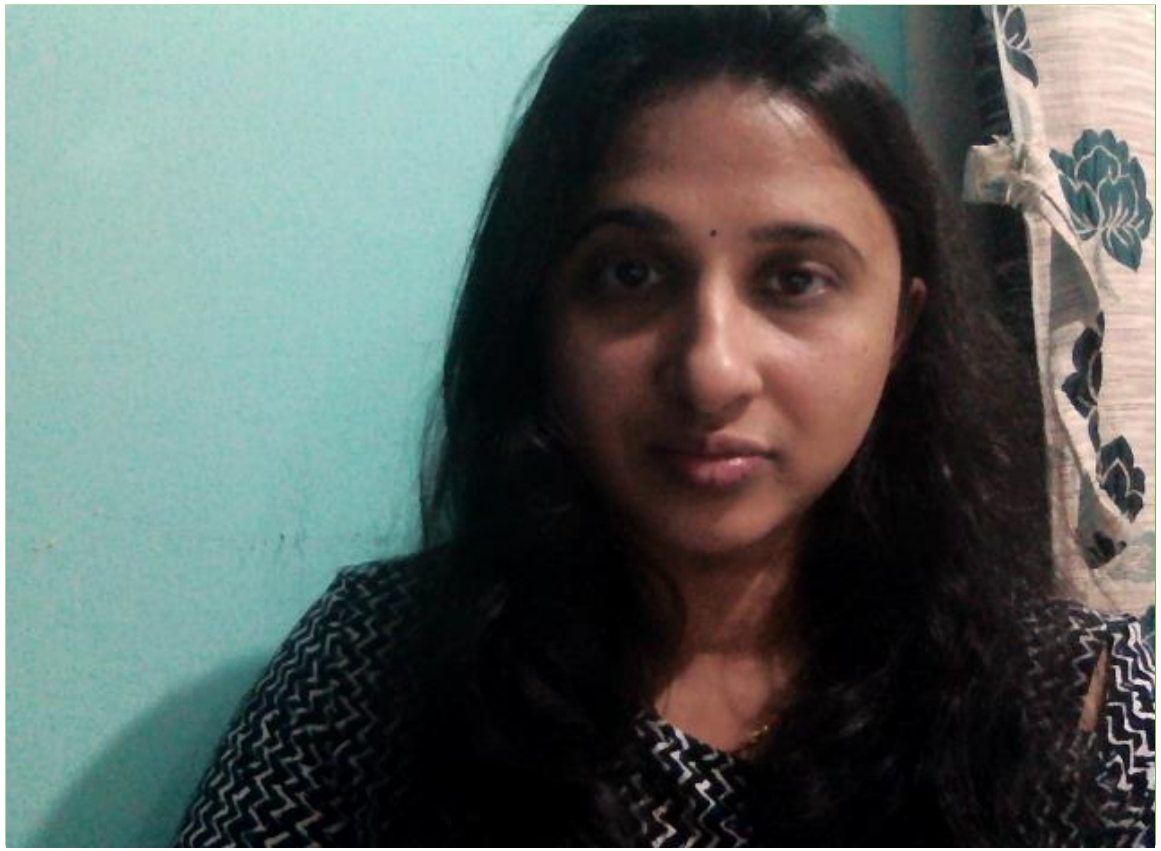
Introduction (max 30 sec)



Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

Machine Learning on Text Data

Ramya Mercy Rajan

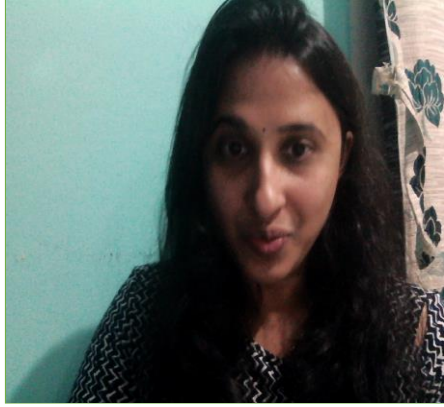
Roll No AA.SC.P2MCA2107434

What have you learned in NLP?

