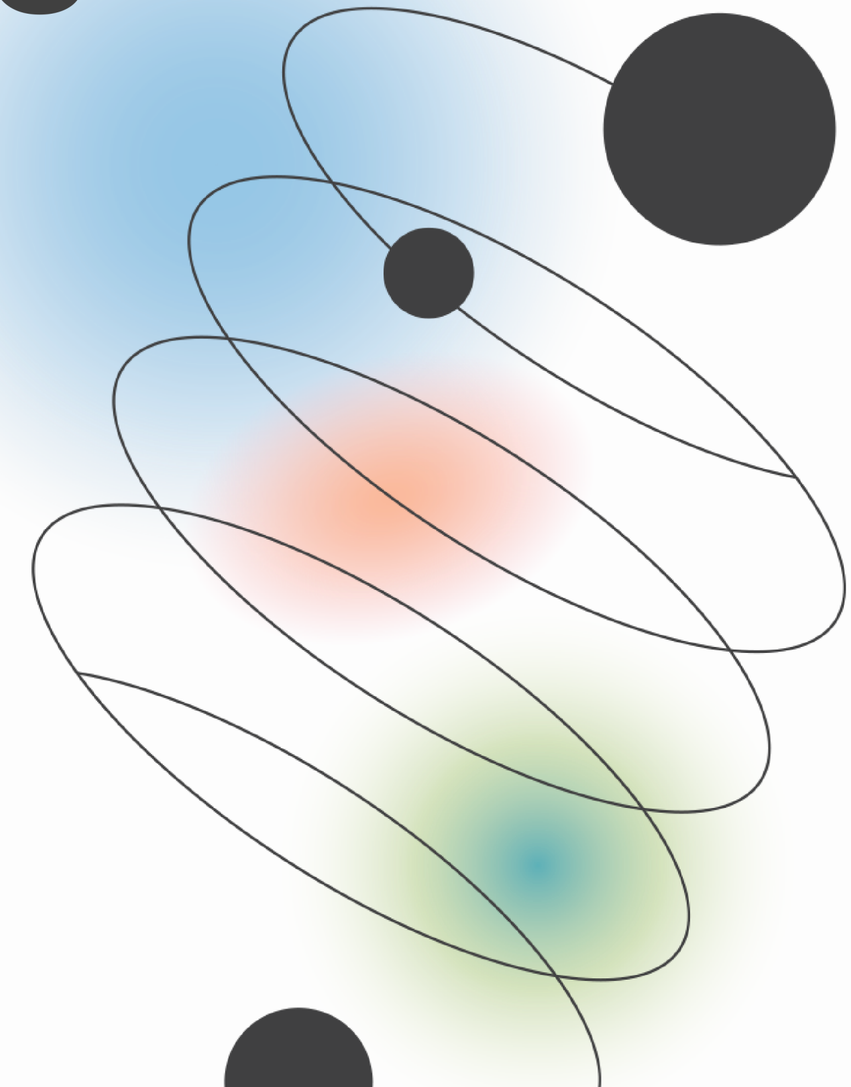
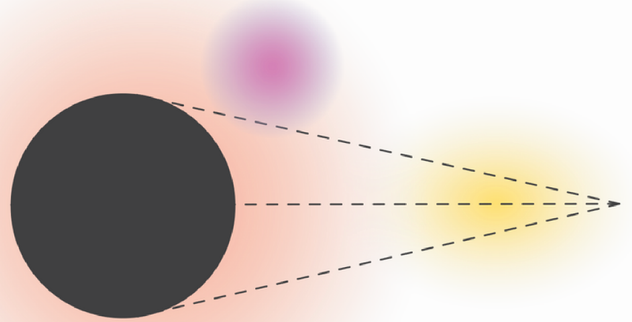
Democratizing NLP



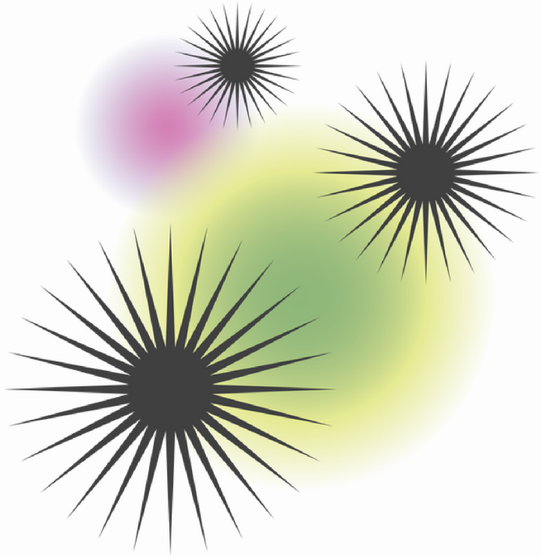
Progress Check Meeting

LITMUS7

Today's Agenda



In this meeting, we'll be going over the following:



