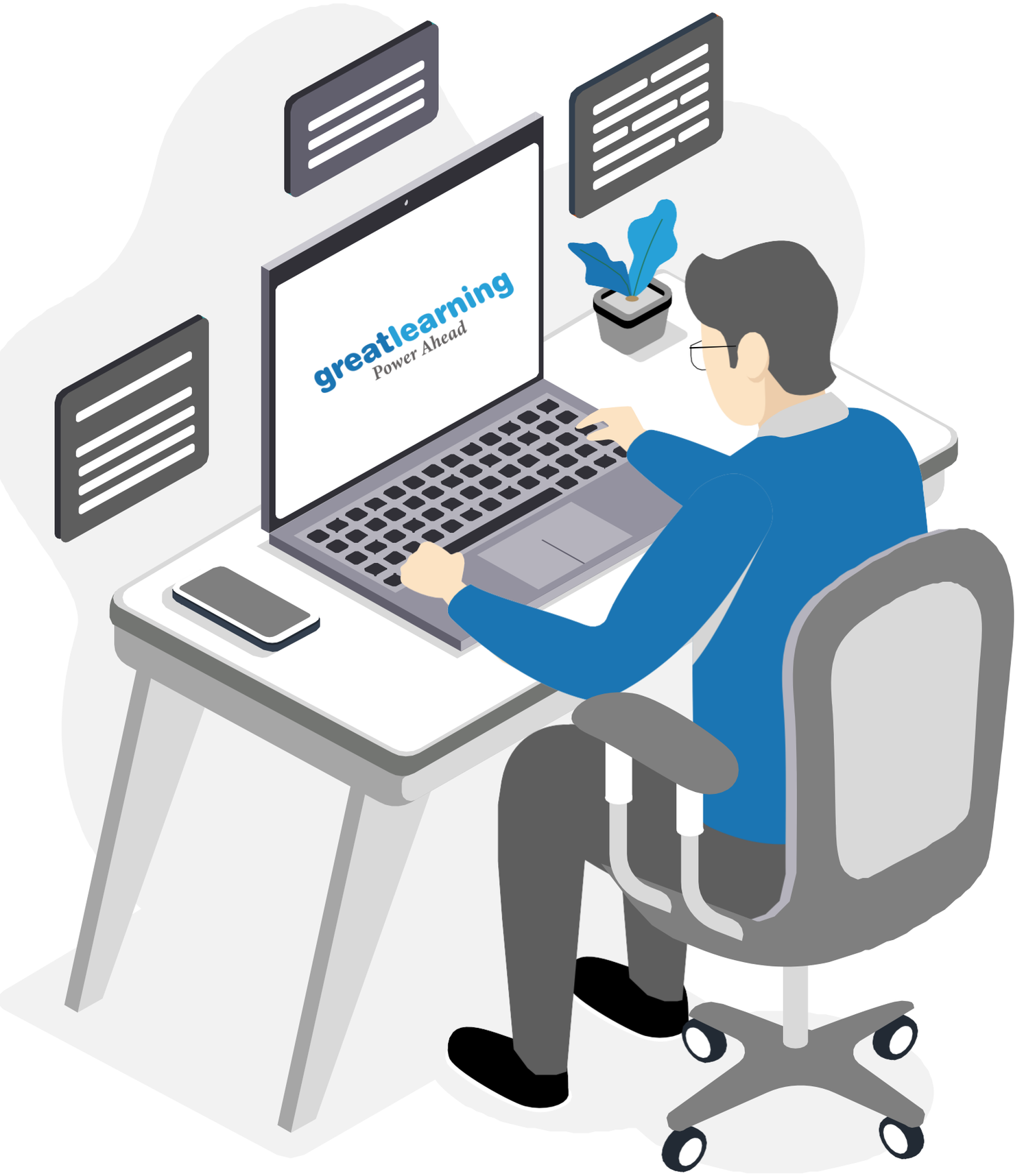
greatlearning
Power Ahead



AIML | MODULE PROJECT

greatlearning
Power Ahead

5 TAKEAWAYS

1. AIML module projects are designed to have a detailed hands on to integrate theoretical knowledge with actual practical implementations.

