# Thinkerbell Labs Assignment (Annie Software Engineering - Python)

Git-hub Link - https://github.com/nayank07/ThinkerbellLab

## Approach Used

- Dataset considered for the assignment is checked properly.
- Then it is imported into python using panda library and pre- processed to remove the duplicate, null values.
- Text values are checked for stop- words, punctuations and converted to meaningful string array using lemmatization.
- Using term frequency(tf) and inverse document frequency(idf) frame-works, all dataset string array is converted into vector array.
- Naive Bayes method is used to train the model and predict the result.

## Further Improvements

- Recently, I started to learn ML and understand about the NLP.
- If I get a chance to work with it further, I can learn more about creation of word vector and train it using multilayer learning model to get better results.
- Currently the model trained cannot predict the correct spam since it only considers the given data set. By using word2vec we can extend it into bag of words and get better results.
- Website Design to be implemented to get input directly from user and give them output regarding spam or not.

Nishchay N
PES1201801669
PES University