DataNeuron

ASSESSMENT FOR DATA SCIENCE

PROBLEM STATEMENT

Dataset (attached with the task)**:** The data contains a pair of paragraphs. These text paragraphs are randomly sampled from a raw dataset. Each pair of sentences may or may not be semantically similar. The candidate is to predict a value between 0-1 indicating the similarity between the pair of text paras. A sample of a similar dataset will be used as test data, therefore it's crucial to the model solution using provided dataset.

Part A

Build an algorithm/model that can quantify the degree of similarity between the two text-based on Semantic similarity. Semantic Textual Similarity (STS) assesses the degree to which two sentences are semantically equivalent to each other.
1 means highly similar

0 means highly dissimilar

Part B

Deploy the Algorithm/Model built-in Part A in any cloud service provider. Your final algorithm should be exposed as a Server API Endpoint. In order to test this API make sure you hit a request to the server to get the result as a response to the API. The request-response body should be in the following format:

Request body: {"text1": "nuclear body seeks new tech .....", "text2": "terror suspects face arrest ..... "}
Response body: {"similarity score": 0.2 }
Note: "text1", "text2", and "similarity score" keys should be kept as it is, without any change.

THE FINAL SUBMISSION MUST INCLUDE THE FOLLOWING -

- - **Live API endpoint(IP Address of hosted app) of the Algorithm Deployed on the Server**
- - Complete Code for Part A and Part B (.py files)
- - 1-2 page short Report explaining only the core approach taken in Part A and Part B.
- - Your updated resume with contact number

INSTRUCTIONS

- Use only Python programming language
- The correctness of similarity scores on test data will be evaluated from the results obtained from the Server Response.
- Task evaluation is equally based on both Part A and Part B. Finally delivery of task A is through task B itself. Therefore it's mandatory to attempt both parts.
- Please ensure the structure of the API endpoint is as per requirement.
- Code must be well commented
- Use any approach to solve algorithms using Statistical models Machine Learning or Deep Learning
- Use any cloud service providers to deploy solutions eg. Azure, GCP, AWS, Heroku, etc.

- Candidates will be judged on three criteria namely the Model/Algo approach, Successfully deployed API, and API response results on test data.
- Time duration: 2 days from the day of receiving the task.

NOTE

1. The given dataset does not contain any label. Therefore, can be treated as an unsupervised learning problem. However, this does not imply that supervised techniques/algorithms are not applicable. The candidate is free to use any technique.
2. Please attach your updated resume and contact information with the submission mail.
3. Your time should start from when this task was sent to you.
4. If you intend to take more than 2 days, you may do so without permission. However, it would be appreciated if you state the reasons for the delay in your report.
5. Every step in the task is self-explanatory to the best of our knowledge. If any part is unclear, use your best judgment and mention it in your report.
6. Your project will not be used for the benefit of the company in any manner. The intention of this task is ONLY to evaluate your skills.
7. Your submission will showcase your skills and knowledge of the said field and help us evaluate your candidature in a better manner, so kindly try to keep the work as original as possible.
8. The deployed server can be closed after the final results are announced. We recommend that candidates should use freely available resources only to deploy their APIs.
9. Final submission must be sent at abhishek.kumar@dataneuron.ai and cc mail@dataneuron.ai Submissions via any other platform will not be considered.
10. We wish you all the best!