Similarity

Datathon

1 Definition

Assume we have two vendors/agencies. In each year, both have numerous transactions that can be divided into different categories based on its "mcc description". Let A and B be two sets, representing all the transactions of the two vendors/agencies. Each element in these sets represents a particular "mcc category". We define a metric for similarity as:

$$S = \frac{|A \cap B|}{|A| + |B|}$$

where the $|A|$ means,for each category in A, sums up its transaction amount.

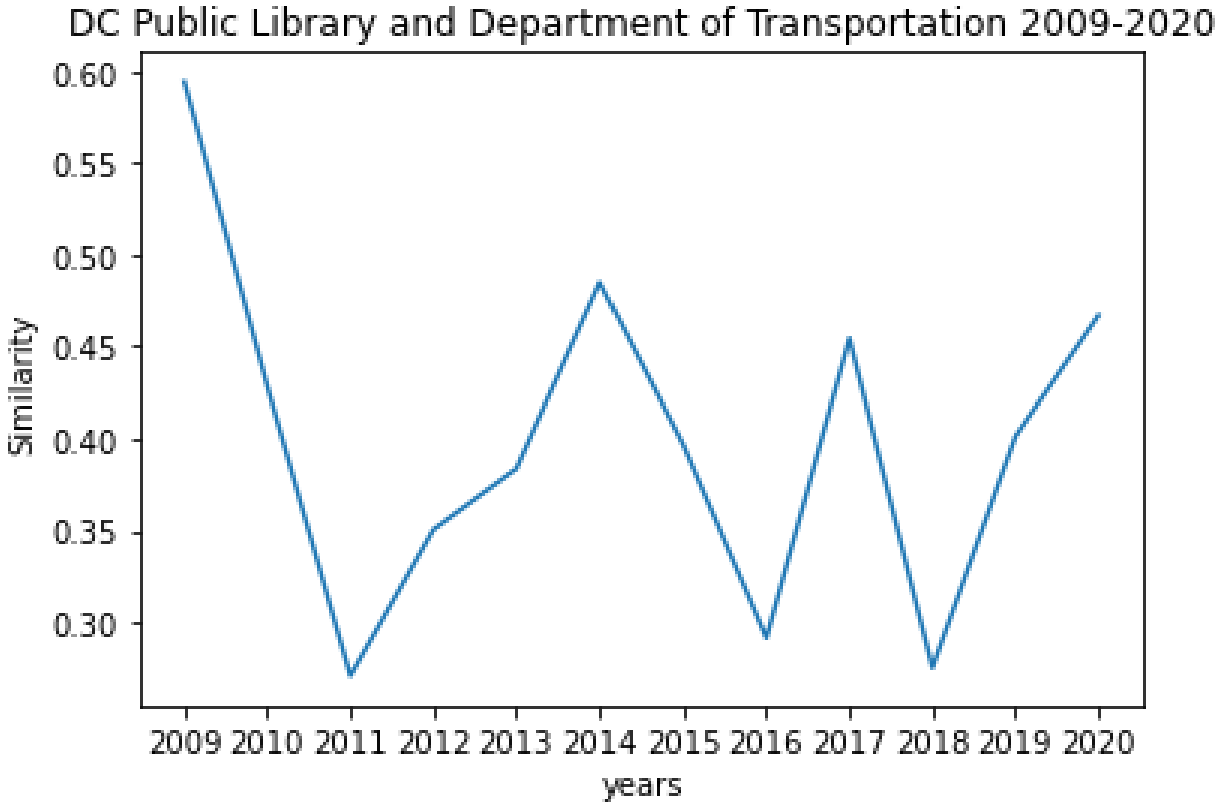
This metric illustrates the similarity in a straightforward way:

If $S = 1$, then the two vendors/agencies have all their transactions in the same mcc category;

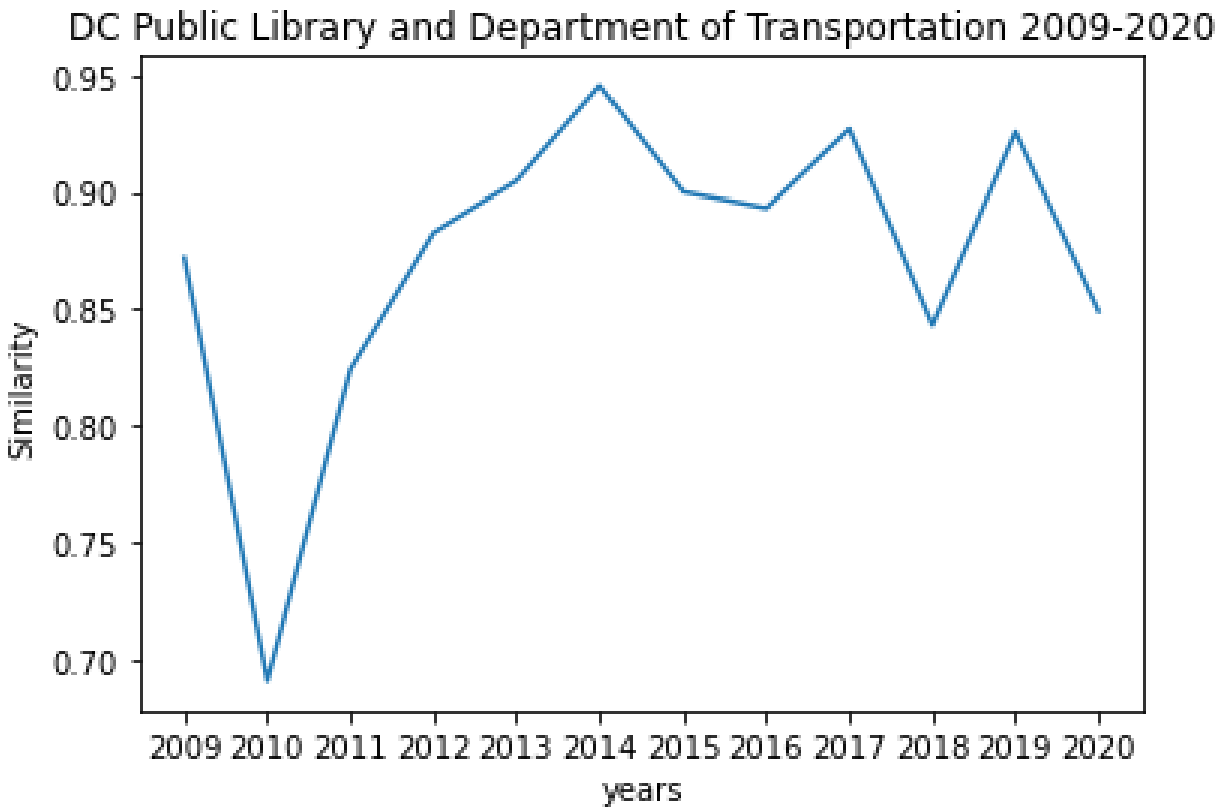
If $S = 0$, then the two vendors/agencies have none of their transactions in the same mcc category;

2 Significance

Imagine the following scenario: we are a vendor who has been doing businesses with the DC Public Library for a while, and now we would like to expand our business. A potential choice could be the department of education. We check their similarity and found:



Their similarity generally falls below 50%. If we check the plot against another agency, Department of transportation



We see this upward trend of increasing similarity. On the first glance education department might seem more related, but their transaction data provide new insights.