Measuring Culinary Diversity of Cities

Girish Ganesan

Introduction

When discussing food in a town, one often hears comments about the diversity of cuisines available. Certainly Manhattan / NYC tops the list of the places where one can find authentic cuisines from most parts of the world. When coming to LA or San Francisco, one wonders whether we have the same level of diversity available there. How about Toronto? Or London?

As a Data Scientist I have always wondered whether we could develop a quantitative metric to measure the diversity of cuisines; an absolute metric that could be used to compare culinary diversity across cities.

With that in mind, I set about to develop a Culinary Diversity Index that would measure the diversity of cuisines available in a city. The key idea is simple: we calculate the percentage of restaurants in a city for each type of cuisine. The most diverse city will have an equal percentage for each cuisine. So if we consider a list of 100 cuisines, the ideal city will have 1% of restaurants with each cuisine.

That brings us the next question: How about a city which has “almost 1%” for this number? How would we measure this error? To answer this, we leverage the concept of *Entropy*. In the next section we will motivate the discussion of Entropy and see how it can be applied to our problem of constructing a Culinary diversity index.

Entropy

In the previous section we looked at equitable distribution of cuisines as a possible measure of diversity. Instead of percentages let us look at the fraction of the total restaurants that serve a particular cuisine. If we pick a random restaurant at a location from Foursquare, this fraction will be indicative of the probability that we pick that particular cuisine. Thus if 10% of the restaurants serve Indian cuisine, the probability that a randomly picked restaurant will have Indian Cuisine is 0.1.

The thoughts we had in the previous section can now be cast in a probability framework. Diversity is maximum when all cuisines are equally probable. The more a city deviates from this equiprobable distribution of cuisines, the less diverse it is.

Even when cuisines are equiprobable, the larger the number of cuisines the larger should be the diversity. That is a city with 100 cuisines (each with probability 1/100) should be more diverse than a city with 10 cuisines (each with probability 1./10).

Thankfully, the science of *Information Theory* has given us a tool to measure this type of randomness: Entropy. Consider a city with cuisines c_1, c_2, c_3, ... c_n. Let the fraction of each cuisine be p_1, p_2, p_3, ... p_n. If we pick a restaurant at random the *event* that we will end up picking a cuisine c_k occurs with a probability p_k. If we consider the set of all possible events, the Entropy of this collection of events is:

$$E = -\sum p_k \log_2(p_k)$$

Where $\log_2$ represents the base 2 logarithm.

Examples of Entropy

Consider an unbiased coin. The possible events that result from the coin toss are Head and Tail, each with a probability ½. The entropy of this event collection is:

$E = -½ \log_2(½) + -½ \log_2(½) = -½ * (-1) + -½ * (-1) = 1$

However if the coin is biased, say the probability of Heads is 0.7, then the entropy of the events is only 0.88. And if the coin has two heads (or two tails) then only one event is possible and the Entropy is 0.

Thus we see that the more the distribution is closer to an equal probability distribution, the higher the entropy.

Now consider a fair six sided die instead of a fair coin. The cast of a die has six outcomes: 1 - 6. The Entropy of this collection of events is:

$E = -\sum ⅙ * \log_2(⅙) = -6 * ⅙ * \log_2(⅙) = 2.585$

This is higher than the Entropy of a fair coin. Thus we see that even among equiprobable event sets, the set with higher number of events has higher Entropy. This ties in with our diversity requirement. A city with 100 equiprobable cuisines should be ranked higher than one with 10 equiprobable cuisines. With the Entropy measure in hand, we are now ready to define the Culinary Diversity Index.

Culinary Diversity Index

Consider a city with cuisines c_1, c_2, c_3, ... c_n. Let the fraction of each cuisine be p_1, p_2, p_3, ... p_n. The *Culinary Diversity Index* (CDI) of the city is defined as:

$$CDI = -\sum p_k \log_2(p_k)$$

The higher the CDI the more diverse the cuisines. This metric gives us a measure to compare different cities in terms of culinary diversity.