01

Project Aim

02

Code Review - Colab

03

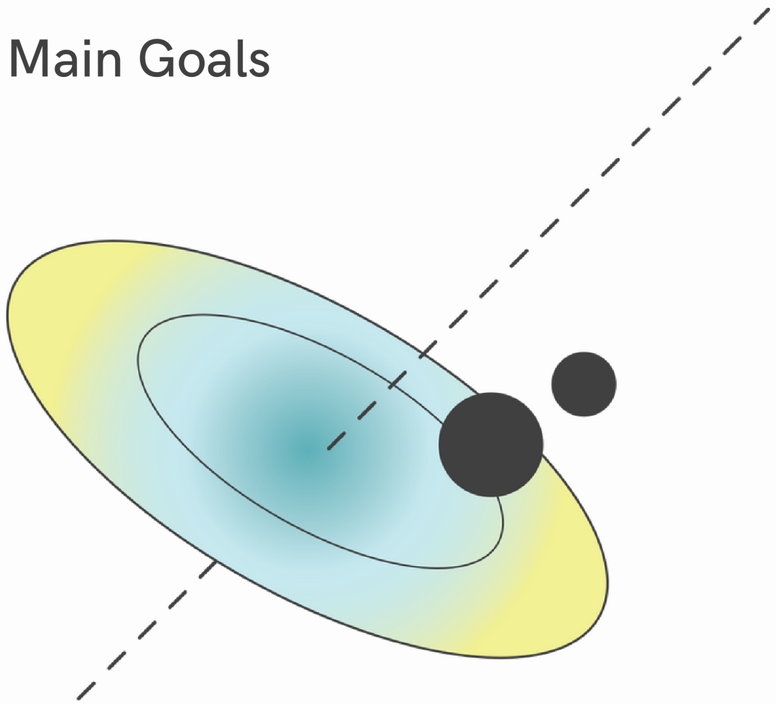
Code Review - Streamlit

04

What's Next

Project Aim

Main Goals



Implement a NLP Framework

Design an interface where a User can upload a Dataset and can select the Natural Language Processing (NLP) steps which can be performed on the Dataset and can download the processed file.

Implement a StreamLit UI

Design a Interface for the smooth functioning of the Program

Project Timeline

WE ARE HERE!

Importing the Dataset

Data Visualization

Data Pre-Processing

Vectorization Techniques

StreamLit UI

Project Completion

Dataset Upload Technique

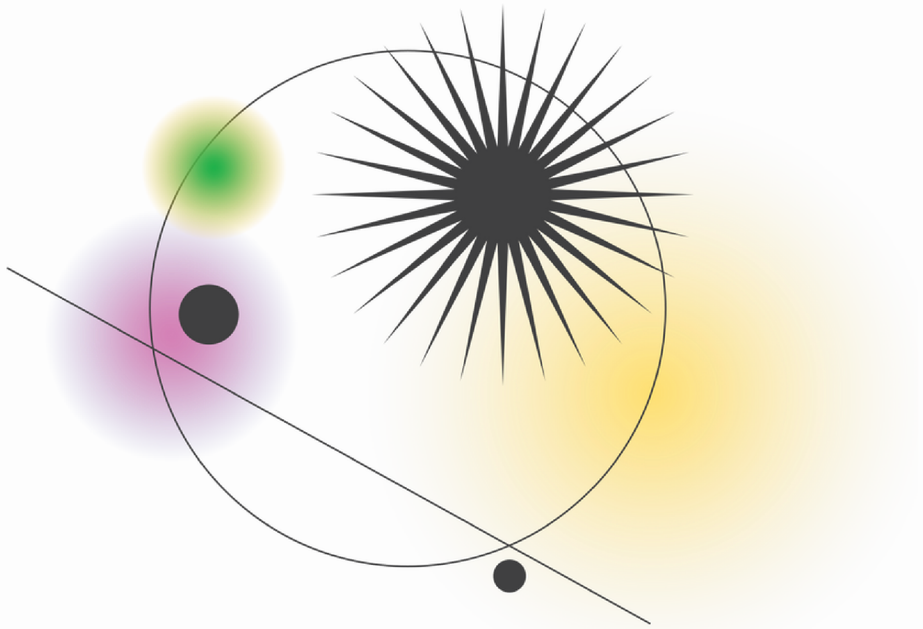
- One or more Datasets can be uploaded by the User. If more than One Dataset present, the User will be prompted to enter the Dataset he wants to process first.

- The Datasets can be of the formats : CSV, JSON or XML. Different functions are invoked for each format. The format will be automatically detected.

Our Current Progress

This is how we're doing so far this year.

Data Visualization



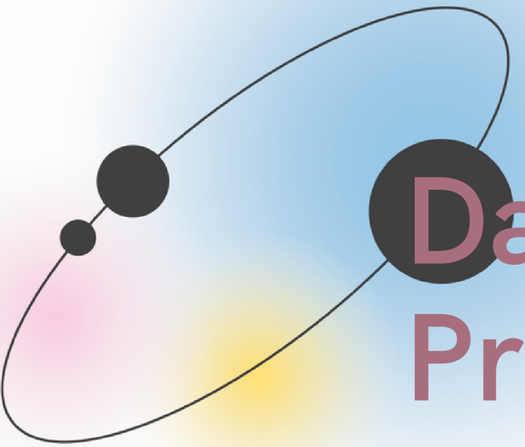
01 Null Values - Plotting Graphs and Removing Null Values

02 Missing Values - Removal

03 Check for Duplicate Values

04 Languages present in the Dataset

05 N Gram Visualization



Data Pre-Processing

01 Removing Null Values and Duplicate Values

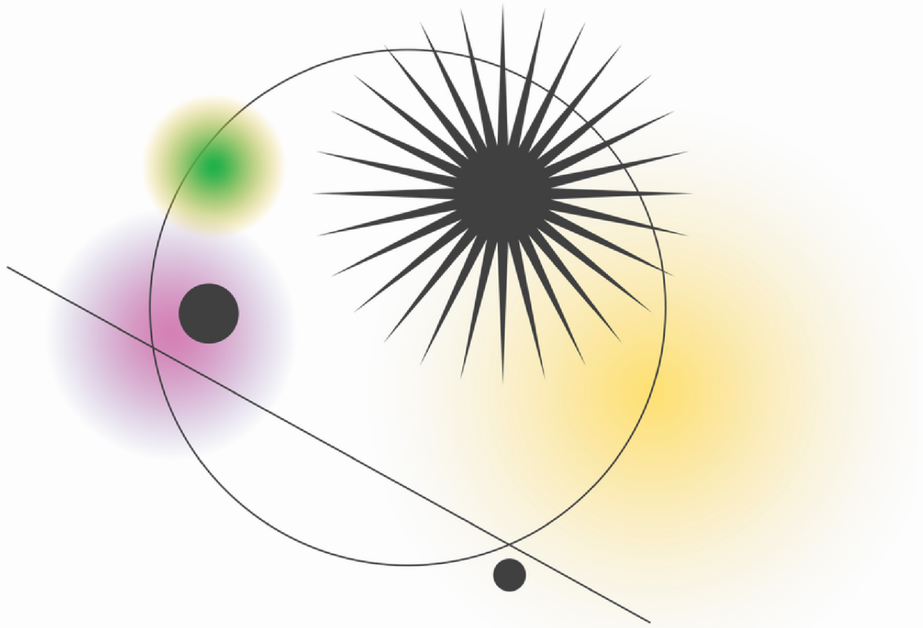
02 Tokenization, Removal of Stop Words, Punctuation

