AIML | MODULE PROJECT

Statistical **NLP**

TOTAL **SCORE** 60

General Instructions:

1. Submission of all the parts is expected in 1 notebook only

2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only

3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.

4. If output for any code cell is missing, 50% marks will be deducted.

5. Any kind of Plagiarism will lead to 0 (zero) Marks.

Submission Format:

1. '.ipynb' (Jupyter Notebook) and

2. '.html' (Jupyter Notebook > File > Download as > HTML)

5 Marks will be deducted if submission in any of the formats is missing.

Part A - 40 Marks

- **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. are 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.
- **PROJECT OBJECTIVE:** To build a NLP classifier which can use input text parameters to determine the label/s of the blog. Specific to this case study, you can consider the text of the blog: 'text' feature as independent variable and 'topic' as dependent variable.

Steps and tasks: [Total Score: 40 Marks]

1. Read and Analyse Dataset. [5 Marks]
   A. Clearly write outcome of data analysis(Minimum 2 points) [2 Marks]
   B. Clean the Structured Data [3 Marks]
      i. Missing value analysis and imputation. [1 Marks]
      ii. Eliminate Non-English textual data. [2 Marks]
      Hint: Refer 'langdetect' library to detect language of the input text)
2. Preprocess unstructured data to make it consumable for model training. [5 Marks]
   A. Eliminate All special Characters and Numbers [2 Marks]
   B. Lowercase all textual data [1 Marks]
   C. Remove all Stopwords [1 Marks]
   D. Remove all extra white spaces [1 Marks]
3. Build a base Classification model [8 Marks]
   A. Create dependent and independent variables [2 Marks]
   Hint: Treat 'topic' as a Target variable.
   B. Split data into train and test. [1 Marks]
   C. Vectorize data using any one vectorizer. [2 Marks]
   D. Build a base model for Supervised Learning - Classification. [2 Marks]
   E. Clearly print Performance Metrics. [1 Marks]
   Hint: Accuracy, Precision, Recall, ROC-AUC
4. Improve Performance of model. [14 Marks]
   A. Experiment with other vectorisers. [4 Marks]
   B. Build classifier Models using other algorithms than base model. [4 Marks]
   C. Tune Parameters/Hyperparameters of the model/s. [4 Marks]
   D. Clearly print Performance Metrics. [2 Marks]
   Hint: Accuracy, Precision, Recall, ROC-AUC
5. Share insights on relative performance comparison [8 Marks]
   A. Which vectorizer performed better? Probable reason?. [2 Marks]
   B. Which model outperformed? Probable reason? [2 Marks]
   C. Which parameter/hyperparameter significantly helped to improve performance?Probable reason?. [2 Marks]
   D. According to you, which performance metric should be given most importance, why?. [2 Marks]

Part B - 20 Marks

- **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. [5 Marks]

  2. Accept dynamic text based questions from the user. Reply back with relevant answer from the designed corpus. [10 Marks]

  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 Marks]

  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/
- **Evaluation:** 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.