OKRA

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 object of this task is to extract the topics that people are talking about. Use whichever tools you feel are appropriate for the task, given the time available.

Tasks

1. Process the Data
2. Extract the main topics people are talking about.
3. Analyze how the topics changed over time. Do you see a difference in the topics or their frequency?
4. Visualize the results

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

Please send any scripts used to automate plotting and analysis, alongside your discussion. Tools you could consider using: Python, NLTK, Spacy, Pandas, etc. Please spend 5-8 hours on this assignment.

Follow-on discussion

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