03 Stemming and Lemmatization, Processing based on Regex - User Input

04 Solving Encoding Issues with ftfy library

05 Language Detection for every line in Dataset

Pre-Processing Menu



01 Standard Pre-Processing

02 Regex for removal of Digits

03 Regex for removal of Alphabets

04 Regex for removal of Alphanumeric

05 Regex for retaining only alphanum

06 Regex as a User Input

07 Regex for Removing Special Characters

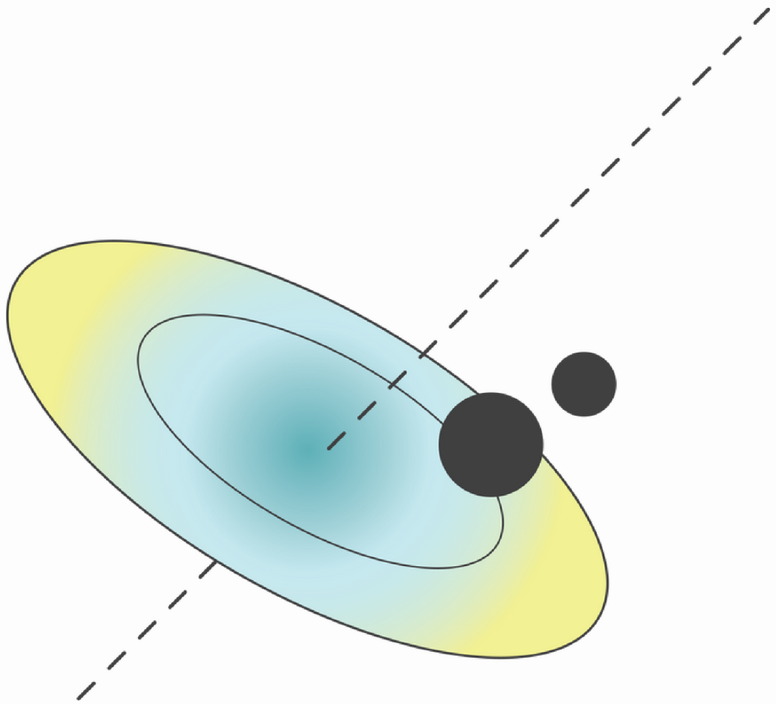
Fixing Encoding

Using ftfy Library

- The Encoding issues can be fixed using the Python3 fyfy library.
- In our code, we iterate through the Dataframe using for loop and invoke the ftfy function to fix the encoding issues

```
#!pip install ftfy
for i,r  in enumerate(df["text"]):
  try:
    df.loc[i,"text"] = ftfy.fix_text(r)
  except:
    continue
```

Language Detection



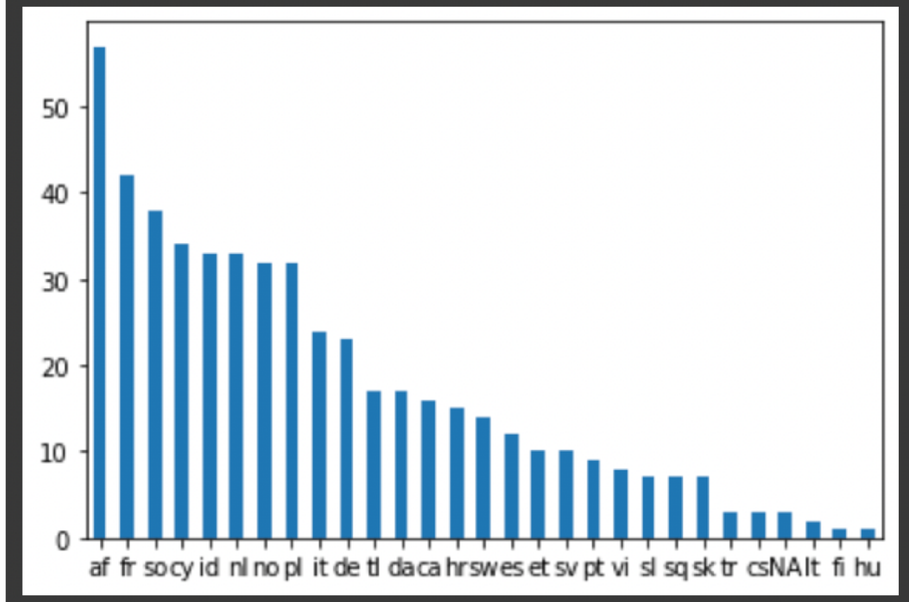
Using langdetect package

- In our code, we are creating a new column "language" for storing the language for each row.
- Therefore, we enumerate through the dataset and using the detect function in langdetect package, we find the language present in the dataset.
- The Translator package is used in the code, and the translate() is used for translation of the text

```
print("Do you wish to translate the text")
tr = input()
if(tr == "yes" or tr == "y"):
  fromlang = input("Enter the language of the given text")
  tolang = input("Enter the language to which the given text has to be translated")
  translator= Translator(to_lang=tolang)
  print(translator.translate("How do you do?"))
  for i,r  in enumerate(df["text"]):
    try:
        df.loc[i,"translation"] = translator.translate(r)
    except:
      df.loc[i,"translation"] = "Unable to Translate"
```

Languages Present in Dataset other than English

```
A = new_df["language"].value_counts()
A = A.drop(labels=["en"])
by = A.plot.bar(x='lab', y='val', rot=0)
#Parameters to be changed
```



