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

greatlearning
Power Ahead

SEQUENTIAL NLP

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

TOTAL SCORE 60

PART ONE | PROJECT **BASED** | TOTAL SCORE | 30

- **DOMAIN:** Digital content and entertainment industry

- **CONTEXT:** The objective of this project is to build a text classification model that analyses the customer's sentiments based on their reviews in the IMDB database. The model uses a complex deep learning model to build an embedding layer followed by a classification algorithm to analyse the sentiment of the customers.

- **DATA DESCRIPTION:** The Dataset of 50,000 movie reviews from IMDB, labelled by sentiment (positive/negative). Reviews have been preprocessed, and each review is encoded as a sequence of word indexes (integers). For convenience, the words are indexed by their frequency in the dataset, meaning the for that has index 1 is the most frequent word. Use the first 20 words from each review to speed up training, using a max vocabulary size of 10,000. As a convention, "0" does not stand for a specific word, but instead is used to encode any unknown word.

- **PROJECT OBJECTIVE:** Build a sequential NLP classifier which can use input text parameters to determine the customer sentiments.

Steps and tasks: **[Total Score: 30 points]**

1. Import and analyse the data set.
   Hint: - Use `imdb.load_data()` method
   - Get train and test set
   - Take 10000 most frequent words
2. Perform relevant sequence adding on the data
3. Perform following data analysis:
   - Print shape of features and labels
   - Print value of any one feature and it's label
4. Decode the feature value to get original sentence
5. Design, train, tune and test a sequential model.

   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.

6. Use the designed model to print the prediction on any one sample.

PART TWO | PROJECT BASED | TOTAL SCORE | 30

- **DOMAIN:** Social media analytics

- **CONTEXT:** Past studies in Sarcasm Detection mostly make use of Twitter datasets collected using hashtag based supervision but such datasets are noisy in terms of labels and language. Furthermore, many tweets are replies to other tweets and detecting sarcasm in these requires the availability of contextual tweets.In this hands-on project, the goal is to build a model to detect whether a sentence is sarcastic or not, using Bidirectional LSTMs.

- **DATA DESCRIPTION**:

  The dataset is collected from two news websites, theonion.com and huffingtonpost.com.

  This new dataset has the following advantages over the existing Twitter datasets:

  Since news headlines are written by professionals in a formal manner, there are no spelling mistakes and informal usage. This reduces the sparsity and also increases the chance of finding pre-trained embeddings.

  Furthermore, since the sole purpose of TheOnion is to publish sarcastic news, we get high-quality labels with much less noise as compared to Twitter datasets.

  Unlike tweets that reply to other tweets, the news headlines obtained are self-contained. This would help us in teasing apart the real sarcastic elements

  Content: Each record consists of three attributes:

  is_sarcastic: 1 if the record is sarcastic otherwise 0

  headline: the headline of the news article

  article_link: link to the original news article. Useful in collecting supplementary data

  Reference: https://github.com/rishabhmisra/News-Headlines-Dataset-For-Sarcasm-Detection

- **PROJECT OBJECTIVE:** Build a sequential NLP classifier which can use input text parameters to determine the customer sentiments.

