AI Evaluation Engineer Technical Task

Context

Consult is a topic modelling tool currently being developed by i.AI to analyse responses to public consultations. It has two main stages:

1. Theme generation, which identifies key themes across consultation responses.
2. Theme mapping, where responses are mapped to the list of themes using one-to-many mapping.

Themefinder is a simplified version of Consult that is being made more widely available. In this task, we will ask you to analyse and evaluate self-generated data using Themefinder.

Task

The task has four components, set out below. Try to ensure that all stages of your work are as reproducible as possible.

As an AI team we encourage the use of AI assisted coding. If you do use any AI tools, we're very interested to learn about how you use them. Therefore please let us know what tools you used as part of completing this task.

You can use the following API key:

```python
AZURE_OPENAI_ENDPOINT="https://iai-azoai-interview.openai.azure.com/"
AZURE_OPENAI_API_KEY="6FrJnxZWsBwkXOpFtb16DmpS5ULgoodO1m3A0DpDIj6Q9PmhnwKCJQQJ99BCACi0881XJ3w3AAABACOGqhml"
OPENAI_API_VERSION="2024-08-01-preview"
DEPLOYMENT_NAME="gpt-4o"
```

Complete the following 4 tasks (they continue over the page!):

1 Generate synthetic data

For the remainder of this task, you will be working with synthetic responses to a public consultation, which you now need to create. **Use an LLM to generate 300 plausible responses to the question:**

What changes would you like to see in the education system in your area over the next five years?

Save any of your working, as well as the responses.

2 Generate themes

Install ThemeFinder from here: https://pypi.org/project/themefinder/

Use ThemeFinder to generate themes for your synthetic data.

If ThemeFinder is taking a while to run, you can use the toy dataset attached in the email to start working on steps 3 & 4. *If, for any reason, you're unable to get Themefinder to generate themes for your responses, then please write a function to randomly assign theme labels to each of your responses. Remember that more than one label can be assigned to each response.*

3 Generate a second set of themes*

Using the themes from the previous step as a starting point, create a second set of theme mappings of the same format that randomly differs from the first. It should be possible to edit your code to change the degree of randomisation in a clear way (i.e. edit a higher/number lower of theme mappings).

This second set of theme mappings will be used in the next step of the task so please read ahead.

4 Compare the two sets of themes*

Imagine that your first set of themes were generated by Themefinder, and the second set were by a human coder.

Produce a summary paragraph describing the variation between the first and second sets of theme mappings. Your paragraph should include at least one metric that aims to quantify this variation (feel free to use more than one). Both technical and non-technical stakeholders will read the summary paragraph.

Deliverables

At the end of time period (2 hours) respond to the task email with the following attachments:

1. A dataset with your synthetic responses and the two theme mappings
2. Code used to create synthetic responses
3. Summary paragraph describing the variation in theme mapping, and any associated code (these can be separate attachments if preferred)

In your email, please specify whether you used AI assistance for this task and, if so, which tools.

***if you're stuck on step 2, you might find this Json file helpful for steps 3 and 4 (also attached to the email).**