Введём функцию $\rho(x, y)$, для которой выполнено три требования:

$$1)\ \rho(x, y) = \rho(y, x);$$

$$2)\ \rho(x, y) \geq 0;$$

$$3)\ \rho(x, y) = 0 \Leftrightarrow x = y.$$

Если a и b - численные атрибуты, то тогда пусть классически ρ(x, y) = |x - y|.

Пусть теперь a и b - атрибуты, состоящие из 1 символа.

Поскольку при указанном хэшировании используются лишь латинские строчные буквы и цифры, то можно их перенумеровать и искать расстояния как между числами.

В соответствие каждой цифре поставим её же саму. Для того, чтобы подчеркнуть различие цифр и букв, поставим букве a в соответствие 21, b - 22, ..., z - 46. Такую функцию назовём ord.

Тогда расстояние фактически находится как $\rho(x, y) = |ord(x) - ord(y)|$.

Пусть теперь x и y - атрибуты, состоящие из n символов, а $x_i, y_i -$ соответствующие $i -$ые символы атрибутов. В таком случае

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (\rho(x_i, y_i)^2)}$$

Если количество символов в атрибутах x и y равны n и k, причём $n \neq k$, то тогда будем полагать, что первые $|n - k|$ символов меньшего из атрибутов нулевые: в таком случае вышеприведённая формула всё ещё останется верной.

Пусть теперь имеется атрибут x и набор атрибутов y длины m. Обозначим $i -$ый элемент y как $y[i]$. В таком случае принимаем, что

$$\rho(x, y) = \frac{\sqrt{\sum_{i=1}^{m} (\rho(x, y[i])^2)}}{m}$$

Для сравнения двух наборов атрибутов x и y длины n воспользуемся следующей формулой:

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (\rho(x[i], y)^2)}$$

Данная функция реализована на языке Python в файле metrics.py.