Precily

ASSESSMENT FOR DATA SCIENCE

Task- Finding Semantic Textual Similarity

PROBLEM STATEMENT

Given two paragraphs, 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. The STS task is motivated by the observation that accurately modelling the meaning similarity of sentences is a foundational language understanding problem relevant to numerous applications including machine translation (MT), summarization, generation, question-answering (QA), short answer grading, semantic search.

STS is the assessment of pairs of sentences according to their degree of semantic similarity. The task involves producing real-valued similarity scores for sentence pairs.

Dataset is attached with the task.

The data contains a pair of paragraphs. These text paragraphs are randomly sampled from a raw dataset. Each pair of the sentence may or may not be semantically similar. The candidate is to predict a value between 0-1 indicating a degree of similarity between the pair of text paras.

1 means highly similar

0 means highly dissimilar

INSTRUCTIONS

- Use any programming language. (Preferred: Python)
- Use the given dataset (link provided above)
- Code must be well commented
- Use any approach you want using Statistical models/ Machine Learning/ Deep Learning
- Time duration: 3 days from the day of receiving the task.

FINAL SUBMISSION MUST INCLUDE THE FOLLOWING -

- **CSV file of similarity scores with Unique_ID. (Columns : Unique_ID, Similarity_Score)**
- Complete Code
- Text Report explaining the approach you implemented and why.
- Your updated resume with contact number

EVALUATION

- Candidates who have used Deep Learning approaches will be preferred for further rounds.
- The correctness of similarity scores will be evaluated from the actual similarity values of the paragraphs.
- Candidate Selection is not based only on the accuracy of the results but also the proposed approach.

Precily

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 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 3 days, you may do so without permission. However, it would be appreciated if you state the reasons for 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 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. Final submission must be sent at anurag.sharma@precily.com. Submissions via any other platform will not be considered.
9. We wish you all the best!