Data Scientist STC Assessment

Department for Outcomes

Instructions

This test is composed of two parts. Please answer both parts to the best of your capabilities.

Please ensure that your solutions are reproducible and readable by submitting commented code and a brief write-up of your approach and findings, in addition to the deliverables specified in the questions. You may use any software, programming language and/or publicly available packages to develop your solutions.

Email your solutions in the form of a zip file containing all relevant materials to Federico Tiberti (ftiberti@worldbank.org) before **April 21st, 10am EST**. If you have any questions while taking the test, please address them to Federico Tiberti.

Part 1: Data Extraction and Exploration

Problem Objective

This problem measures the candidate's competency in automation, understanding and accessing APIs, unstructured data processing, and visualization.

Problem Description

- Using the World Bank Documents & Reports API, extract metadata related to documents under the document type "Implementation Completion Report Review", under the major document type "Project Documents", published from January 1st, 2019 to April 15th, 2025.
- "Implementation Completion Report Review", or ICRR, is a document prepared by the Independent Evaluation Group (IEG), an agency within the WBG that conducts evaluation of the Bank's operational and analytical products. The ICRR evaluates the quality of the ICR (Implementation Completion and Results Review), a document produced by projects' task teams at the time of the closure of a project outlining the projects main milestones and results.
- Then, extract country names from the document title (not from any other metadata component).
- Produce an animated chart showing the evolution of the number and the percentage of documents by country and year.

Part 2: Natural Language Processing and Text Analysis

Problem Objective

This problem measures the candidate's competency in natural language processing, text classification and inference.

Problem Description

- Using the URL metadata extracted from the API in Part 1, run a bulk download of the documents in .txt format.
- Choose one of (1) Outcome rating, (2) M&E Quality Rating.
- Based on your choice, extract from the document the relevant section that leads up to the rating assignment (ie, "Outcome" or "M&E Design, Implementation, & Utilization") and the corresponding rating assignment.
- Produce an analysis of your choice linking the rating to text features from the document section and, if you choose, document and project metadata. Your analysis should attempt to identify key textual patterns, sentiment/topic indicators, or technical terminology that correlate with specific ratings. If you model this relationship, evaluate the performance of your model using appropriate metrics and briefly discuss potential limitations and biases in your approach.
- Produce visualizations of what you consider to be the main takeaways of your analysis.