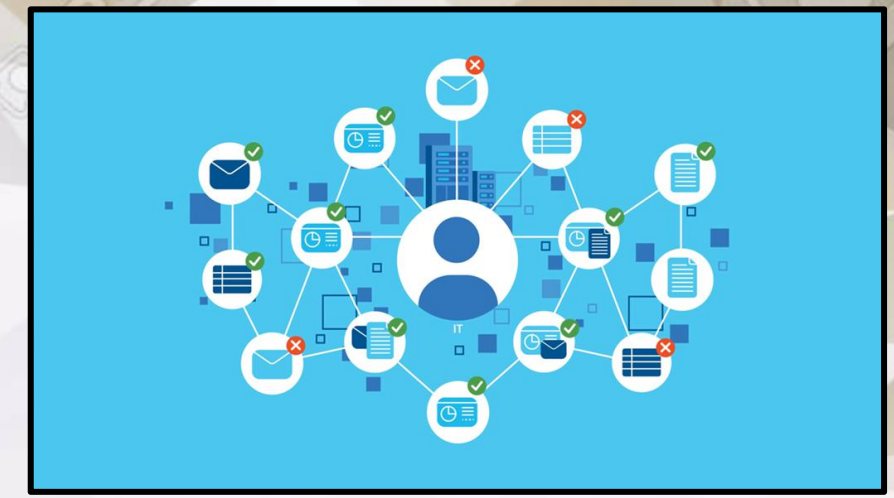
A Neural Network and Machine Learning Presentation on

Fake News Classifier

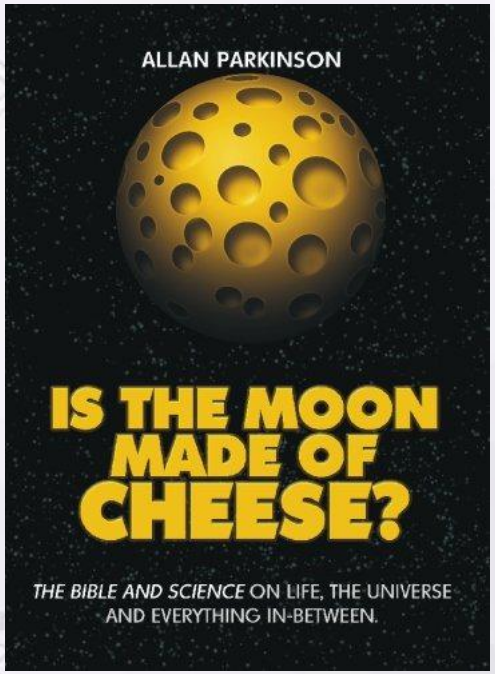
Using Natural Language Processing(NLP)

Overview

- In todays world. Everything that our eyes perceive and our mind interprets is some kind of information. Information or News can be acquired from many sources, especially in our era the trail of information is unending.



- But how does one classify that the information you're perceiving is legit or fake. You can hear a lot of things on the internet. When we were kids, we might have believed the fact that the moon is made up of cheese.



- So in this project we have decided to create a system that can easily select you're preferred article from a particular database and provide you it's authentication.

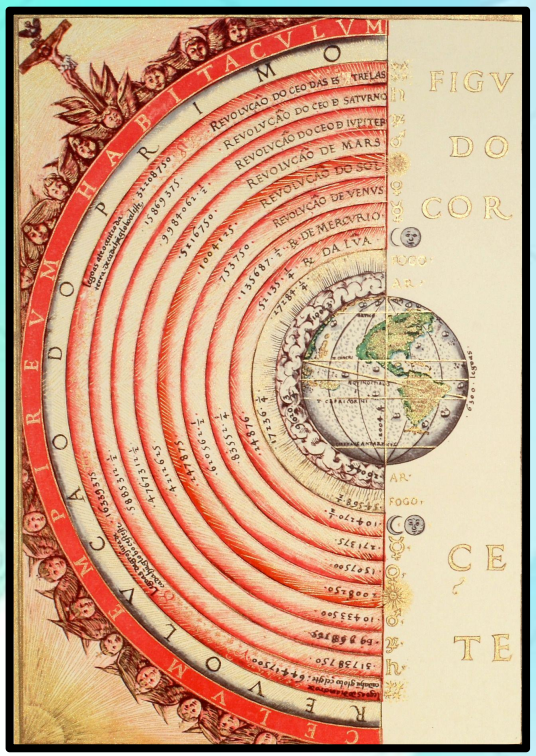
Fake News

- Before diving into the technical aspects of this presentation, lets take a brief moment to understand what exactly is a Fake News and what is its origin.