- **Natural Language Processing (NLP)** is a branch of Artificial Intelligence (AI) that enables machines to understand the human language.
- NLP analyzes the grammatical structure of sentences and the individual meaning of words, then uses algorithms to extract meaning and deliver outputs.
- NLP primarily comprises of natural language understanding (human to machine) and natural language generation (machine to human).
- 5 common techniques used in Natural Language Processing

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434



Natural Language Processing

1. ## Named Entity Recognition

   The most basic and useful technique in NLP is extracting the entities in the text. Named Entity Recognition identifies entities such as people, locations, organizations, dates, etc. from the text.

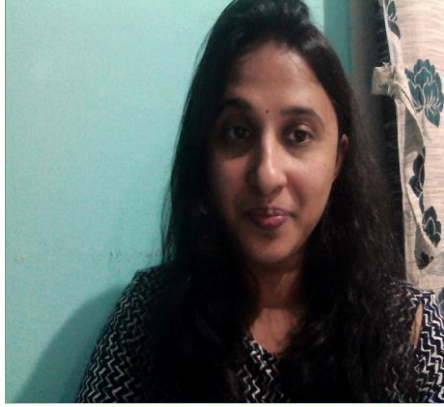
2. ## Sentiment Analysis

   Sentiment Analysis is most useful in cases such as customer surveys, reviews and social media comments where people express their opinions and feedback. The simplest output of sentiment analysis is a 3-point scale: positive/negative/neutral.

3. ## Text Summarization

   Techniques that help summarize large chunks of text. Text Summarization is mainly used in cases such as news articles and research articles.

Name : Ramya Mercy Rajan
Roll No : AA.SC.P2MCA2107434



4. Aspect Mining

Aspect mining identifies the different aspects in the text. One of the easiest methods of aspect mining is using part-of-speech tagging.

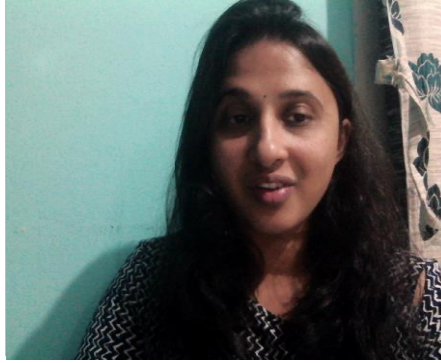
5. Topic Modeling

Topic modeling is one of the more complicated methods to identify natural topics in the text. A prime advantage of topic modeling is that it is an unsupervised technique. Model training and a labeled training dataset are not required.

There are quite a few algorithms for topic modelling, two of them are

- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)

Name : Ramya Mercy Rajan
Roll No : AA.SC.P2MCA2107434



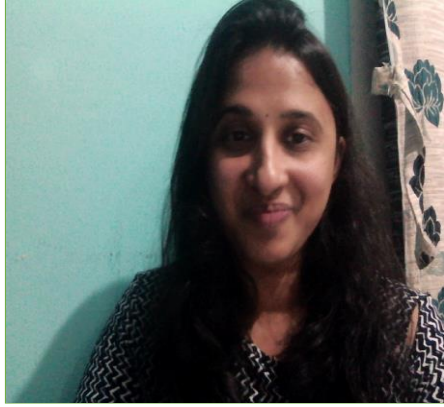
Project Description(1 min)

- This projects includes extraction of tweets from twitter and have performed pre-processing and text representation. Machine Learning Algorithms have been applied on it.
- Length of dataset is 13192
- Tweets dataset consist of 13192 rows × 3 columns
- Columns consist of id, text and label.
- Labels are those 5 classes (labels) selected and extracted tweets from twitter.
- The 5 classes includes

1. Digital electronics
2. Robotics
3. Artificial Intelligence
4. Computer Vision
5. Bioinformatics

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434



- The stopwords was cleaned and removed from the tweets text.
- URL's, punctuations, numeric numbers were removed as a part of Preprocessing stage.
- Tweets text were tokenized , stemming and Lemmatization was applied.
- Plotted a cloud of words for negative tweets.
- Data was splitted in to Train and Test subset
- Data was transformed using TF-IDF Vectorizer.
- Word2vec was the another technique/model used to produce word embedding. It is a natural language processing method that captures a large number of precise syntactic and semantic word relationships.

