

## AIML Online

### Project Milestone – Natural Language Processing (Project 1)

This is a comprehensive project for all the topics you have learned so far from all the modules. This project will give a good grip on most of Natural Language Processing techniques you have learned in this module.

Criteria	Marks	Week	Topic
Part A.1 a - Clearly write outcome of data analysis (Minimum 2 points)	2 pts	Python	Pandas functions
Part A.1 b i) - Missing value analysis and imputation.	1 pts	Python	Pandas functions
Part A.1 b ii) - Eliminate Non-English textual data.	2 pts	Python	Pandas functions
Part A.2 a - Eliminate All special Characters and Numbers	2 pts	Pre-work	RegEx
Part A.2 b - Lowercase all textual data	1 pts	Python	Pandas functions
Part A.2 c - Remove all Stopwords.	1 pts	Week 1	NLP basics
Part A.2 d - Remove all extra white spaces	1 pts	Week 1	NLP Basics
Part A.3 a - Create dependent and independent variables	2 pts	SUL	Hands On Linear Regression-Model building
Part A.3 b - Split data into train and test.	1 pts	SUL	Hands On Linear Regression-Train test split
Part A.3 c- Vectorize data using any one vectorizer.	2 pts	Week 1	TF-IDF
Part A.3 d - Build a base model for Supervised Learning - Classification.	2 pts	SUL/EST	Any model of your choice. Ex: Decision Tree
Part A.3 e - Clearly print Performance Metrics.	1 pts	SUL	Performance Measures - Precision and Recall, Evaluation of Models
Part A.4 a - Experiment with other vectorisers.	4 pts	Week-1	Explore sklearn feature extraction.
Part A.4 b - Build classifier Models using other algorithms than base model.	4 pts	SUL/EST	Any model of your choice

Part A.4 c - Tune Parameters/Hyperparameters of the model/s.	4 pts	FMT	Ridge/Lasso, Hyper-parameter tuning.
Part A.4 d - Clearly print Performance Metrics.	2 pts	FMT	Model Performance measures.
Part A.5 a - Which vectorizer performed better? Probable reason?	2 pts	Week-1	Your observations
Part A.5 b - Which model outperformed? Probable reason?	2 pts	Week-1	Your observations
Part A.5 c - Which parameter/hyperparameter significantly helped to improve performance? Probable reason?	2 pts	Week -1	Your observations
Part A.5 d - According to you, which performance metric should be given most importance, why?	2 pts	Week -1	Your reasoning
Part B 1. Start chat session with greetings and ask what the user is looking for	5 pts	Python	Functions and loops
Part B 2. Accept dynamic text-based questions from the user. Reply back with relevant answer from the designed corpus.	10 pts	Week 1, SUL	Text preprocessing, Bag of words, Classifier of your choice
Part B 3. End the chat session only if the user requests to end else ask what the user is looking for. Loop continues till the user asks to end it.	5 pts	Python	Functions, while, for loops.

**Wishing the best for every step in this journey!**

\*\*\*\*\*Happy Learning\*\*\*\*\*