- Using barplot, we showcase the different languages present
- English has been removed, as English forms the vast majority of the languages detected
- Therefore, to maintain the readability of the bar graph, English is removed

Initial View of the Dataframe

- Un-Processed Text
- Spam/Ham

	text	spam
0	Subject: naturally irresistible your corporate...	1.0
1	Subject: the stock trading gunslinger fanny i...	1.0
2	Subject: unbelievable new homes made easy im ...	1.0
3	Subject: 4 color printing special request add...	1.0
4	Subject: do not have money , get software cds ...	1.0
...	...	...
11301	This is the 2nd time we have tried 2 contact u...	1.0
11302	Will b going to esplanade fr home?	0.0
11303	Pity, * was in mood for that. So...any other s...	0.0
11304	The guy did some bitching but I acted like i'd...	0.0
11305	Rofl. Its true to its name	0.0

11306 rows × 2 columns

Final View of Dataframe

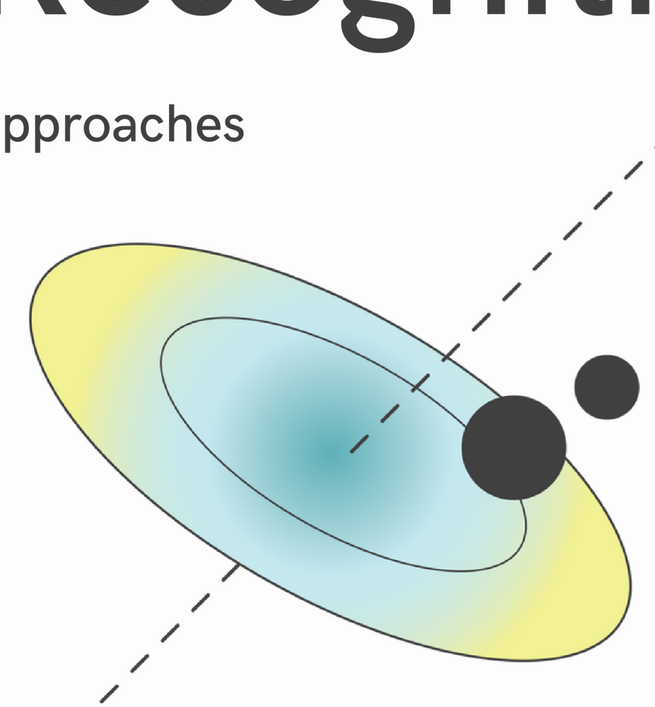
	text	spam	language	Pre-Processed
0	Subject: naturally irresistible your corporate...	1.0	en	subject natur irresist corpor ident lt realli ...
1	Subject: the stock trading gunslinger fanny i...	1.0	en	subject stock trade gunsling fanni merril muzo...
2	Subject: unbelievable new homes made easy im ...	1.0	en	subject unbeliev new home made easi im want sh...
3	Subject: color printing special request addi...	1.0	en	subject color print special request addit info...
4	Subject: do not have money , get software cds ...	1.0	en	subject money get softwar cd softwar compat gr...
...	...	...	...	...
10807	NaN	NaN	id	nan
10825	NaN	NaN	en	nan
10838	NaN	NaN	en	nan
10839	NaN	NaN	en	nan
10842	NaN	NaN	en	nan

11252 rows × 4 columns

- Un-Processed Text
- Spam/Ham
- Language of Data
- Pre-Processed Data

Named Entity Recognition

Approaches



We have advocated 2 approaches for NER:

- Transformers
- Spacy

- The line where NER has to be performed is taken as input from the user

Word Cloud Visualization

With Stopwords

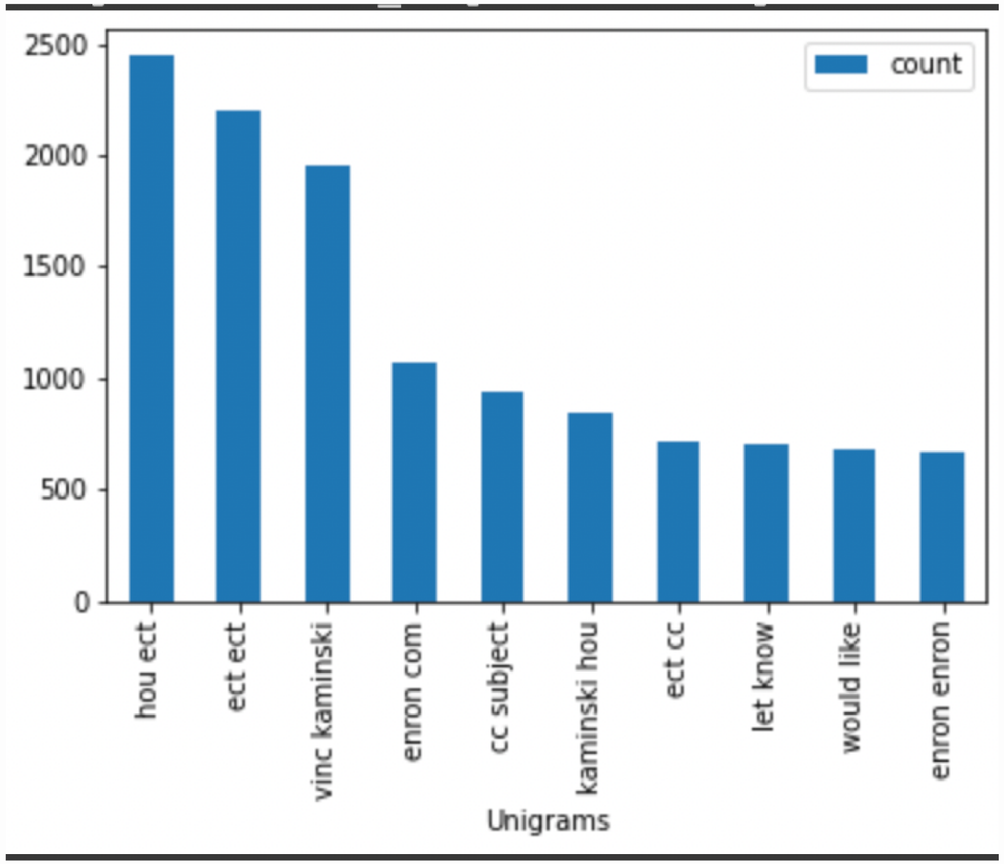
- The Word Cloud Visualization is done based on the Stopwords.