- Fake news is not a term that gained leverage with the invention of internet. Well there are multiple fake news sources nowadays but its origin is way old than the internet itself. Take a look at the last image at the right hand side. What kind of information do you perceive?

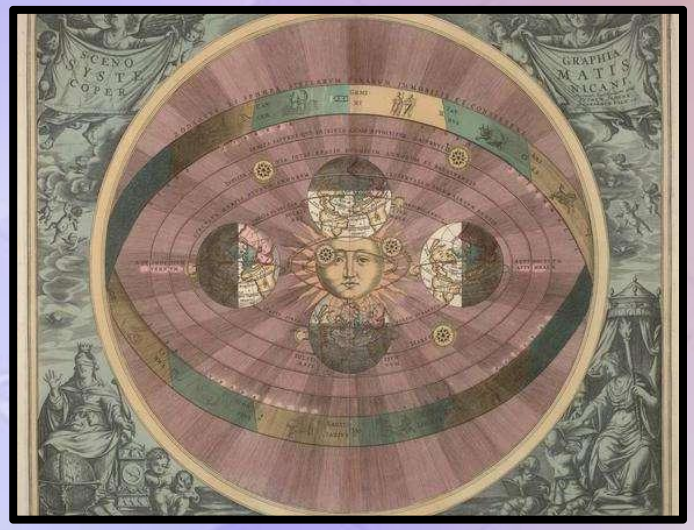
- In the next slide, we will explain a very famous debate reagrding the solar system that was termed the biggest hoax of the 2nd Century.

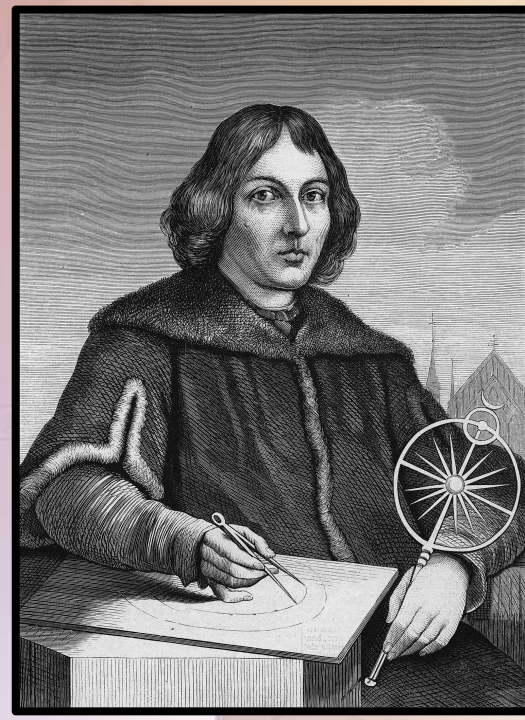
Geocentric Model (Fake)



The Geocentric Model was founded by the Ptolemy of Alexandria in the late 2nd Century. According to his model, the earth was the center of the solar system and all the other planets including the sun revolved around it. It was quite obvious to belive because people would see the sun and the moon moving and the earth staying still so the news was deemed true and also accepted by the Catholic Astronomers.

Heliocentric Model (Legit)





The Heliocentric model was proposed by Copernicus after 1400 years of the Geocentric model dominance.This model stated the sun as the center of the solar system and all the other planets as its children. He was opposed by the Catholic Astronomers and some even tried to attack him during his research. The heliocentric model turned out to be the true astronomical model and is still the model on which scientist research till now.

Database

- For this project we referred to kaggle.com for the database on fake nfews. It can be accesed using the link:
  https://www.kaggle.com/c/fake-news

- The database contains approximately 20,000 articles labeled 0's and 1's where 0 means True and 1 means False.

- The only limitation of this project is that it is limited to the dataset provided to it and it can only authenticate those articles and it cannot predict the authenticity of a foreign article.

