Data Engineer Interview Project

OVERVIEW

In this interview project, candidates will be presented with a real-world scenario involving the parsing and analysis of financial data from SEC's EDGAR (Electronic Data Gathering, Analysis, and Retrieval) database. The goal of this task is to assess the candidate's ability to extract relevant information from complex financial documents, demonstrate programming skills, and showcase problem-solving abilities.

PROJECT DESCRIPTION

Candidates will be provided with a set of sample EDGAR filings in HTML format, representing financial reports such as 8-K filings. The objective is to parse these filings, systematically extract companies' quarterly EPS (Earning Per Share) for the latest quarter, and subsequently present this data in a structured format. The primary emphasis is on creating a versatile parser capable of seamlessly handling distinct filing formats.

PROVIDED DOCUMENTS

1. The project description (this document)
2. Training Filings: A total of 50 filings are included.
3. Sample Output for 5 Filings: output_example.csv

COMMONLY ASKED QUESTIONS

1. When both diluted EPS and basic EPS are present in the filing, prioritize outputting the basic EPS figure.
2. In cases where both adjusted EPS (Non-GAAP) and unadjusted EPS (GAAP) are provided in the filing, opt to output the unadjusted (GAAP) EPS value.
3. If the filing contains multiple instances of EPS data, output the net or total EPS.

4. Notably, enclosed figures in brackets indicate negative values. For instance, in the majority of filings, (4.5) signifies -4.5.
5. In scenarios where the filing lacks an earnings per share (EPS) value but features a loss per share, output values of the loss per share. Remember the output values should always be negative.

OUTPUT FORMAT

Please generate a CSV file containing two fields: "filename" and "EPS." Presented below is an illustrative output example for five filings.

Example:

filename,EPS

0001564590-20-019726.html,0.08

0000066570-20-000013.html,1.12

0000008947-20-000044.html,-0.41

0001564590-20-019431.html,1.08

0001564590-20-019396.html,-3.15

PROJECT SUBMISSION

Submission must include:

1. A main script named **parser.py** which takes 2 command-line arguments
    a. Input directory path. Example: /home/trex/Training_Filings/
    b. Output file path. Example: /home/trex/output.csv
2. Comprehensive report for the approach and any other details.
3. Output CSV file generated by your parser for all the 50 provided filings.

ASSESSMENT

During the assessment phase, candidates' parsers will be thoroughly evaluated. This evaluation process encompasses the functionality of their parsers on two fronts: **first, parsing and extracting EPS data from the 50 provided sample filings, and second, ensuring the parser's adaptability to handle previously unseen filings.** The ability to seamlessly navigate both the familiar and the unknown filing formats will play a pivotal role in determining the success and robustness of the candidates' parsers. This comprehensive assessment ensures that the parsers not only excel in extracting data from the provided samples but also demonstrate a high degree of generalization to handle an array of unforeseen filing formats.