Spectrum Data Science Interview Project

Objective:

Our objective here is generally two-fold: for you to demonstrate your approach to Data Science projects, problem solving, and communications while also learning more about the type of problems we tackle at Spectrum and how we collaborate.

Problem Definition:

At Spectrum, one of our main challenges is working with unstructured (or semi-structured) data and extracting insights as well as building models for our client's particular ecosystem. In this project, we'll use the public 4chan site as our ecosystem because it is both public and it's data can be easily gathered from their API.

4chan has various channels. We'll focus on a few: Technology (/g/), TV & Film (/tv/), Food & Cooking (/ck/), and Literature (/lit). Pull the latest threads (up to 100) for each and perform the following steps:

1. **Extraction/Cleaning**: Extract the comment text (ignore images) and relevant metadata, as time permits (e.g., timestamps). Perform any additional cleaning or pre-processing you feel is necessary or helpful.
2. **Exploratory Analysis**: For each channel, derive insights on common words, messages per thread, and two or more other points of interest you may find.
3. **Discriminative Analysis**: Find the words and signals that differentiate the channels. Use the method you feel appropriate, but be prepared to explain your approach and show concrete examples that back up your insights.
4. **Modeling [as time permits]:** Apply a simple model (e.g., BOW) to classify which channel a post is likely to have come from. For this step, we're not looking for the best model, but rather for a model, metrics on performance, and how you would iterate and refine the model.

Misc:

You may use the programming language and packages that you're most comfortable with. While there are many packages that may perform the above tasks largely off-the-shelf, what we are looking for is that you can understand the tools you're using and can interpret the results and have an idea on how to iterate. Similarly, think about how to present your results. For example, if you're using Python, you may consider using a Jupyter Notebook, but for Java you may share your code via Github and organize and prepare your results in a shared Google Doc (or text document).

We anticipate that the above steps would take roughly an hour or so each to accomplish. If you find yourself taking more than two hours, either move on to the next step or reach out to Jon Purnell (jonathan@getspectrum.io) for help. This time frame means that you may need to compromise; perhaps the data won't be 100% cleaned or there will be some additional exploratory questions you won't have time to answer. During your review, you can highlight what you would do with additional time or what different steps you might take but didn't have time to pursue.

If you have any questions or need clarification at any point, please feel free to email Jon Purnell at any time. He'll respond as quickly as he can and be available as much as possible. We hope you enjoy this project!

For reference, 4chan's API is defined here: https://github.com/4chan/4chan-API
Your code does not need to pull live data; you may manually download data via the API and then have your code ingest it from a local file. For example, to get the posts for Outdoors (/out/) you could use https://a.4cdn.org/po/catalog.json, either saving from your browser to a json file or, if you're familiar with unix, using the `wget` command. While we'd like to see how you extract and clean the text, writing a scraper is beyond the scope of this project.