Importing Libraries

1.Numpy

- *Essential when dealing with multidemensional arrays.*
- *Contains High level computational functions.*

2.Pandas

- *Pandas is a very important library when trying to manipulate or analyze datasets.*
- *In this project, we have mostly used pandas for our data pre processing works.*

3. RE

- *Regular Expression or RE is used for specifyinhg a set of strings that match.*
- *It is used to form a search pattern and can be used for stemming or converting words.*

4. NLTK

- *NLTK comes under NLP techniqe and is the most used feature for computational linguistics.*
- *NLTK also has more essential sub-libraries that will be explained as we move on with the slides.*

5. Sci-Kit

- *Sci-Kit learn is another important python library used for statistical modelling of the dataset and has various classification techniques such as Regression and SVM that are helpful in calculating the accuracy score of a model.*

6. Logistic Regression

- *Logistic Regression is a statistical model that is used mainly for binary classification even though it has more complex extensions. It is more helpful in calculating the accuracy score of models with binary output, such as ours.*

Note

All these Libraries have their own sub libraries that are used for in-depth processing and data cleaning. These sub libraries will also be explained in the upcoming slides.

Data Pre-Processing

- Data Pre-Processing is much needed when we're trying to get a good accuracy score for our model.

- The first step towards data pre-processing is to eliminate all the NaN values or empty cells inside the dataset.To find the NaN values inside our dataset, we used isnull command and found out approx 2500 NaN values. Since some of the cells were empty, the model might have assumed them as 0 and could have provided us with a wrong or a very low accuracy score.

- So, to counter this issue, we decided to use a command inside Pandas Library known as isnull and fillna for finding the null values and filling them up with blank spaces. With this step, the entire dataset could be fed to the machine for proper analysis.

```
#Finding the null values in the dataset
dataset.isnull().sum()
```

```
id          0
title     558
author   1957
text       39
label       0
dtype: int64
```

```
#substituting the null values with blank spaces
dataset=dataset.fillna('')
```

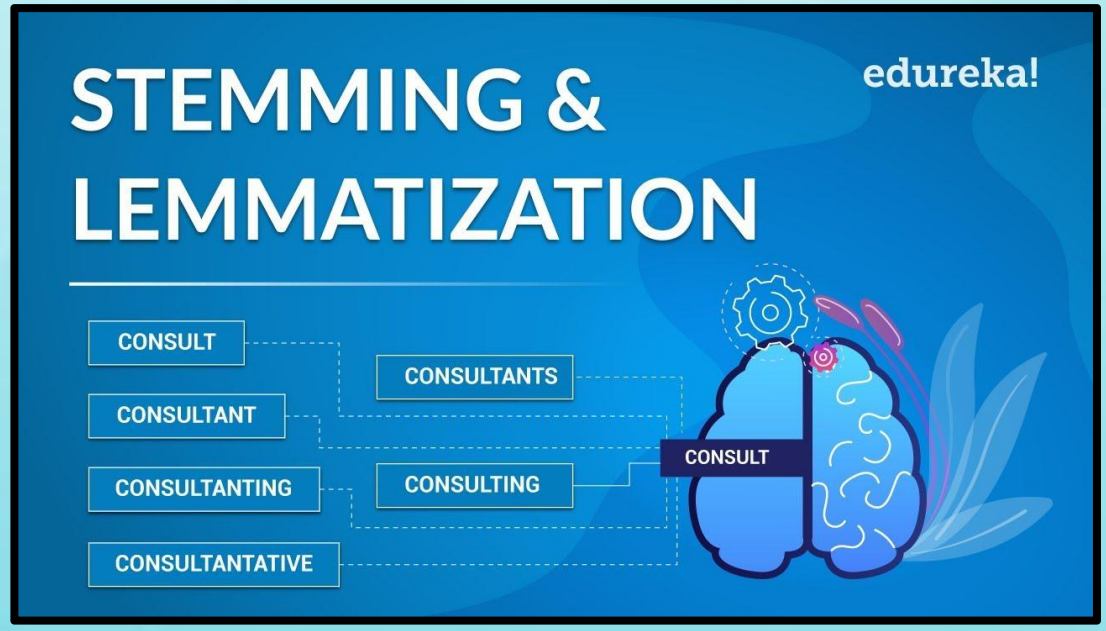
Left side shows the features/colums and the right side shows the number of NaN values inside that feature

NLTK Corpus Features

Stopwords: Inside our dataset, there are a lot of sentences. The use of Stopwords basically is to remove words that have no importance but to construct a sentence. *Words such as "the", "is", "can", etc are termed as Stopwords in NLTK* and it is used for making the model precise and fast in calculating the accuracy score.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Stemmer: *Stemming is a term that converts any word to its root origin.* For example, Flying and Flew can be converted to 'Fly', Runing and Ran can be converted to 'Run'. This also makes the model very effective in calculating the accuracy score.