N Grams

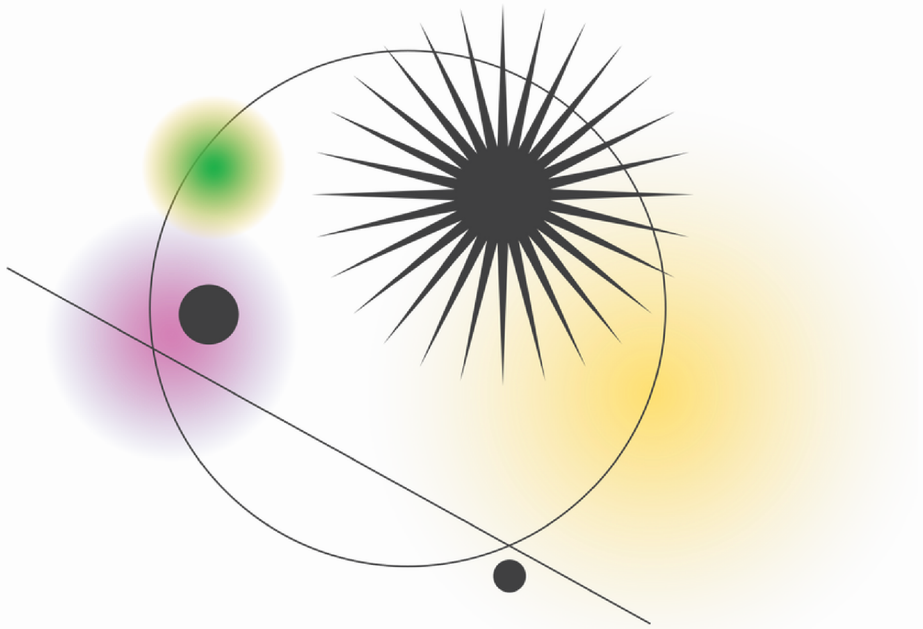
N Gram Analysis on the Dataset

```
top_n = 20
ngram_range = (1,1)
uni_grams = get_top_n_ngrams(new_df["Pre-Processed"], top_n, ngram_range)
unigram_df = pd.DataFrame(uni_grams, columns = ['Unigrams' , 'count'])
#plot top 20
```

	Unigrams	count
0	enron	6211
1	subject	5585
2	ect	5224
3	vinc	4000
4	com	3008



ML Algorithm



01 K Nearest Neighbour

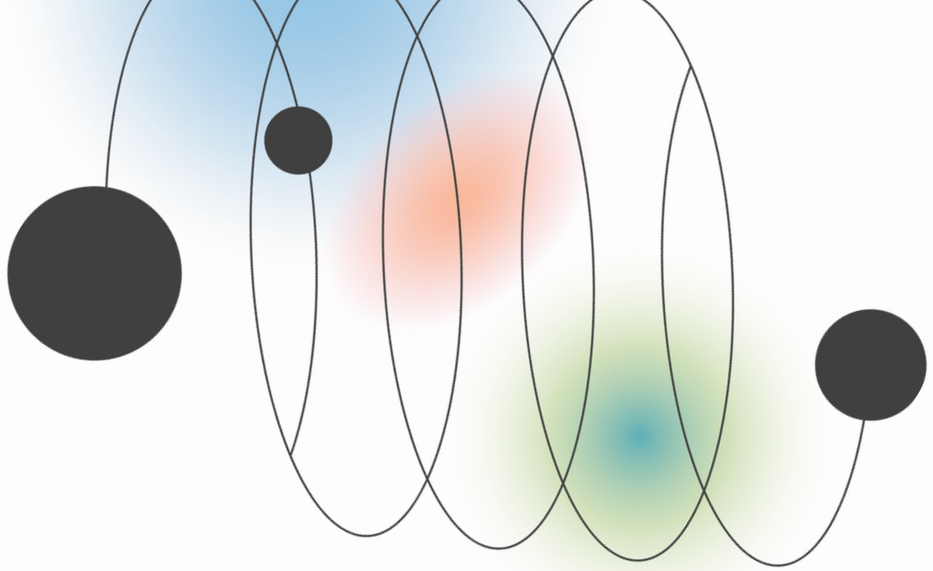
02 Logistic Regression

03 Support Vector Machine

04 Naive Bayes

05 Random Forest

Libraries Used



01

Plot

MatplotLib, MissingNo, Seaborn

02

Pre-Processing

re, ftfy,nltk, pandas, PorterStemmer

03

Language

LangDetect, Translator

04

Vectorization, NER

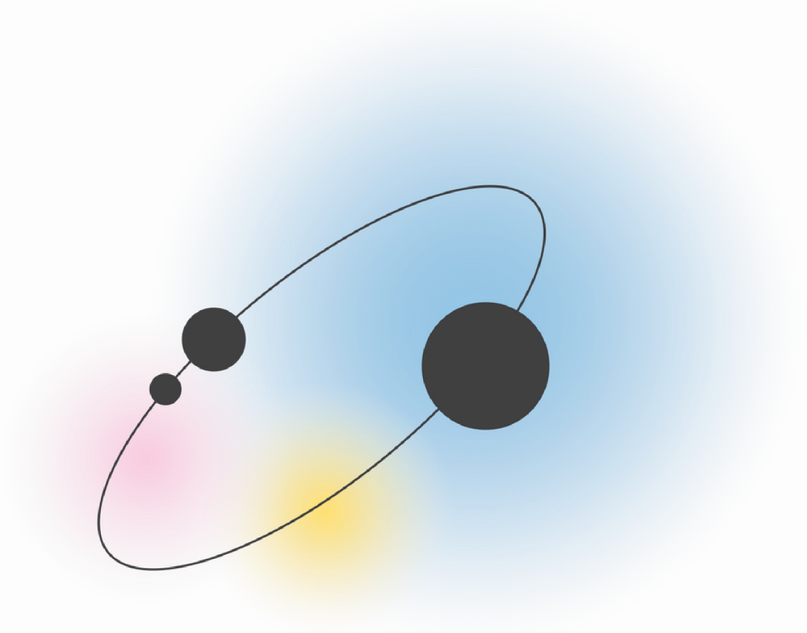
Word2Vec, CountVectorizer, WordCloud, spacy, transformers