The Battle for Culinary Diversity

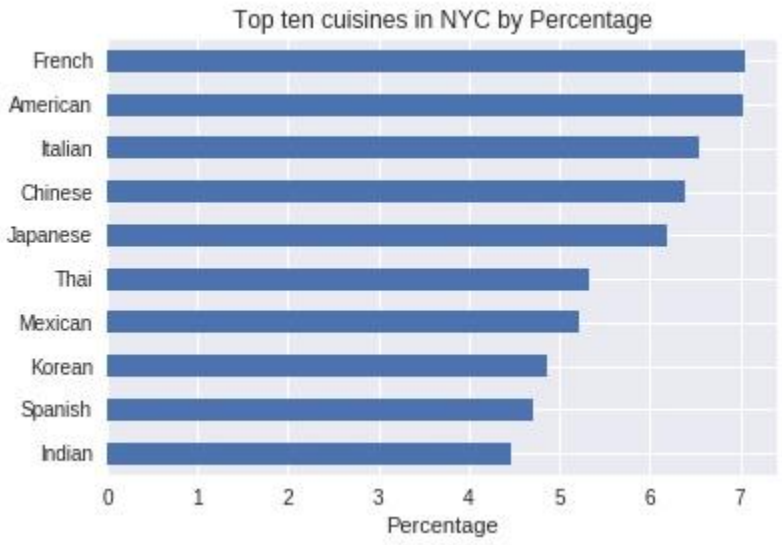
For the battle of Neighborhoods project I compare the culinary diversity of three major cities: New York, San Francisco and Toronto. I plan to calculate the CDI for each of the cities and see which one is more diverse in terms of CDI ranking.

Data

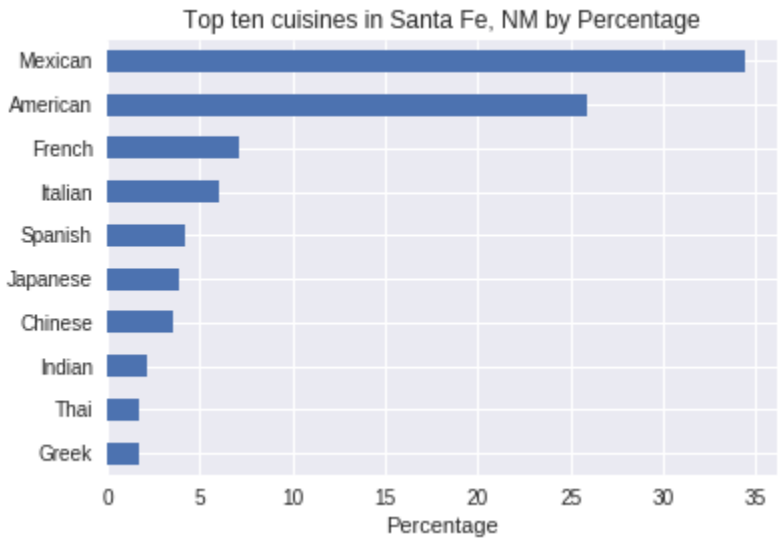
To calculate the CDI for a city, we need a list of cuisines and the number of restaurants serving that cuisine. This data is obtained from Foursquare. The cuisines are obtained from the Food subsection of the venue category hierarchy of Foursquare. Not all cuisines are used. For example, I consider Indian cuisine but do not dive into other sub-categories. The list of cuisines considered and their Foursquare codes will be provided in the Python code.

The data for a given location / cuisine is found by querying the Foursquare places API with the location and cuisine code.

As an example, consider New York City. We download the data from Foursquare with location as New York and by cuisine. The top ten cuisines by percentage are shown below:



To contrast this with a smaller city, let us do the same experiment for Santa Fe, NM, USA. The top ten cuisines by percentage are shown below:



Comparing NYC and Santa Fe, NM we can see that the cuisine in Santa Fe is predominated by Mexican (presumably this also includes New Mexican; Foursquare does not have a separate category for New Mexican) and American (ok, God knows what Foursquare includes here!).

