This document lists some of the available distance metrics and metric spaces. The document is not a theoretically intensive and thorough study of the metrics, but a quick refresher on the topic. A distance metric is a function

$$D : X \times X \to \mathbb{R} \tag{1}$$

that measures the distance between two members (x,y) of a set X. The distance metric is also known as a distance function, a metric and a distance. A set with a metric is called a metric space. A distance metric can be created using a norm if the norm is available for the given set. Therefore, a normed space (a set with a norm) is always a metric space. However, a metric space is not always a normed space. A metric D is a function that satisfies the following conditions:

$$\begin{aligned} D(x,y) &\geq 0 \ (\forall x, y) && (2)\\ D(x,y) &= D(y,x) && (3)\\ D(x,y) &= 0 \iff x = y && (4)\\ D(x,y) + D(y,z) &\geq D(x,z) && (5) \end{aligned}$$

Verbal summary of the previous conditions:

1. The distances can not be negative.
2. The distance is symmetric.
3. The identical members of a set have a zero distance.
4. The triangle inequality. The shortest path between two points is a line.

A norm is a function that defines the magnitude of a vector. A normed vector space is a vector space, which is equipped with a norm. A distance metric can always be defined using a norm, but not all metric spaces are normed spaces. Let

$$N : X \to \mathbb{R} \tag{6}$$

be a norm of a set X. A distance metric using a norm for the set X is defined as

$$D(x,y) = N(x-y) \tag{7}$$

where $x \in X$ and $y \in X$.

For demonstration purposes, this document defines the metrics for a real coordinate space of three dimensions ($\mathbb{R}^3$). A bolded symbol ($\mathbf{x}$ and $\mathbf{y}$) denotes a three-dimensional vector ($\mathbf{x} = \{x_1, x_2, x_3\} \in \mathbb{R}^3$). The vectors $\mathbf{x}$ are the members $\mathbf{x} \in \mathbb{R}^3$ of the set $\mathbb{R}^3$. The following items are a subset of the metrics in a d-dimensional vector space ($\mathbb{R}^d$), which are defined for the $\mathbb{R}^3$:

Manhattan distance

$$D(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{d} |x_i - y_i| = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3| \tag{8}$$

Euclidean distance

$$D(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2} \tag{9}$$

Chebyshev distance

$$D(\mathbf{x},\mathbf{y}) = \forall i \ max(|x_i - y_i|) = max(|x_1 - y_1|, |x_2 - y_2|, |x_3 - y_3|) \tag{10}$$

L_p distance (generalizes the Manhattan, Euclidean and Chebyshev distance)

$$D(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{d} (x_i - y_i)^p)^{\frac{1}{p}} = ((x_1 - y_1)^p + (x_2 - y_2)^p + (x_3 - y_3)^p)^{\frac{1}{p}} \tag{11}$$

Mahalanobis distance

$$D(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})} \tag{12}$$

Cosine distance

$$D(\mathbf{x},\mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \ ||\mathbf{y}||} = \frac{x_1 y_1 + x_2 y_2 + x_3 y_3}{\sqrt{x_1^2 + x_2^2 + x_3^2}\sqrt{y_1^2 + y_2^2 + y_3^2}} \tag{13}$$

Pearson's distance

$$D(\mathbf{x},\mathbf{y}) = 1 - \frac{(\mathbf{x} - \bar{\mathbf{x}})^T (\mathbf{y} - \bar{\mathbf{y}})}{||\mathbf{x}|| \ ||\mathbf{y}||} \tag{14}$$

Canberra distance

$$D(\mathbf{x},\mathbf{y}) = \frac{1}{d} \sum_{i=1}^{d} \frac{|x_i - y_i|}{x_i + y_i} = \frac{1}{d}\frac{|x_1 - y_1|}{x_1 + y_1} + \frac{1}{d}\frac{|x_2 - y_2|}{x_2 + y_2} + \frac{1}{d}\frac{|x_3 - y_3|}{x_3 + y_3} \tag{15}$$

$$\tag{16}$$

The choice of the metric depends on the set X and the problem at hand. For example, consider a task where you need to categorize a collection of documents. The cosine distance does not depend on the magnitude of the vectors because it normalizes the vectors to the unit vectors. The commonly used Euclidean distance measures only the difference of the magnitude of the vectors. Therefore, one could use the cosine distance to categorize the documents because a category for the documents shares a collection of specific words. However, the magnitude of the words (Euclidean distance) might not define the categories as efficiently.