05

ML Algo

sklearn

Downloading the Model



ML Model

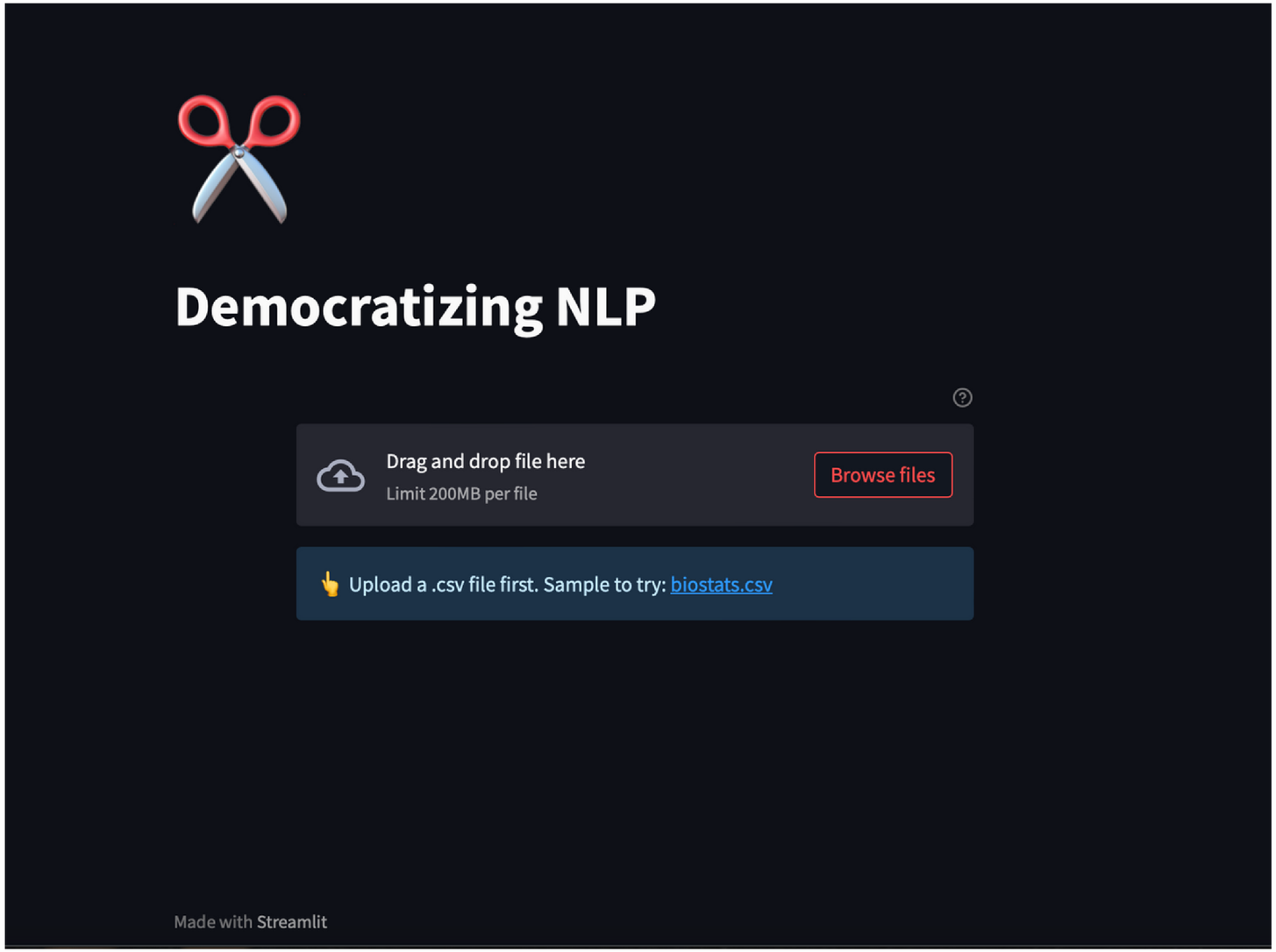
A provision has been provided in the UI for downloading the required model. For each algorithm, the User can check the Accuracy Score of the Model using the training and testing Data Sets and then can download the required model, preferably the most accurate model.

