***** IN DASK *****

Q1. Solve 35 points:
Datasets: Restaurants_in_Durham_County_NC.csv
durham-nc-foreclosure-2006-2016.json

1.1. Find food service and active restaurants ("status" = "ACTIVE" and ""rpt_area_desc" = "Food Service") closest to the following coordinate: of 35.994914, -78.897133, and show it.

1.2. With that restaurant in (a) as your center point, find the number of foreclosures within a 1 mile radius

You can use external libraries for calculating coordinate distances. For Python notebooks, the haversine library is available in Jupyterhub's bigdata environment.

Q2. Language Models – 50 points
Dataset: hw1text.zip (provided in class website)

P(A, B) Joint Probability: the probability of both events A and B happening simultaneously. It can be calculated using the formula: P(A, B) = P(A) * P(B) if A and B are independent events.

P(A|B) Conditional Probability: the probability of event A occurring given that event B has already occurred. It can be calculated using the formula: P(A|B) = P(A, B) / P(B).

In NLP work, we are not as interested in the probability of a single word, but instead on the conditional probability of, say, 'york' given that the word 'new' precedes it.

P(york|new) = P("new york") / P(new)
P("new york") = count("new york")/count of bigrams seen
P(new) = count(new)/count of words (unigrams) seen

SOLVE (using DASK):

Compute the conditional probability of the top 3 most likely words following "new".