_state=0)
```

Model Fitting:

Supervised learning

```
>>> lr.fit(X, y)
>>> knn.fit(X_train, y_train)
>>> svc.fit(X_train, y_train)
```

Unsupervised Learning

```
>>> k_means.fit(X_train)
>>> pca_model = pca.fit_transform(X_train)
```

Prediction:

Supervised Estimators

```
>>> y_pred = svc.predict(np.random.random((2,5)))
>>> y_pred = lr.predict(X_test)
>>> y_pred = knn.predict_proba(X_test))
```

Unsupervised Estimators

```
>>> y_pred = k_means.predict(X_test)
```

Evaluate Your Model's Performance

Classification Metrics:

Accuracy Score

```
>>> knn.score(X_test, y_test)
>>> from sklearn.metrics import accuracy_score
>>> accuracy_score(y_test, y_pred)
```

Confusion Matrix

```
>>> from sklearn.metrics import confusion_matrix
>>> print(confusion_matrix(y_test, y_pred)))
```

Regression Metrics:

Mean Absolute Error:

```
>>> from sklearn.metrics import mean_absolute_error
>>> y_true = [3, -0.5, 2])
>>> mean_absolute_error(y_true, y_pred))
```

Mean Squared Error:

```
>>> from sklearn.metrics import mean_squared_error
>>> mean_squared_error(y_test, y_pred))
```

Tensorflow: TensorFlow is an open source library for fast numerical computing.It was created and is maintained by Google and released under the Apache 2.0 open source license. The API is nominally for the Python programming language, although there is access to the underlying C++ API. Unlike other numerical libraries intended for use in Deep Learning like Theano, TensorFlow was designed for use both in research and development and in production systems, not least RankBrain in Google search and the fun DeepDream project. It can run on single CPU systems, GPUs as well as mobile devices and large scale distributed systems of hundreds of machines.

Conclusion: The internship exercise was mainly to enable me acquire practical skills and link theory to practice. I have been able to acquire practical skills like creating dataset, working on models, algorithms like linear regression, logistic etc .

ANNEXURE A

CODE:

In [1]:
```
import tweepy
consumer_key="hmL3tbCl4Y7JgpKsDHKCpTEm9"
consumer_secret="zGJXewHINvlNE7ViTA10YuGsjZcOJamA1dgdF3V2aN0uF54zc2"
access_token="741301816361635840-p46wy3cSUEfuoozXvYNDvrfhpQjjt07"
access_token_secret="Edvzh6AN9KZdZtIClohE1K3KWPMlzY8u3rRRRxKJn8ztW"
```

In [2]:
```
auth=tweepy.OAuthHandler(consumer_key,consumer_secret)
auth.set_access_token(access_token,access_token_secret)
```

In [3]:
```
api=tweepy.API(auth)
user=api.user_timeline(screen_name="KushKnows")
```

In [4]:
```
tmpp=[]
tweets_for_csv =[tweet.text for tweet in user] # CSV file created
for j in tweets_for_csv:
        tmpp.append(j)
```

In [5]:
```
import numpy as nm
# Cleaning Spaces from each and every tweets
for  i in range(len(tmpp)):
     tmpp[i]=tmpp[i].replace('\n','')
```

In [6]:
```
import re

for i in range(len(tmpp)):
    x=re.compile(r'https://t.co/*[a-zA-Z0-9]*',re.DOTALL)
    tmpp[i]=re.sub(x,'',tmpp[i])
    print(tmpp[i])
```

```
RT @TripathiiPankaj: प्रतिभाओं(Talent) के चार चरण ।१ - अनुपयोग२ - दुरुपयोग३ - उपयोग४ - सदुपयोग अर्थात ईमानदारी और प्रेम से कार्यरत र…
RT @PMNehru: Only if I could ask Dhoni how it feels when everyone puts the blame on you. #NZvsIND
RT @TripathiiPankaj: .....
RT @varungrover: Jab 2003 sah liya toh 2019 kya bigaad lega humaara. Phir kabhi. #CWC19
Everywhere on social media, we are watching only 'Jai Shree Ram' chant videos by 'akal ke andhe Bhakts' , this screwed up me !!
RT @varungrover: नक्शे में जंगल है पेड़ नहींनक्शे में नदियां हैं पानी नहींनक्शे में पहाड़ है पत्थर नहींनक्शे में देश है लोग नहींसम…
RT @nadeemkhanUAH: Smile is saying all court order of prashant kanojia
@katelynnacon Happy birthday!!
गुड़िया खेलने की उम्र थी उसकी ,पर दरिन्दों ने उसकी ज़िन्दगी के साथ खेला !#TwinkleSharma
RT @anuragkashyap72: Chernobyl scripts .. here
RT @GabbbarSingh: From निंदा to डंडा 😂
Election Commission of India seems to be Hacked !!!
@GameOfThrones
Now again, a water bottle had been seen in GOT final episode.
RT @AisiTaisiDemo: Kaafi cool artwork featuring Aisi Taisi Democracy with a Cow-Aadhaar being made, by Kolkata based artists @M
Bobbying and…
RT @AAPPunjab: AAP Convenor @ArvindKejriwal paying regards and taking blessings of the legendary Maharaj Agrasen in Sunam.#Kejr
iwalinPunja…
Just promises remains with people ; no waste to energy plant in Ghazipur, Delhi. Note:These types of issue should n…
RT @Showbiz_IT: Game of Thrones ending will leave fans in worst pain possible, says Sophie Turner  …
RT @AamAadmiParty: .@raghav_chadha briefing the media about large scale bogus voting at a school in Tughlakabad.
Guruji : Hindu Muslim Sab Atapi Vatapi hai .
```

In [7]:
```
for i in range(len(tmpp)):
    tmpp[i]=re.sub(r'@[a-zA-Z_.0-9]*| #[a-zA-Z0-9]*','',tmpp[i])
    print(tmpp[i])
```