Dataset

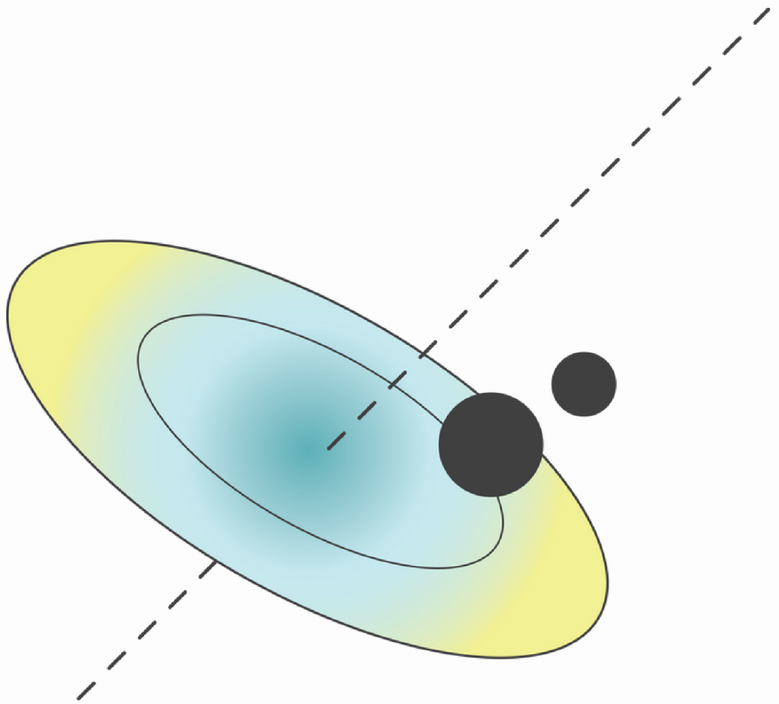
The Pre-Processed Dataset ca also be downloaded from the StreamLit UI. The dataset, which has undergone various pre-processing standards like Stemming, Stop Words removal etc can be downloaded.