2. AIML module projects are designed to enable you as a learner to work on realtime industry scenarios, problems and datasets.

3. AIML module projects are designed to enable you simulating the designed solution using AIML techniques onto python technology platform.

4. AIML module projects are designed to be scored using a predefined rubric based system.

5. AIML module projects are designed to enhance your learning above and beyond. Hence, it might require you to experiment, research, self learn and implement.

AIML MODULE PROJECT

STATISITCAL NLP

PART I

AIML module project part I consists of industry based NLP dataset which can be used to design a text classifier using NLP and AIML techniques and models.

PART II

AIML module project part II consists of industry based problem statement which can be solved by designing a semi-rule based NLP chatbot utility.

TOTAL SCORE 60

PART ONE | PROJECT BASED | TOTAL SCORE | 40

- **DOMAIN:** Digital content management
- **CONTEXT:** Classification is probably the most popular task that you would deal with in real life. Text in the form of blogs, posts, articles, etc. is written every second. It is a challenge to predict the information about the writer without knowing about him/her. We are going to create a classifier that predicts multiple features of the author of a given text. We have designed it as a Multi label classification problem.
- **DATA DESCRIPTION:** Over 600,000 posts from more than 19 thousand bloggers The Blog Authorship Corpus consists of the collected posts of 19,320 bloggers gathered from blogger.com in August 2004. The corpus incorporates a total of 681,288 posts and over 140 million words - or approximately 35 posts and 7250 words per person. Each blog is presented as a separate file, the name of which indicates a blogger id# and the blogger's self-provided gender, age, industry, and astrological sign. (All are labelled for gender and age but for many, industry and/or sign is marked as unknown.) All bloggers included in the corpus fall into one of three age groups:
  - 8240 "10s" blogs (ages 13-17),
  - 8086 "20s" blogs(ages 23-27) and
  - 2994 "30s" blogs (ages 33-47)

  For each age group, there is an equal number of male and female bloggers.
  Each blog in the corpus includes at least 200 occurrences of common English words. All formatting has been stripped with two exceptions. Individual posts within a single blogger are separated by the date of the following post and links within a post are denoted by the label url link. Link to dataset: https://www.kaggle.com/rtatman/blog-authorship-corpus
- **PROJECT OBJECTIVE:** The need is to build a NLP classifier which can use input text parameters to determine the label/s of of the blog.

  Steps and tasks: **[Total Score: 40 points]**

  1. Import and analyse the data set.
  2. Perform data pre-processing on the data:
     - Data cleansing by removing unwanted characters, spaces, stop words etc. Convert text to lowercase.
     - Target/label merger and transformation
     - Train and test split
     - Vectorisation, etc.
  3. Design, train, tune and test the best text classifier.
  4. Display and explain detail the classification report
  5. Print the true vs predicted labels for any 5 entries from the dataset.

  Hint: The aim here Is to import the text, process it such a way that it can be taken as an inout to the ML/NN classifiers. Be analytical and experimental here in trying new approaches to design the best model.

PART TWO | PROJECT BASED | TOTAL SCORE | 20

- **DOMAIN:** Customer support
- **CONTEXT:** Great Learning has a an academic support department which receives numerous support requests every day throughout the year. Teams are spread across geographies and try to provide support round the year. Sometimes there are circumstances where due to heavy workload certain request resolutions are delayed, impacting company's business. Some of the requests are very generic where a proper resolution procedure delivered to the user can solve the problem. Company is looking forward to design an automation which can interact with the user, understand the problem and display the resolution procedure [if found as a generic request] or redirect the request to an actual human support executive if the request is complex or not in it's database.
- **DATA DESCRIPTION:** A sample corpus is attached for your reference. Please enhance/add more data to the corpus using your linguistics skills.
- **PROJECT OBJECTIVE:** Design a python based interactive semi - rule based chatbot which can do the following:
  1. Start chat session with greetings and ask what the user is looking for.
  2. Accept dynamic text based questions from the user. Reply back with relevant answer from the designed corpus.
  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.

