IBM Data Science Capstone Proposal

Jeremy Nathan

1. Business Understanding

For my capstone project, I would like to be able to generate a sample menu for a given neighborhood based on a representative sample of restaurants' menus in that neighborhood. This sample menu would reflect the culture of that neighborhood by highlighting differential ingredients and preparations. For instance, a sample menu for a neighborhood that has historically been inhabited by fishermen/is located along a coast might prominently feature seafood as an ingredient; sample menus for Italian ethnic enclaves, on the other hand, might highlight tomatoes and pasta as ingredients.

A sample menu would be of great use for people looking to start a restaurant in a particular neighborhood. For instance, if someone wishes to open a restaurant in the neighborhood, they would be able to draw upon this sample menu for inspiration and start by including menu items or ingredients popular with residents in the neighborhood. Alternatively, if an ambitious restauranteur would like to open an establishment that serves different dishes from what already exists in the neighborhood, a sample menu would prove useful to them as well (by providing a list of what is already common in the area).

2. Analytical Approach/Data Understanding

The problem calls for predictive text generation. In the context of the problem, this can be achieved in two parts: First, getting a list of ingredients mentioned in the various restaurant menus in the neighborhood, and then using a subset of that list as a starting point for predictive text generation. Each of the selected ingredients would be fed into the predictive text generator as an input, and the output would be full menu item description. For instance, I could feed in the word "lobster" and the output could be "lobster ravioli with mushrooms and garlic butter". As a starting point, both parts of the problem could be solved by using Recurrent Neural Networks (RNNs), which are the industry standard for named entity recognition (first part) and predictive text generation (second part).

Data will be obtained from Foursquare. I will be specifically making use of the "explore" endpoint to obtain a list of restaurants, then use the "menu" endpoint on each restaurant to obtain a list of its menu items. I will then gather the data into a dataframe with 2 columns: one column denoting the neighborhood, and another column which contains menu item descriptions for each menu item found in the neighborhood. As the menus endpoint is a premium endpoint, I will only be focusing on 2 neighborhoods with distinct restaurant scenes to demonstrate my solution as a proof of concept. Not only will this reduce the number of premium calls I will have to make, but this will drastically reduce model development time, since I will only have to build 2 neural networks.