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

- I have used three different models respectively :

1. Bernoulli Naive Bayes
2. SVM (Support Vector Machine)
3. Logistic Regression

- The idea behind choosing these models is that I want to try all the classifiers on the dataset ranging from simple ones to complex models and then try to find out the one which gives the best performance among them.

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

Data Creation and Description

- Data Acquisition technique

➢Length of dataset is 13192

➢Tweets dataset consist of 13192 rows × 3 columns

➢Columns consist of id, text and label.

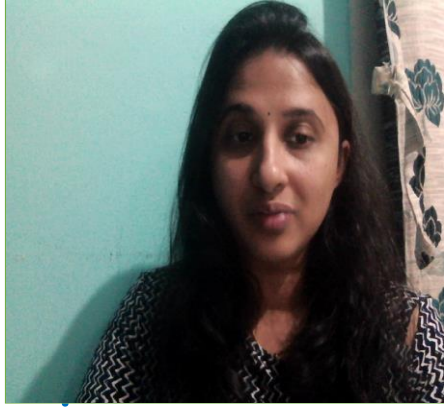
➢Labels are those 5 classes selected and extracted tweets from twitter.

➢The 5 classes includes

- ❖Digital electronics
- ❖Robotics
- ❖Artificial Intelligence
- ❖Computer Vision
- ❖Bioinformatics

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

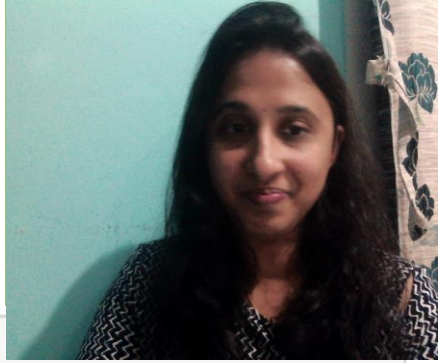


- Pre-processing applied - The stopwords, URL’s, punctuations, numeric numbers were removed as a part of Preprocessing stage.
- Data description
  - Number of records- 13192 rows × 3 columns
  - Number of tokens- 109923

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

Dataset creation - twitter data extraction



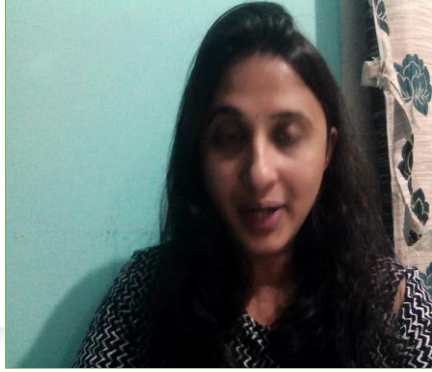
```
query_list=['digital electronics','robotics','artificial intelligence','computer vision','bioinformatics']
```

```
%%time
tweets_response_list = []
for q in query_list:
    for page in tweepy.Cursor(API.search_tweets, q=q + " lang:en -filter:retweets", count=100, tweet_mode='extended').pages(50):
        for response in page:
            tweets_response_list.append([response.id_str, response.full_text, q])
    df = pd.DataFrame(tweets_response_list, columns = ['id_str', 'text', 'label'])
    df.to_csv("tweets_ramya_new.csv", index=False)

df.to_csv("tweets_ramya_new.csv", index=False)
```

```
CPU times: total: 12.7 s
Wall time: 2min 30s
```


Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434



- Dataset

```
Wall time: 2min 30s
```

```
df
```

	id_str	text	label
0	1521120082444926976	GAOMON M10K PRO 10 x 6.25 Inches Art Digital G...	digital electronics
1	1521116610764832768	man i hate digital electronics so damn muchhh ...	digital electronics
2	1521113104804941827	https://t.co/dRPEBmWJBr Riptunes Portable Cass...	digital electronics
3	1521113064099172352	12. ISMC semiconductor\n\n📊 Detailed Stats: ht...	digital electronics
4	1521105470873739264	@AtteroIndia @Navyavegi 5 Electronics:\n\nLAPT...	digital electronics
...	...	...	...
13187	1518164863461335040	Our Bioinformatics Masters (MSc/MRes) courses ...	bioinformatics
13188	1518133273096081408	#BioIT #BioInformatics What does it mean by pr...	bioinformatics
13189	1518133265726685185	#BioIT #BioInformatics Answer: Biostar under s...	bioinformatics
13190	1518127369546092544	Postdoctoral Funded Position in Microbial Ecol...	bioinformatics
13191	1518124056792678400	When I read "bioinformatic analysis revealed" ...	bioinformatics