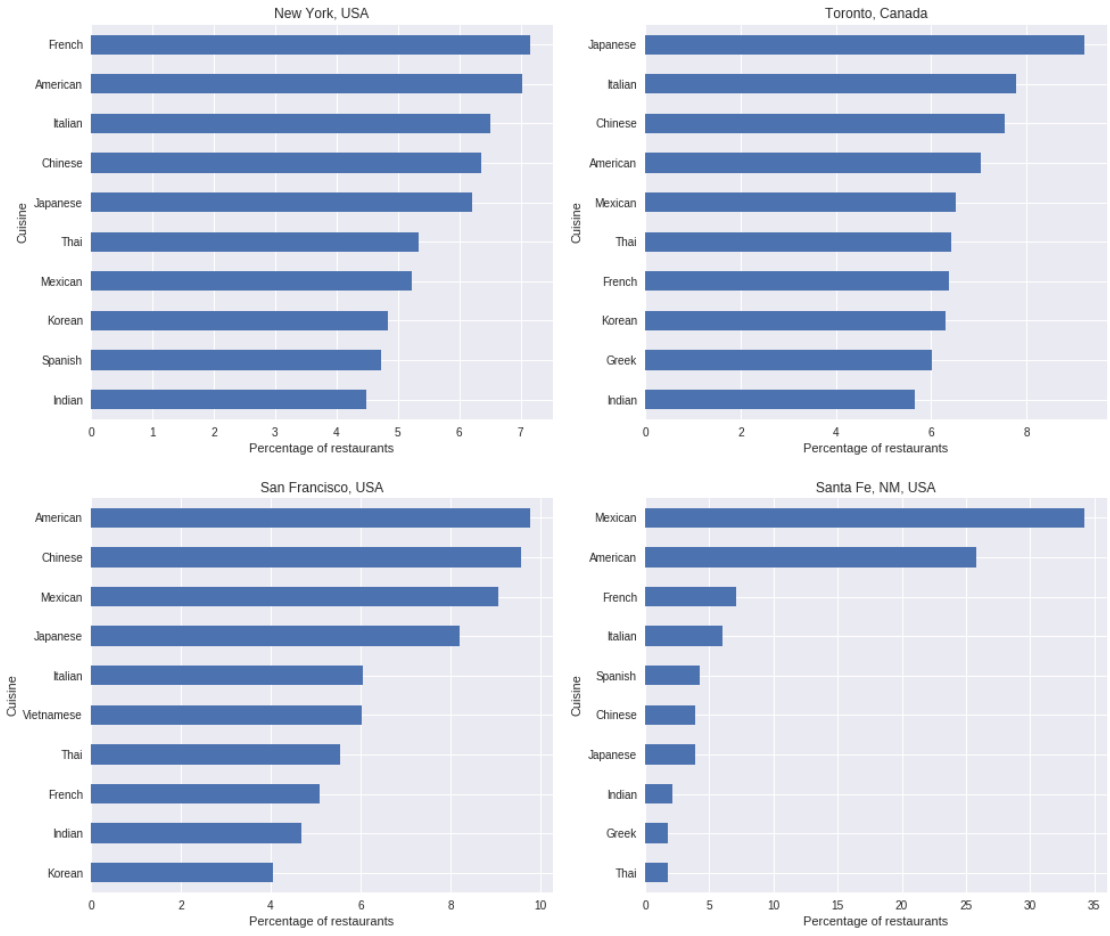
Therefore we would expect that the Culinary Diversity Index for Santa Fe, New Mexico will be smaller than New York. In the next part of the project we will explore if that is tru.

Aside

In Data Science any analysis will only be as good as the data. A good example is the failure of Foursquare to recognize *New Mexican* restaurants from Mexican restaurants (not even a sub-category). They are not the same. As the residents will tell you *'New Mexico: Ain't new, ain't Mexico'*. So if you are looking for Mexican food, be assured Santa Fe, NM has much fewer options than suggested by Foursquare. If you do make it to this part of the world and would like to try *Mexican* food, do visit Los Potrillos.

