Applied ML Semantic Search Exercise

Overview:

Grainger is North America's leading broad line supplier of maintenance, repair and operating (MRO) products. For nearly 100 years, we have helped customers access useful information to find the products they need to get their jobs done.

Market:

Our customer base is diverse. Every business in the US buys the types of products that Grainger sells and for that reason, Grainger sells into every industrial segment – from companies developing new uses for nanotechnology to companies involved in anthracite mining.

Business:

Our success is based on our expertise. It's our ability to understand the customer, the products they need, the services they require and the channel in which they prefer to interact with us that has helped Grainger achieve our financial strength. We have a proud history of being an early adopter and innovator of technology, and we're really excited about the road ahead.

Exercise

This exercise is intended to serve as the technical component of the Grainger ML Engineer interview process. Based on your performance on this exercise, you, as the candidate, may be invited to explain and explore your line of thinking, discuss your approach to the problem, explain the analysis that you did leading up to the model building phase and the steps you took to get to a solution. As in any Machine Learning problem, there are no right or wrong answers, there are only iteratively better ones.

Some pointers:

Be ready to explain why you took a certain approach in the case review round that will follow
Do your best to write explainable, modular code.
Feel free to make assumptions but be ready to back them up with reasoning.
Better presentation of your results leads to more productive case reviews
Think of ways of improving your methodologies and be prepared to talk about them on the subsequent call

Problem Statement

Refer Amazon's esci-data link to generate dataset of products and search queries. The features and volume of the product data is comparable to that which we deal with at Grainger. For this exercise we are looking for candidates to build a basic semantic search application and to report the quality of the solution.

Below are the details for the task:

1. Select the training dataset applicable to Task 1 - Query-Product Ranking, with the 'us' product locale and the 'E' esci_label.
2. Create a sample dataset consisting of approximately 500 rows with around 50 unique queries from point number 1. If this doesn't yield the desired dataset, you may use the following steps to generate the sample dataset.
    a. Determine a random sample of 50 unique queries from the dataset derived from point number 1.

   b. Filter the dataset derived from point number 1 to contain only the unique queries from point number 2.a.
   c. Create a sample dataset of 500 rows from the dataset derived from point number 2.b

3. The goal of this project is to create a vector index for the product dataset derived in point number 2 (i.e. columns starting with prefix "product_") and to assess the quality of that index against the search queries provided.
4. A solution to this problem will require: a vector index, an embedding function, and quantified metrics about search performance. If it is beneficial the solution might also contain some secondary ranking logic.
5. Choose an external persistent vector embedding storage option (e.g., LanceDB, Milvus Lite) if in-memory storage is unsuitable.
6. Metrics of particular interest for product search would be HITS@N (N=1,5,10) and MRR.
7. To accomplish the goal of the project a candidate will likely have to iterate over approaches to embedding or different indices to improve search performance
8. The data science team at Grainger have benchmarked a solution to this assignment using a typical Macbook (16GB memory), alternatively feel free to use Google Colab
9. Candidates are expected to provide the following:
   a. Dataset derived in point number 2.
   b. Repository of working code or a notebook that fulfils the above tasks with adequate documentation to explain any design decisions they took. Verifying that the code as submitted can be run is a requirement of this exercise.
   c. Specify and justify any assumptions you are making about the data or design decisions.

Time/Duration

You have ***a week*** to complete the exercise and get back to us with a solution. We expect that a reasonable submission should take between 2-5 hours.

Should you have any questions about the interview, role, company benefits – or otherwise, please feel free to chat with your assigned recruiter. He/She will also work with you to schedule the interview.

Thanks for making time for us and the effort that you are putting in to help us understand your qualifications/expertise/credentials more clearly. We look forward to meeting you in person.