13192 rows × 3 columns

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

Vectorization on Dataset

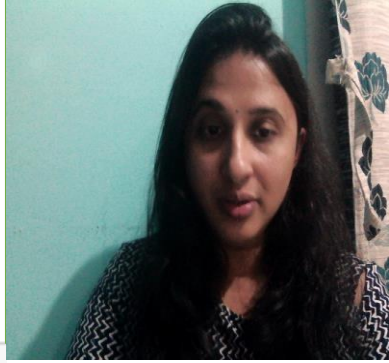
Screenshot and explanation of vectorization techniques used

- Term Frequency — **Inverse Document Frequency** (TFIDF) is a technique for text vectorization based on the Bag of words (BoW) model. It performs better than the BoW model as it considers the importance of the word in a document into consideration. The main limitation is that it does not capture the semantic meaning of the words.
- Word2vec **technique/model of word embedding was used**. It is a natural language processing method that captures a large number of precise syntactic and semantic word relationships.

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

Transforming Dataset using TF-IDF Vectorizer



Fit the TF-IDF Vectorizer

[45]:
```
%%time
vectoriser = TfidfVectorizer(ngram_range=(1,2), max_features=500000)
vectoriser.fit(X_train)
print('No. of feature_words: ', len(vectoriser.get_feature_names()))
```

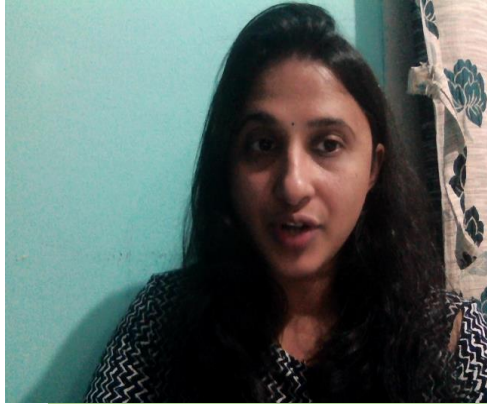
```
No. of feature_words:  109923
CPU times: total: 1.19 s
Wall time: 1.18 s
```

Transform the data using TF-IDF Vectorizer

[46]:
```
X_train = vectoriser.transform(X_train)
X_test  = vectoriser.transform(X_test)
```


Name : Ramya Mercy Rajan

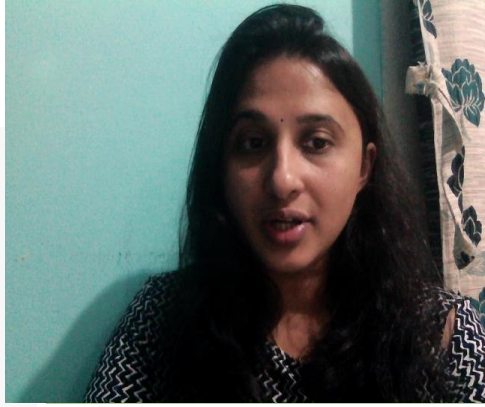
Roll No : AA.SC.P2MCA2107434



```python
class Word2VecVectorizer:
    def __init__(self, file_path):
        print("Loading in word vectors...")
        binary_file = file_path.endswith(".bin")
        self.word_vectors = KeyedVectors.load_word2vec_format(
            file_path, encoding="utf-8",
            binary=binary_file
        )
        print("Finished loading in word vectors")

    def fit(self, data):
        pass

    def transform(self, data):
        # determine the dimensionality of vectors
        v = self.word_vectors.get_vector('king')
        self.D = v.shape[0]

        X = np.zeros((len(data), self.D))
        n = 0
        emptycount = 0
        for sentence in data:
            tokens = sentence.split()
            vecs = []
            m = 0
            for word in tokens:
```


Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

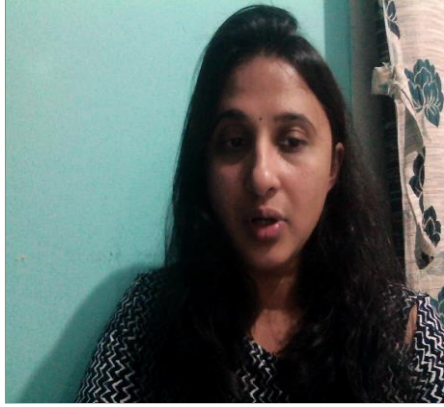


```
            try:
                # throws KeyError if word not found
                vec = self.word_vectors.get_vector(word)
                vecs.append(vec)
                m += 1
            except KeyError:
                pass
        if len(vecs) > 0:
            vecs = np.array(vecs)
            X[n] = vecs.mean(axis=0)
        else:
            emptycount += 1
        n += 1
    print("Number of samples with no words found: %s / %s" % (emptycount, len(data)))
    return X


def fit_transform(self, data):
    self.fit(data)
    return self.transform(data)
```


Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

ML on dataset
Train Test data used in ML



Splitting our data into Train and Test Subset

In [41]:
```
# Separating the 95% data for training data and 5% for testing data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, shuffle = True, stratify = y, random_state = 26105111)
```

In [42]:
```
X_train.shape, X_test.shape
```

Out[42]:
```
((9234,), (3958,))
```

In [43]:
```
y_train.shape, y_test.shape
```

Out[43]:
```
((9234,), (3958,))
```

In [44]:
```
X_train
```

Out[44]:
```
8334     artifici intellig classroom rachel dene poth p...
9491              govern announc am artifici intellig push
9777     rzenelzld giveawayhost project creat maximum r...
11021    nga s mark munsel deputi director data digit i...
4609     narvuntien mustbejosh hakimi hakimi wayneallan...
                               ...
5131     dji halt russia ukrain busi prevent drone misu...
8013     ebmhead ricfulop lorakolodni stevelevin know r...
8788     legitgraci unigridorg solida project project b...
9947     feder bank agenc tri ensur ai ml benefit rathe...
3827     appreci elonmusk inspir innov climat vision in...
Name: text_stem, Length: 9234, dtype: object
```


Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

- Screenshots of ML algorithms used
- Evaluation/Accuracy chart

```
In [72]: list_accuracy = [[Test_Accuracy_bnb1, Test_Accuracy_svc1, Test_Accuracy_lr1],
                          [Test_Accuracy_bnb2, Test_Accuracy_svc2, Test_Accuracy_lr2],
                          [Test_Accuracy_bnb3, Test_Accuracy_svc3, Test_Accuracy_lr3]]
         df_accuracy = pd.DataFrame(list_accuracy, columns = ['BernoulliNB', 'LinearSVC', 'LogisticRegression'])
         df_accuracy.index = ['TF-IDF', 'Word Embedding 1', 'Word Embedding 2']
         df_accuracy[df_accuracy.columns] = df_accuracy[df_accuracy.columns].applymap(lambda x: x*100)
         df_accuracy[df_accuracy.columns] = df_accuracy[df_accuracy.columns].applymap("{0:.2f}%".format)
         df_accuracy
```

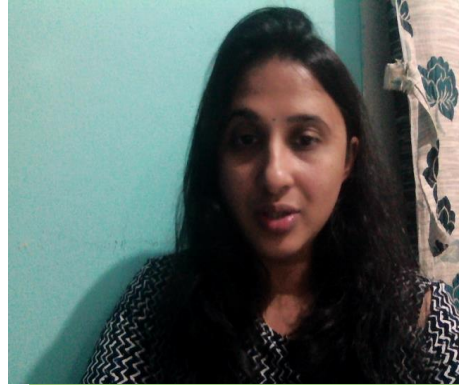
Out[72]:

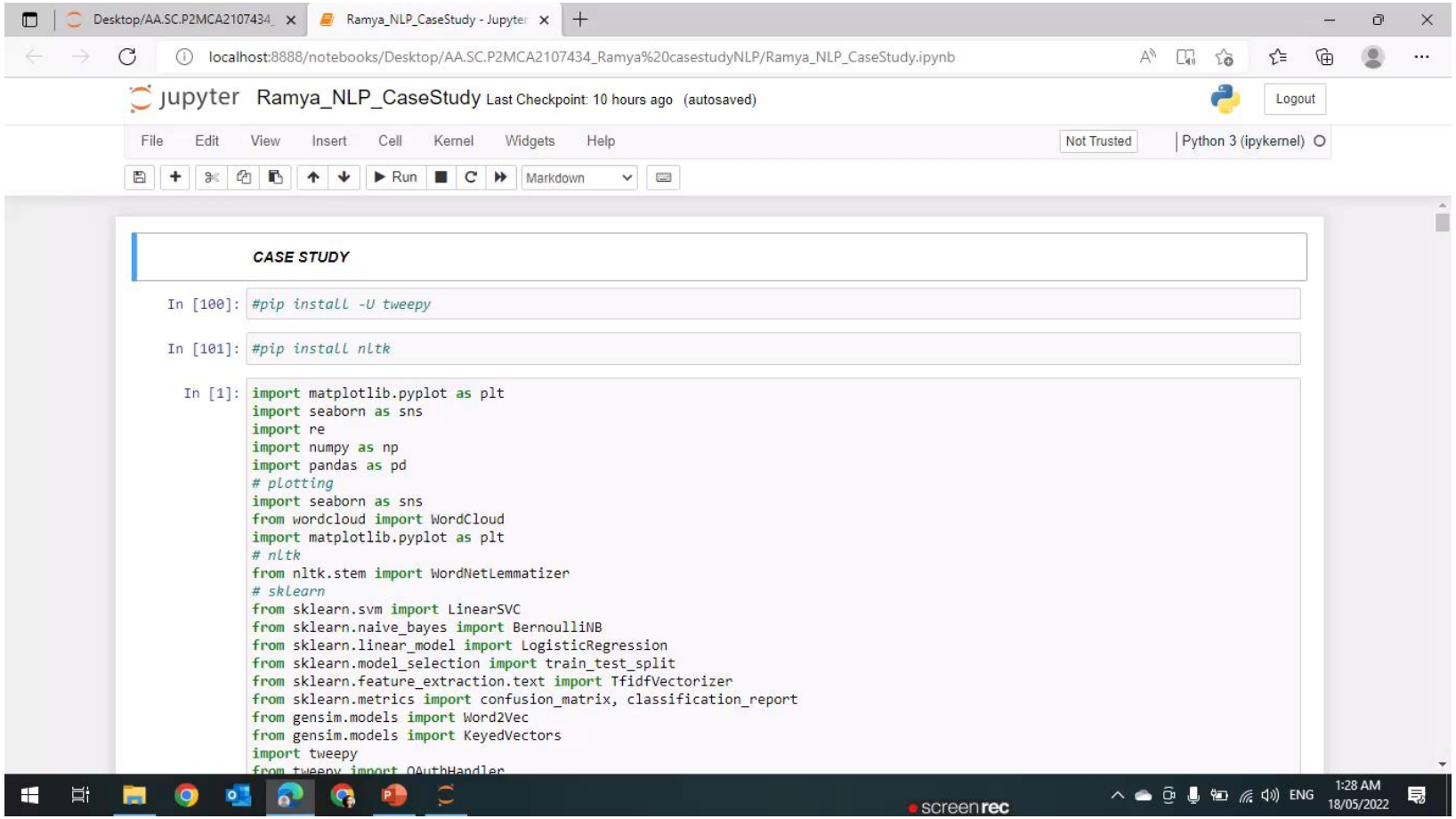
	BernoulliNB	LinearSVC	LogisticRegression
TF-IDF	73.29%	96.08%	95.60%
Word Embedding 1	89.69%	94.90%	94.44%
Word Embedding 2	88.45%	94.52%	94.39%

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

Demo of the project





Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434

Conclusion

Name : Ramya Mercy Rajan

Roll No : AA.SC.P2MCA2107434