horizontal line



Democratizing NLP - A No Code Framework

**Aug 3, 2022**

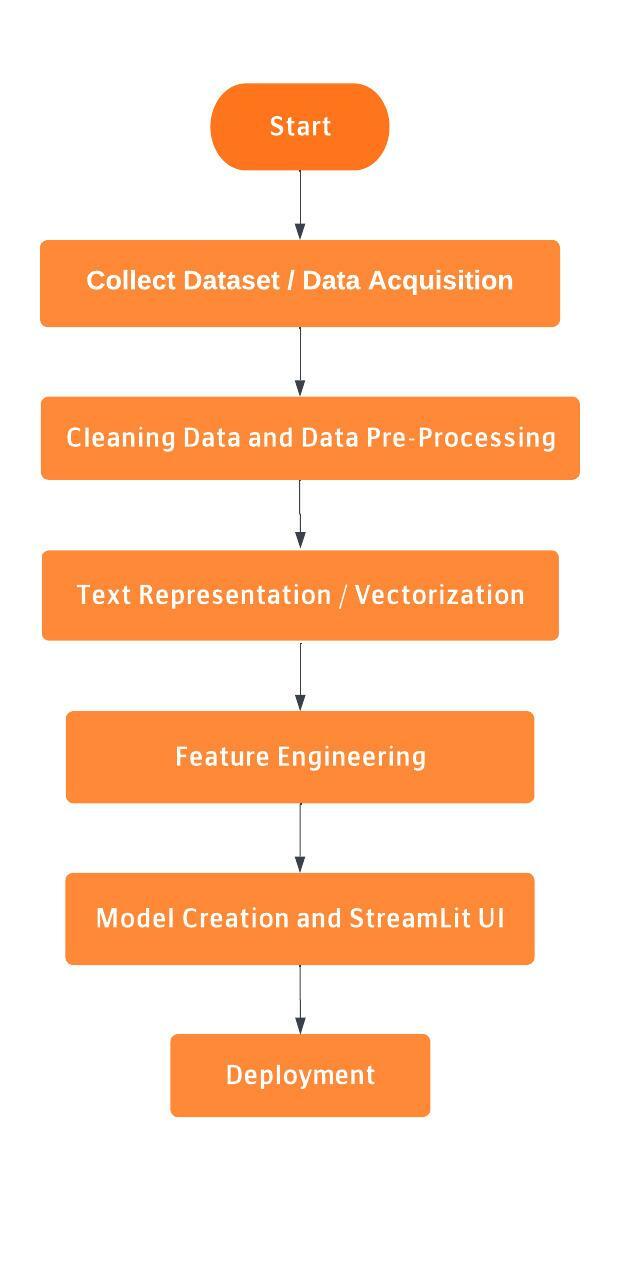
# OVERVIEW

* Perform an initial Research on the Project - Democratizing NLP - A No Code Framework and submit the report

# GOALS

1. Identify common tasks in NLP
2. Identify data analysis done as part of most NLP Projects
3. Identify Text Pre - Processing / Cleaning Activities done

# PROJECT PIPELINE



# COMMON TASKS IN NATURAL LANGUAGE PROCESSING

# Text Pre - Processing

## Text Cleaning

The Dataset may contain unwanted data. Therefore, all the characters will be converted to lowercase, and the punctuation characters will be removed.

* **Stop Words Removal**

Stop Words are the words which are most commonly used in any language, and they do not provide any contribution to the meaning of the sentence. Therefore, these stop words are removed from the dataset for giving more focus to the important words.

* **Removal of Null Values**

If the Dataset contains any null values within the data, the entire row will be deleted so as to not create errors while processing the dataset.

* **Removal of Duplicate Values**

If the dataset contains any row which is exactly same as another row in the dataset, then only one of these rows is retained, and the other rows are deleted.

**Tokenization**

Tokenization of the corpus can be done in 2 ways : Word Tokenizer and Sentence Tokenizer. According to the input given by the user, the corpus can be tokenized in either of the two ways. Named Entity Recognition (NER) can be used for obtaining a better result in Tokenization. NER can be used to extract important entities such as names of People, Location and Organizations.

**Text Normalization**

Converting Tokens into Standard Form. We use both Stemming and Lemmatization. These operations reduce the volume of the text. The User can choose which methodology to use, or can select both.

* **Stemming**

The suffixes from the words are removed to give the stem word. The stem words are inflected with the affixes to produce the given word. Porter Stemmer, Snowball Stemmer and Lancaster Stemmer are few commonly used Stemming Algorithms. The problem with Stemming is that there are words which do not have a proper stem word after removal of the suffix, therefore, they are meaningless.

* **Lemmatization**

The process of mapping all different forms of the word into the root word. All root words are lexicographically correct, unlike Stemming. Lemmatization can be implemented using WordNet. Lemmatization used where the meaning of the stem words are given importance.

**POS Tagging**

If the Machine Learning approach is towards obtaining the meaning of the words, Parts of Speech Tagging can be used. POS Tagging obtains the meaning of the sentence. It tags each word in the sentence with their respective POS - noun, verb.

**Syntactic Parsing**

Recognize a sentence and to generate a grammatical structure to it.

**Vectorization**

* **One Hot Encoding**
* **N - grams**
* **BoW ( Bag of Words)**
* **Term Frequency - Inverse Document Frequency (TF - IDF)**
* **Word2Vec**
* **Glove**

**DATA ANALYSIS DONE IN NLP PROJECTS**

Python provides many packages for visually representing the data. Therefore, it really helps which doing Exploratory Data Analysis on the Dataset.

**Some of the Data Analysis methods are:**

* **Word Frequency Analysis**

Using barplots, we can represent the frequency of each feature using bars. Therefore, we can compare the frequencies of different features.

The average length of each row of each feature can also be represented using bar plots.

* **Stop Word Analysis**

The average occurrence of stop words can be calculated and represented using bar plots. If the stop words count is high, then the dataset can be reconsidered for applying ML algorithm.

* **Latent Dirichlet Allocation (LDA)**

LDA can be used for Topic Modelling. Each document is represented by the distribution of topics and each topic is represented by the distribution of words.

* **Python Word Cloud**

For Text Data, Word Cloud is a good way for representing Textual Data. For this, the data will be converted into a corpus. The size and color of each word that appears in the wordcloud indicate it’s frequency or importance.

* **Sentiment Analysis**

It is used to determine whether the sentence is positive, negative or neutral.

Therefore, by using Bar Plots, Sea born, Pie charts etc, we can display the results of the Exploratory Data Analysis on the Dataset and ascertain various features of the dataset.

* **Readability of a Text Document**

The readability of a text document can be checked using various readability indices like Flesch Reading Ease, Dale Chall Readability Score, and Gunning Fog Index. For this, we can use the textstat library in Python.

Therefore, since there is no hard and fast rule for Exploratory Data Analysis, a number of operations can be performed on the Dataset.

**Project Mentor**