Exploring CDI

In this part we look at the city of three major cities, New York, USA, San Francisco, USA, and Toronto, Canada. For comparison we also look at a much smaller city Santa Fe, NM USA. As a first sep we collect the data and plot the top ten cuisines by percentage for each city. The results are shown below:



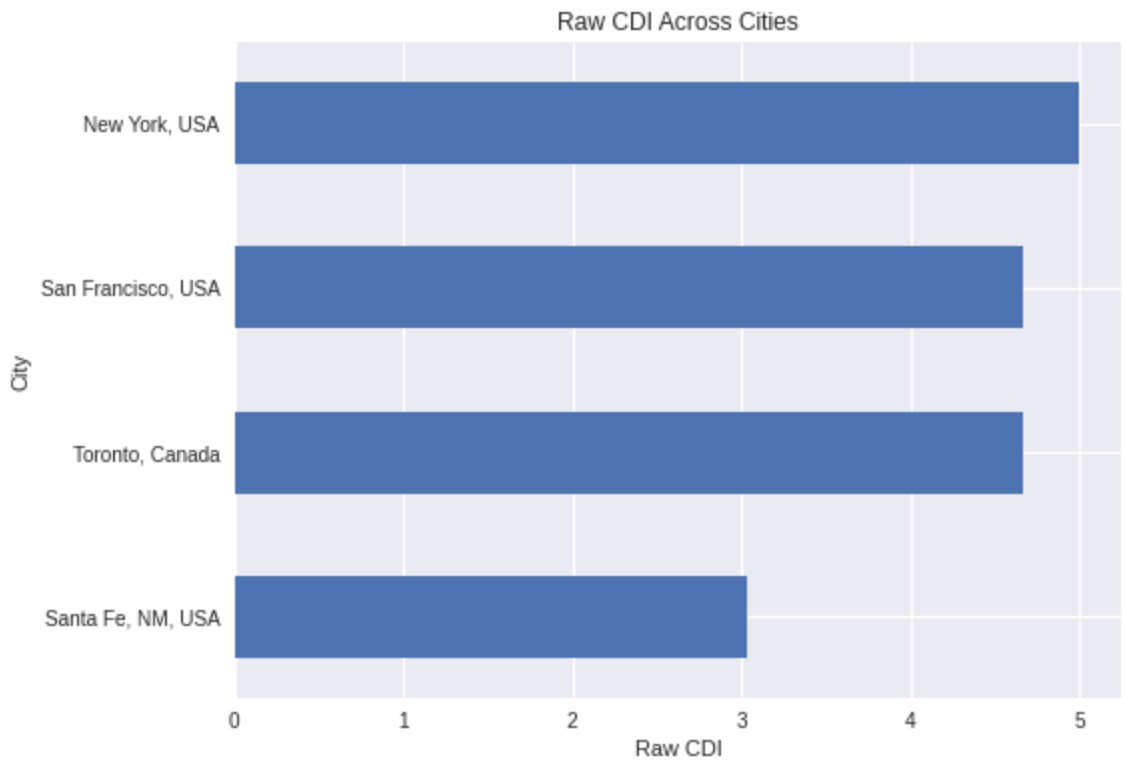
Looking at the data we can see that in the major cities, the predominant cuisines are only 7% - 8% of the total restaurants. However in Santa Fe, NM the top two cuisines account for more than 60% of the restaurants. So we would expect the CDI to be lower for Santa Fe, NM.

Calculating raw CDI

Now we calculate the CDI according to the formula:

$$\text{CDI} = -\sum p_k \log_2(p_k)$$

Where p_k is the fraction of restaurants with cuisine c_k. The values for the various cities are plotted in the following figure:



The raw CDI values are:

City	Raw CDI
New York, USA	4.99
San Francisco, USA	4.66
Toronto, Canada	4.65
Santa Fe, NM, USA	3.03

As seen from the figure and the above table, New York tops the list. San Fancisco and Toronto are very close and Santa Fe, NM is at the bottom. As expected the CDI of Santa Fe, NM is way lower than that of the major cities.