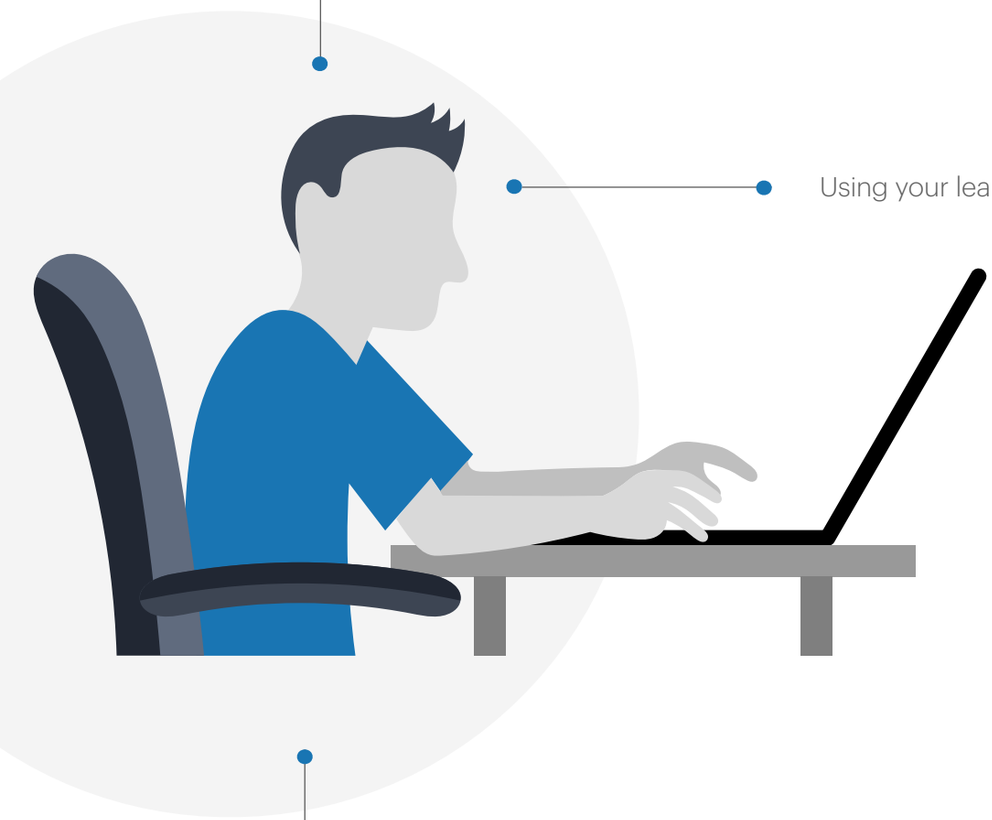
Steps and tasks: **[Total Score: 30 points]**

1. Read and explore the data
2. Retain relevant columns
3. Get length of each sentence
4. Define parameters
5. Get indices for words
6. Create features and labels
7. Get vocabulary size
8. Create a weight matrix using GloVe embeddings
9. Define and compile a Bidirectional LSTM model.

   Hint: Be analytical and experimental here in trying new approaches to design the best model.
10. Fit the model and check the validation accuracy

LEARNING OUTCOME

Hands on experience on importing, pre-processing and computing a text dataset using python.



Using your learnings on text embeddings.

Realtime experience working on designing, training, tuning and testing sequential NLP classifiers.

" Put yourself in the shoes of an actual "

DATA SCIENTIST

THAT's YOU

Assume that you are working at the company which has received the above problem statement from internal/external client. Finding the best solution for the problem statement will enhance the business/ operations for your organisation/project. You are responsible for the complete delivery. Put your best analytical thinking hat to squeeze the raw data into relevant insights and later into an AIML working model.

PLEASE NOTE

Designing a data driven decision product typically traces the following process:

1. Data and insights:
   Warehouse the relevant data. Clean and validate the data as per the the functional requirements of the problem statement. Capture and validate all possible insights from the data as per the the functional requirements of the problem statement. Please remember there will be numerous ways to achieve this. Sticking to relevance is of utmost importance. Pre-process the data which can be used for relevant AIML model.
2. AIML training:
   Use the data to train and test a relevant AIML model. Tune the model to achieve the best possible learnings out of the data. This is an iterative process where your knowledge on the above data can help to debug and improvise. Different AIML models react differently and perform depending on quality of the data. Baseline your best performing model and store the learnings for future usage.
3. AIML end product:
   Design a trigger or user interface for the business to use the designed AIML model for future usage. Maintain, support and keep the model/ product updated by continuous improvement/training. These are generally triggered by time, business or change in data.

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