OKRA.ai

Train travel comments dataset

One of our customers would like to evaluate the new train schedules in the UK. As you might know, these have recently been changed. The customer would like to know what people are discussing and what they could change to improve their service.

There are various data sources online which can help you answer this question. We have taken review data from Trustpilot.com. It consists of about 2000 reviews of various UK train companies. The data is stored in a JSON file and contains the comment itself, when it was submitted, to which site it was submitted, and the review score the user gave.

The objective of this task is to extract the topics that people are talking about and provide insights that help the client calibrate their internal strategy on how to improve customer experience. Use whichever tools you feel are appropriate for the task, given the time available.

Deliverables and Evaluation

Please send your work in a single Jupyter notebook, structure it in logical steps, and make sure that the solution addresses the business questions raised. We evaluate your assignment on:

1. Your understanding of the data
2. Data preprocessing and clean-up
3. Your approach to identifying the main topics of customer reviews
4. Presentation of insights that can drive internal strategy
5. Overall code quality

Please provide any discussion points and future directions along with any scripts used to automate plotting and analysis. Tools you could consider using: Python, NLTK, Spacy, Pandas, etc.

Important: Please make sure to include sufficient comments and discussion in your output to enable us to understand your thinking and how you tackle this assignment.

Please try to avoid spending more than 5-8 hours on this assignment.

Follow-on discussion

If you're successful, we'll ask you to present your findings in the next interview round with 2 members of our team.