CONTELLIGENZ

Data Scientist Internship Assignment

Problem Statement:

Given a list of 4000 Hindi movies - Bollywood Movies Dataset.xlsx, you need to write a python script that can perform the below expected functionality

- Search and fetch the IMDB URLs of as many movies as possible
- Fetch the content metadata details from the IMDB website either by web scraping or by using any of the available open APIs or python libraries for each of the above movies and store them in a csv file. Details required are:
    - Title
    - IMDB ID
    - Date of release
    - Genre
    - Cast
    - Crew
    - Plot summary
    - Plot keywords
    - IMDB Rating
    - IMDB Votes
    - Reviews
- Perform additional data processing to come up with more derived fields
    - Age of content - time since release of the content
    - Popularity of content - a score which can be a combination of IMDB rating and votes OR try to innovatively come up with a new definition
    - Cast popularity score - score of the popularity of all of the cast
    - Crew popularity score - score of the popularity of all of the crew
    - Sentiment score - score based on the NLP based sentiment analysis of the reviews
    - Review keywords - extract the top most relevant keywords from the reviews of the movie
- Perform exploratory data analysis to come up with some of the below insights
    - Genre distribution of titles

    - Top 10 most acted actors etc.
- Build a REST API to perform the below expected functionality
    - GET top n popular movies
    - GET top n popular movies with additional filtering options using below
        - Year of release
        - Genres
        - Plot keywords

Guidelines for the submission:

- Prepare a python notebook with detailed and meaningful comments for each of the operations performed
- Specify all the packages that are required
- Need the most optimised code possible
- Documents to be submitted as part of the submission process:
    - Python Notebook
    - Any additional documents used - readme, requirements etc.

Submission to be emailed at jayadev@contelligenz.com