OKRA

Okra Technical Assignment

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. Use whichever tools you feel are appropriate for the task, given the time available. Please be sure to include sufficient comments and discussion in your output to enable us to understand your thinking as you tackle this assignment.

Tasks

1. Extract the main topics people are talking about.

2. Analyse how the topics changed over time. Do you see a difference in the topics or their frequency?

3. Visualise the results and prepare a short (15 minute) presentation.

Please send any scripts used to automate plotting and analysis, alongside your discussion. Tools you could consider using: Python, NLTK, Spacy, Pandas etc. You should not spend longer than three hours on this assignment.

Follow-on discussion

During the interview we would like a demonstration of the code and a discussion of findings, which may include any of the following:

· What where the main topics in the data, were they different across the various companies?

· Do you see change of topics over time?

· What other sources could be gathered to help answer the research question?