```
RT : प्रतिभाओं(Talent) के चार चरण ।१ - अनुपयोग२ - दुरुपयोग३ - उपयोग४ - सदुपयोग अर्थात ईमानदारी और प्रेम से कार्यरत र…
RT : Only if I could ask Dhoni how it feels when everyone puts the blame on you.
RT : .....
RT : Jab 2003 sah liya toh 2019 kya bigaad lega humaara. Phir kabhi.
Everywhere on social media, we are watching only 'Jai Shree Ram' chant videos by 'akal ke andhe Bhakts' , this screwed up me !!
RT : नक्शे में जंगल है पेड़ नहींनक्शे में नदियां हैं पानी नहींनक्शे में पहाड़ है पत्थर नहींनक्शे में देश है लोग नहींसम…
RT : Smile is saying all court order of prashant kanojia
 Happy birthday!!
गुड़िया खेलने की उम्र थी उसकी ,पर दरिन्दों ने उसकी ज़िन्दगी के साथ खेला !#TwinkleSharma
RT : Chernobyl scripts .. here
RT : From निंदा to डंडा 😂
Election Commission of India seems to be Hacked !!!


Now again, a water bottle had been seen in GOT final episode.
RT : Kaafi cool artwork featuring Aisi Taisi Democracy with a Cow-Aadhaar being made, by Kolkata based artists  and…
RT : AAP Convenor  paying regards and taking blessings of the legendary Maharaj Agrasen in Sunam.#KejriwalinPunja…
Just promises remains with people ; no waste to energy plant in Ghazipur, Delhi. Note:These types of issue should n…
RT : Game of Thrones ending will leave fans in worst pain possible, says Sophie Turner  …
RT : . briefing the media about large scale bogus voting at a school in Tughlakabad.
Guruji : Hindu Muslim Sab Atapi Vatapi hai .
```

In [8]:
```
from textblob import TextBlob
```

In [9]:
```
po=[]
for i in range(len(tmpp)):
    analysis=TextBlob(tmpp[i])
    po.append(analysis.sentiment.polarity)
    print(po[i],tmpp[i])
```

```
0.0 RT : प्रतिभाओं(Talent) के चार चरण ।१ - अनुपयोग२ - दुरुपयोग३ - उपयोग४ - सदुपयोग अर्थात ईमानदारी और प्रेम से कार्यरत र…
0.0 RT : Only if I could ask Dhoni how it feels when everyone puts the blame on you.
0.0 RT : .....
0.0 RT : Jab 2003 sah liya toh 2019 kya bigaad lega humaara. Phir kabhi.
0.016666666666666666 Everywhere on social media, we are watching only 'Jai Shree Ram' chant videos by 'akal ke andhe Bhakts' ,
this screwed up me !!
0.0 RT : नक्शे में जंगल है पेड़ नहींनक्शे में नदियां हैं पानी नहींनक्शे में पहाड़ है पत्थर नहींनक्शे में देश है लोग नहींसम…
0.3 RT : Smile is saying all court order of prashant kanojia
1.0  Happy birthday!!
0.0 गुड़िया खेलने की उम्र थी उसकी ,पर दरिन्दों ने उसकी ज़िन्दगी के साथ खेला !#TwinkleSharma
0.0 RT : Chernobyl scripts .. here
0.0 RT : From निंदा to डंडा 😂
0.0 Election Commission of India seems to be Hacked !!!
0.0
0.0 Now again, a water bottle had been seen in GOT final episode.
0.35 RT : Kaafi cool artwork featuring Aisi Taisi Democracy with a Cow-Aadhaar being made, by Kolkata based artists  and…
1.0 RT : AAP Convenor  paying regards and taking blessings of the legendary Maharaj Agrasen in Sunam.#KejriwalinPunja…
0.1 Just promises remains with people ; no waste to energy plant in Ghazipur, Delhi. Note:These types of issue should n…
-0.4666666666666666 RT : Game of Thrones ending will leave fans in worst pain possible, says Sophie Turner  …
0.21428571428571427 RT : . briefing the media about large scale bogus voting at a school in Tughlakabad.
0.0 Guruji : Hindu Muslim Sab Atapi Vatapi hai .
```

In [10]:
```
import pandas as pd
sentiment=pd.DataFrame(po,columns=["Polarity"])
```

In [11]:
```
senti=pd.DataFrame({'pol':po,'Tweets':tmpp},columns=["pol","Tweets"])
senti.head()
```

Out[11]:

	pol	Tweets
0	0.000000	RT : प्रतिभाओं(Talent) के चार चरण ।१ - अनुपयोग...
1	0.000000	RT : Only if I could ask Dhoni how it feels wh...
2	0.000000	RT : .....
3	0.000000	RT : Jab 2003 sah liya toh 2019 kya bigaad leg...
4	0.016667	Everywhere on social media, we are watching on...

In [12]:
```
import matplotlib.pyplot as plt
import seaborn as sns
x_axis=[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20]
plt.plot(x_axis, senti.pol)
plt.show()
```