Re.sub: *Using this command, we can substitute everything inside our dataset* with blank spaces (according to our code). This is mostly used for eliminating all the symbols and numbers from the dataset since our dataset is of Corpus format.

The need for Vectorization

- As we all know, *machines can only understand numbers*. So, as a result, they may have trouble understanding a database that is entirely based on text format.

- This is where we decided to use another Sci-Kit learn feature by the name of ***TFIDF(Term Frequency-Inverse Document Frequency) Vectorizer.***

- The main purpose of this Vectorizer is to convert each word to numbers based on the frequency of their occurence.

```
[ ] vectorizer=TfidfVectorizer()
    vectorizer.fit(X)

    X=vectorizer.transform(X)

    print(X)

    (0, 15686)    0.28485063562728646
    (0, 13473)    0.2565896679337957
    (0, 8909)     0.3635963806326075
    (0, 8630)     0.29212514087043684
    (0, 7692)     0.24785219520671603
    (0, 7005)     0.21874169089359144
    (0, 4973)     0.233316966909351
    (0, 3792)     0.2705332480845492
    (0, 3600)     0.3598939188262559
    (0, 2959)     0.2468450128533713
    (0, 2483)     0.3676519686797209
    (0, 267)      0.27010124977708766
    (1, 16799)    0.30071745655510157
    (1, 6816)     0.1904660198296849
    (1, 5503)     0.7143299355715573
    (1, 3568)     0.26373768806048464
    (1, 2813)     0.19094574062359204
    (1, 2223)     0.3827320386859759
    (1, 1894)     0.15521974226349364
    (1, 1497)     0.2939891562094648
    (2, 15611)    0.41544962664721613
```

Calculation of Accuracy Score

- The most suitable method to get a high accuracy score was to apply ***Logistic Regression*** Model since data set was of ***binary type***, i.e, outputs were either True(0) or False(1)

- By applying this model, we were able to get a accuracy score of *98%* or *0.987* in the train and as well as the test models. This accuracy score is close to 1 or 100% so it can be considered quite precise.
- To get accuracy as 100% some more hyperparameter tuning might be needed which is not necessary since our predictive system can be built on this accuracy score.

```
In [26]:
# Accuracy score of training data
X_train_prediction=model.predict(X_train)
training_data_accuracy=accuracy_score(X_train_prediction, Y_train)

In [27]:
print("Accuracy score of training data: ",training_data_accuracy)

Accuracy score of training data:  0.9865985576923076

In [28]:
# Accuracy score of testing data
X_test_prediction=model.predict(X_test)
testing_data_accuracy=accuracy_score(X_test_prediction, Y_test)

In [29]:
print("Accuracy score of testing data: ",testing_data_accuracy)

Accuracy score of testing data:  0.9790865384615385
```

Designing a Predictive System

The function of our predictive model is to request the article number the user wants to verify and then based on the article number provided, the model gives the authenticity as the output.Please keep in mind that the model only works for the articles inside the dataset and won't recognize any foreign subject.

Designing a predictive system

In [30]:
```
News=X_test[5]
prediction=model.predict(News)
print(prediction)

if (prediction==0):
  print("The news is Real.")
else:
  print("The news is Fake.")
```

```
[1]
The news is Fake.
```

In [31]:
```
print(Y_test[5])
```

```
1
```

Conclusion

With this slide, our presentation comes to an end. Below you will find the reference websites from where we gained insights about our projects if you wish to explore this topic for yourselves.

Thank you very much for your patience.

Reference

1. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
2. https://www.hindawi.com/journals/complexity/2020/8885861/
3. https://towardsdatascience.com/deep-learning-techniques-for-text-classification-78d9dc40bf7c

Group Details

Group 9	
Anshul Kumar	1907010
Saurabh Singh	1907042
Shivam Singh	1907046
Yash Gautam	1907060