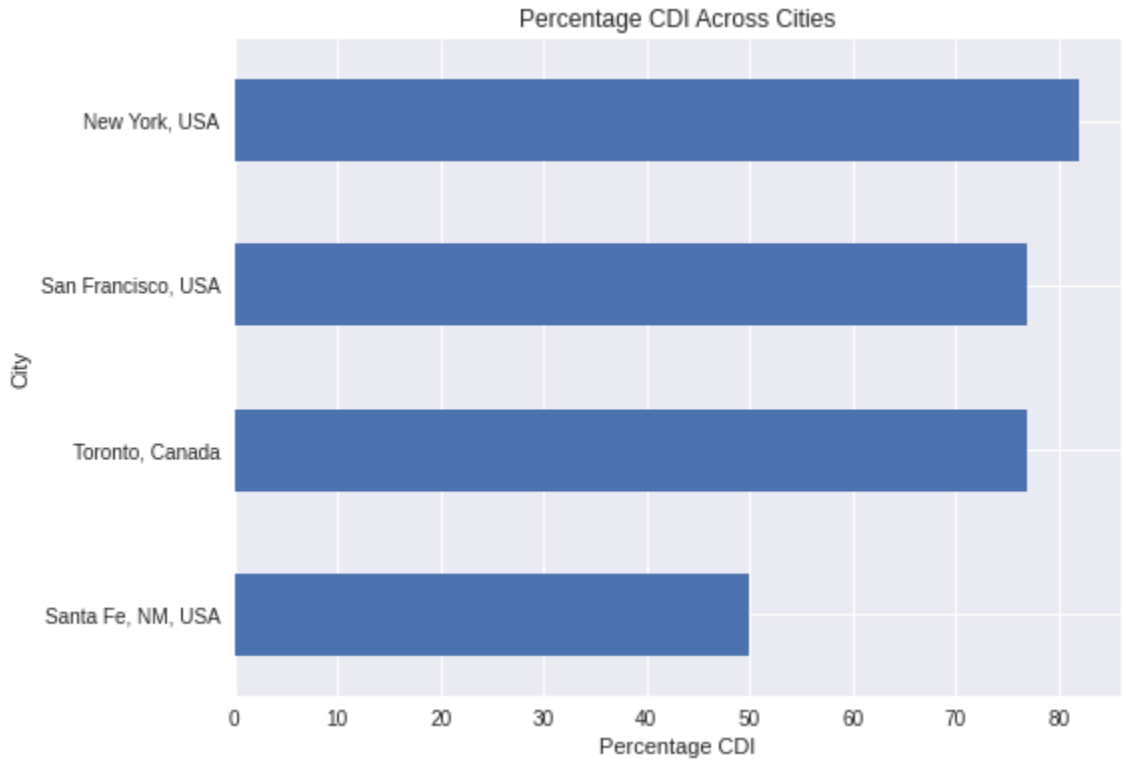
While the raw CDI is enough to rank cities, it lacks a clear interpretability. For example, the CDI of New York is almost 5. However we have no clue what that means in an absolute sense. We need to have a base measure against which we can compare these numbers.

An Interpretable CDI

To provide some interpretability to these results, we ask the question what is the maximum possible CDI value. We have a total of sixty five cuisines. If a city had equal proportions of restaurants for each cuisine, then the CDI of that ideal city would be $-\log_2(1/65) = 0.602$. This is the maximum value of the CDI for this given set of 65 cuisines. Therefore we normalize by this value and get the percentage of this value as the *Percentage CDI.* Thus,

$$\text{Percentage CDI} = 100 * -\sum p_k \log_2(p_k) \; / \; -\log_2(1/65)$$

The plot of Percentage CDI is plotted for the various cities below:



The values are:

City	Percentage CDI
New York, USA	82%
San Francisco, USA	77%
Toronto, Canada	77%
Santa Fe, NM, USA	50%

We see that even though the CDI of New York is high, it is only 82%. So there is some room for improvement since the maximum value is 100%.

One thing to remember is that the percentage CDI is dependent on the length of the cuisines we use. So if Foursquare adds more cuisines to their list, then we need to recalculate the percentage CDI.

Directions for Further Work

This project is just a beginning. An insight into how data science can be used to derive quantitative metrics to quantify qualitative ideas like “food diversity”. Many possible directions exist to further this investigation.

For example, while entropy gave us a CDI there is no reason why a monotonic function of entropy can be used. This would broadly fall under the category of *utility functions.* Widely used in Economics, utility functions measure how much consumers prefer one thing over another. In our case maybe there is no big difference between 755% and 100% CDI but there might be a huge jump in satisfaction if we go from 50% to 55%.

A utility function often has to be determined considering consumer behaviours. Sometimes the parameters have to be obtained through experimentation. The quest never ends.