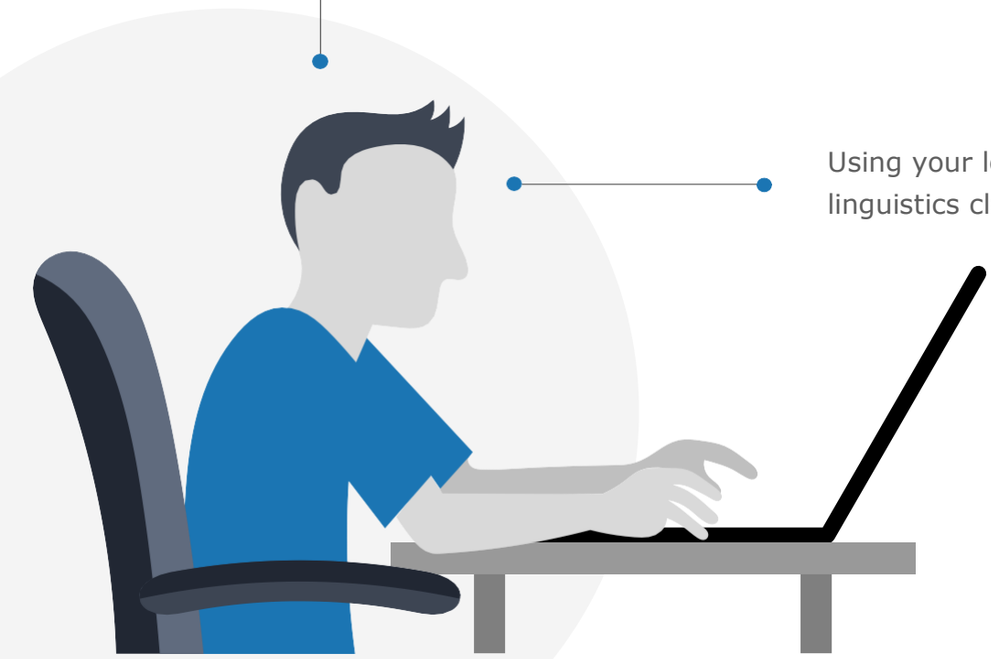
  Please use the sample chatbot demo video for reference.
- **EVALUATION:** GL evaluator will use linguistics to twist and turn sentences to ask questions on the topics described in **DATA DESCRIPTION** and check if the bot is giving relevant replies.

Hint: There are a lot of techniques using which one can clean and prepare the data which can be used to train a ML/DL classifier. Hence, it might require you to experiment, research, self learn and implement the above classifier. There might be many iterations between hand building the corpus and designing the best fit text classifier. As the quality and quantity of corpus increases the model's performance i.e. ability to answer right questions also increases.

Reference: https://www.mygreatlearning.com/blog/basics-of-building-an-artificial-intelligence-chatbot/

LEARNING OUTCOME

Hands on experience on importing, cleansing, pre-processing and computing with huge text dataset. Realtime experience on designing and deploying a text corpus from scratch.



Using your learnings from ML/DL to build a linguistics classifier.

Realtime experience on designing, refining and deploying a text corpus from scratch.

Hands on experience on building a text corps and semi-rule based linguistic chatbot utility using python and AIML models.

IMPORTANT POINTERS

Project should be submitted as a single ".html" and ".ipynb" file. Follow the below best practices where your submission should be:

- ".html" and ".ipynb" files should be an exact match.
- Pre-run codes with all outputs intact.
- Error free & machine independent i.e. run on any machine without adding any extra code.
- Well commented for clarity on code designed, assumptions made, approach taken, insights found and results obtained.

Project should be submitted on or before the deadline given by the program office.

Project submission should be an original work from you as a learner. If any percentage of plagiarism found in the submission, the project will not be evaluated and no score will be given.

greatlearning Power Ahead

HAPPY LEARNING