StreamLit UI

User-Interface

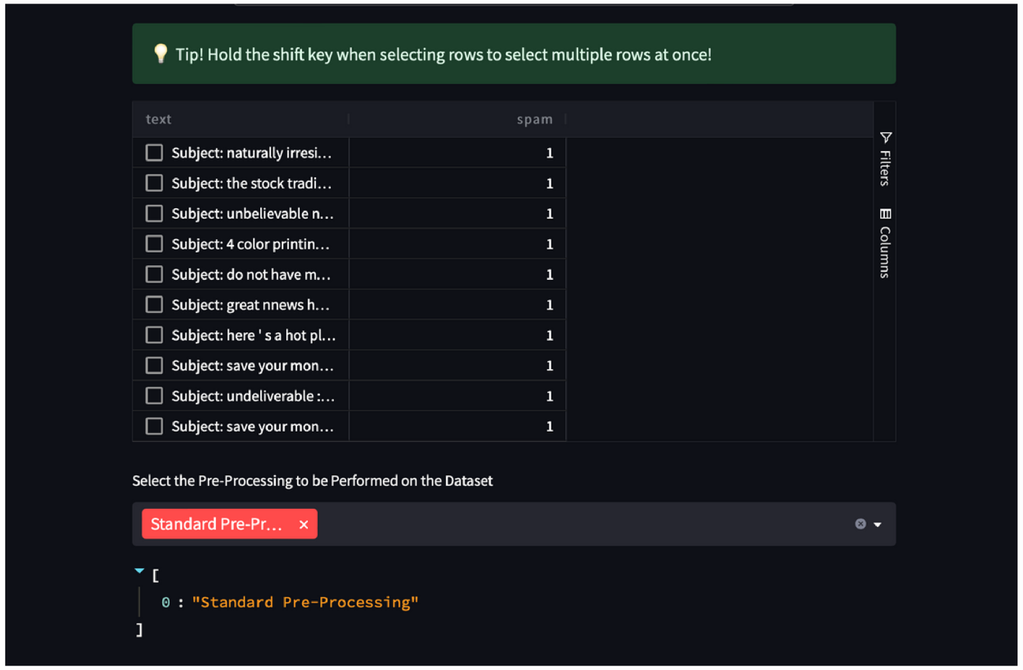


Phase 1



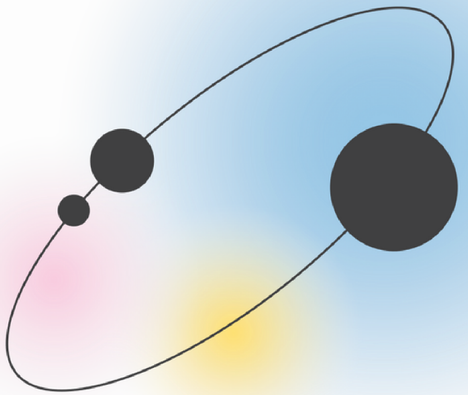
Upload the Dataset

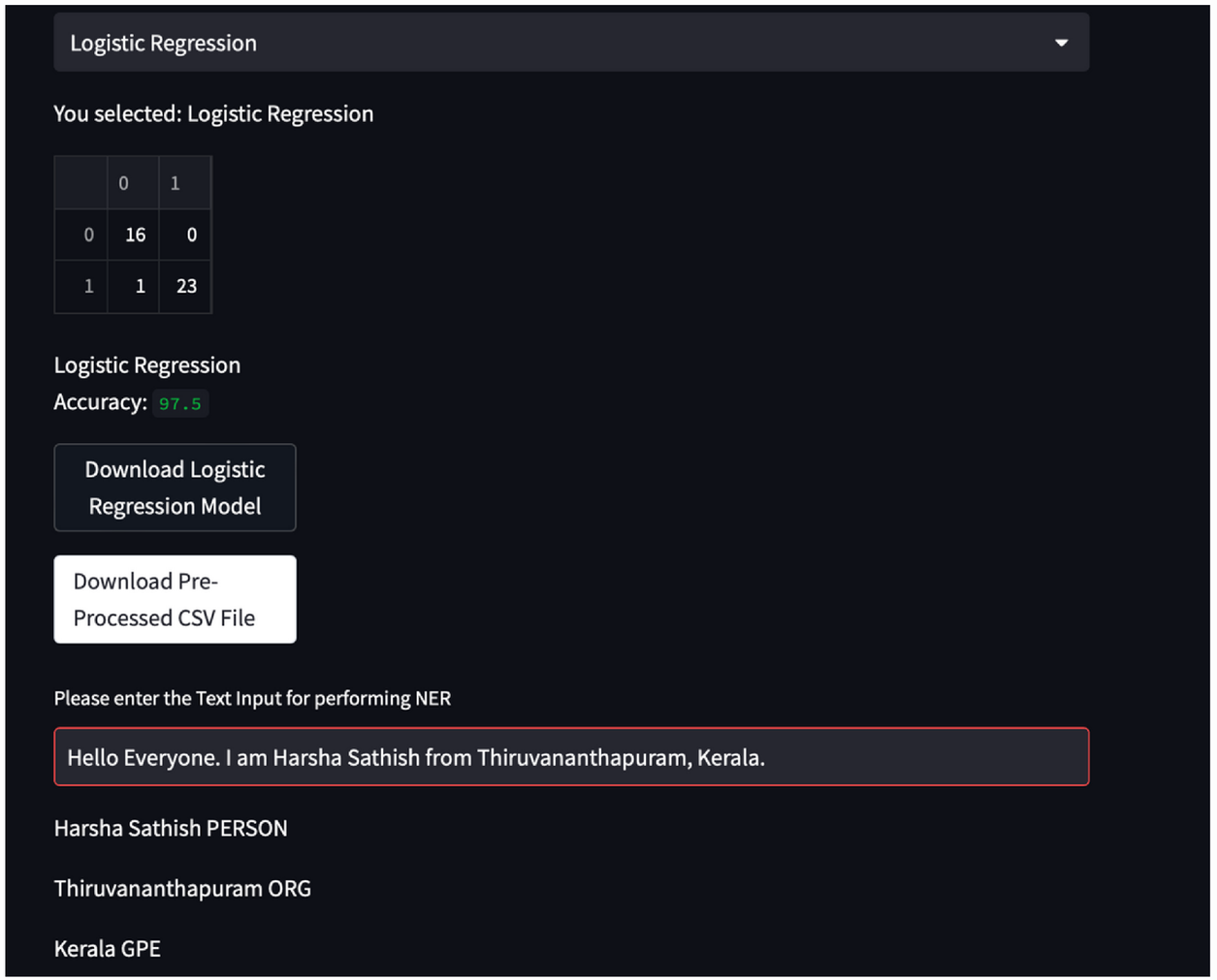
- Upload the Dataset which has to be Pre-Processed
- The Uploaded Dataset will be displayed
- The Pre-Processing mechanisms which are to be used, can be selected using the drop down menu
- The selected options will be displayed



Phase 2

- The Algorithm to be used in the Dataset can be selected from the Drop-Down Menu.
- The Model and the Pre-Processed File can be downloaded as a pickle file.
- NER is given as a text box.



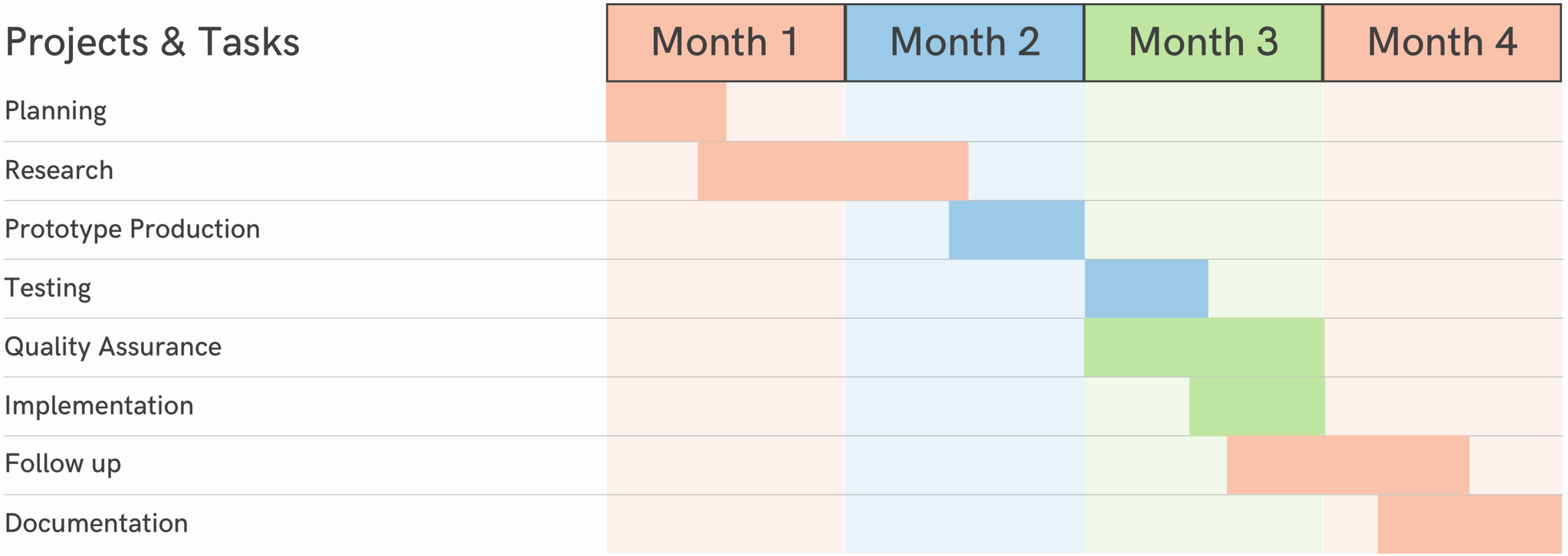


Thank you for attending!

For more detailed reports,
please send us a message!!

Our Current Focus

These are the projects our teams will focus on in the coming months.



Other Important Updates

We've hit 2,000,000 in total guests over DYP's lifetime

Present with ease and wow any audience with Canva Presentations. Choose from over a thousand professionally-made templates to fit any objective or topic. Make it your own by customizing with text and photos.

We're growing to 10,000 employees across 90 hotels globally

Apply page animations and transitions to emphasize ideas and make it even more memorable. Find the magic and fun in presenting by pressing C for confetti, D for drumroll, and O for bubbles.

We're 1 new hotel ahead of schedule this year

Collaborate in real-time and feel like you’re in the same room as your teammates or co-presenters. Share tasks and work simultaneously to create a powerful presentation.

Resource Page

Use these icons and illustrations
in your Canva Presentation.
Happy designing!