Mathematics for Data Analysis

– Distances –

Matteo Gorgone

University of Messina, Department MIFT

email: mgorgone@unime.it

Why Distance?

At the core of most data analysis tasks and their formulations is a distance. This choice anchors the meaning and the modeling inherent in the patterns found and the algorithms used. However, there are an enormous number of distances to choose from!
Firstly, let us provide the basic notions to define distances.

Let X be a vector space over the field $\mathbb{R}$ with the scalar product

$$\begin{aligned} \cdot : X \times X &\to \mathbb{R} \\ (\mathbf{x}, \mathbf{y}) &\mapsto \mathbf{x} \cdot \mathbf{y} \end{aligned}$$

i.e., X is an Euclidean space, and we can introduce also a vector norm

$$\begin{aligned} \| \cdot \| : X &\to \mathbb{R} \\ \mathbf{x} &\mapsto \|\mathbf{x}\| \end{aligned}$$

Distance

In simple terms, a distance is a numerical measurement of how far apart objects or points are. In general, we can introduce distances by means of similarity measures!

Similarity measures

A similarity measure or similarity metric is a real–valued function that quantifies the similarity between two objects. Although no single definition of a similarity exists, usually such measures are in some sense the inverse of distance metrics (or dissimilarities): they take on large values for similar objects and either zero or a negative value for very dissimilar objects.

Similarity and Dissimilarity

Similarity and Dissimilarity are important because they are used by a number of data mining techniques, such as clustering nearest neighbor classification and anomaly detection. The term proximity is used to refer to either similarity or dissimilarity.

Definition: Similarity

The similarity between two objects is a numeral measure of the degree to which the two objects are alike. Consequently, similarities are higher for pairs of objects that are more similar.

Definition: Dissimilarity

The dissimilarity between two objects is the numerical measure of the degree to which the two objects are different. Dissimilarity is lower for pairs of objects that are more similar.

Remark

Frequently, the term distance is used as a synonym for dissimilarity. Dissimilarities (and also similarities) sometimes fall in the interval $[0, 1]$, but it is also common for them to be in the interval $[0, +\infty[$.

Definition: Proximity measures

Proximity measures are defined to have values in the interval $[0, 1]$.
Both similarity and dissimilarity measures can be transformed to proximity measures!

From similarity to proximity!

If the similarity between objects can range from 1 (not at all similar) to p (completely similar), we can make them fall into the interval $[0,1]$ by using the formula

$$s' = \frac{s - \min_s}{\max_s - \min_s},$$

where s and s' are the original and the new similarity values, respectively, $\min_s$ and $\max_s$ are the minimum and maximum similarity values, respectively.

From dissimilarity to proximity!

Likewise, dissimilarity measures with a finite range can be mapped to the interval $[0,1]$ by using the formula

$$d' = 1 - \frac{d - \min_d}{\max_d - \min_d}.$$

Transformation for proximity measures with infinite range

If the proximity measure originally takes values in the interval $[0,\infty]$, then we usually use the formula

$$d' = \frac{d}{1+d},$$

so the proximity measure falls in the range $[0,1]$.

Similarity measures may also satisfy metric axioms!

Metric

A metric on a vector space X over the field $\mathbb{R}$ is a binary operation

$$\mathrm{d} : X \times X \to \mathbb{R}$$

that, for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in X$, satisfies the following axioms:

(i) $\mathrm{d}(\mathbf{x}, \mathbf{y}) \geq 0$ (non-negativity);

(ii) $\mathrm{d}(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$ (identity);

(iii) $\mathrm{d}(\mathbf{x}, \mathbf{y}) = \mathrm{d}(\mathbf{y}, \mathbf{x})$ (symmetry);

(iv) $\mathrm{d}(\mathbf{x}, \mathbf{y}) \leq \mathrm{d}(\mathbf{x}, \mathbf{z}) + \mathrm{d}(\mathbf{z}, \mathbf{y})$ (triangle inequality).

A metric gives a distance between each couple of points of a set. (X, d) is called metric space.

Definition

- A distance that satisfies (i), (iii), and (iv) (but not necessarily (ii)) is called pseudometric.
- A distance that satisfies (i), (ii), and (iv) (but not necessarily (iii)) is called quasimetric.
- A distance that satisfies (i), (ii), and (iii) (but not necessarily (iv)) is called semimetric.
- A distance that satisfies (i) and (iv) (but not necessarily (ii) and (iii)) is called generalized metric or Lawvere metric.

Remark

In an Euclidean space X, distances can be defined also in terms of vector norms (L_1 norm, Euclidean norm, L_p norm, max norm, etc.), and, in particular, we can introduce an induced distance by a norm.

Induced distance by a vector norm

Let X be an Euclidean distance with a vector norm $\|\cdot\| : X \to \mathbb{R}$. It is possible to introduce an induced distance by the vector norm $\|\cdot\|$ defined as

$$\begin{aligned} &\mathrm{d} : X \times X \to \mathbb{R} \\ &\mathrm{d}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|. \end{aligned}$$

This function satisfies the axioms of distance. In fact:

(i) $\mathrm{d}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| \geq 0$;

(ii) $\mathrm{d}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = 0 \Leftrightarrow \|\mathbf{x}-\mathbf{y}\| = \|\mathbf{0}\| \Leftrightarrow \mathbf{x}=\mathbf{y}$;

(iii) $\mathrm{d}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \|(-1)(\mathbf{y}-\mathbf{x}\| = |-1|\|\mathbf{y}-\mathbf{x}\| = \|\mathbf{y}-\mathbf{x}\| = \mathrm{d}(\mathbf{y},\mathbf{x})$;

(iv) $\mathrm{d}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \|\mathbf{x}-\mathbf{z}+\mathbf{z}-\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{z}\| + \|\mathbf{z}-\mathbf{y}\| = \mathrm{d}(\mathbf{x},\mathbf{z}) + \mathrm{d}(\mathbf{z},\mathbf{y})$.

Open ball

For any point $\mathbf{x}$ in a metric space X and any real number $r > 0$, the open ball of radius r around $\mathbf{x}$ is defined to be the set of points that are at most distance r from $\mathbf{x}$:

$$\mathcal{B}_r(\mathbf{x}) = \{\mathbf{y} \in X \ : \ \mathrm{d}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| < r\}.$$

This is a natural way define a set of points that are relatively close to $\mathbf{x}$. Therefore, a set $Y \subseteq X$ is called a neighborhood of $\mathbf{x}$ (*i.e.*, it contains all points close enough to $\mathbf{x}$) if it contains an open ball of radius r around $\mathbf{x}$ for some $r > 0$.

Example: discrete metric

$$\mathrm{d}(x,y)=\begin{cases}0 & \text{if} \quad x=y\\ 1 & \text{otherwise}\end{cases}$$

Example: quasimetric

$$\mathrm{d}(x,y)=\begin{cases}x-y & \text{if} \quad x\geq y\\ 1 & \text{otherwise}\end{cases}$$

Łukaszyk-Karmowski distance

The Łukaszyk-Karmowski distance is a function defining a distance between two random variables or two random vectors that satisfy the following axioms:

(i) $\mathrm{d}(\mathbf{x},\mathbf{y})>0$ (positivity);

(ii) $\mathrm{d}(\mathbf{x},\mathbf{y})=\mathrm{d}(\mathbf{y},\mathbf{x})$ (symmetry);

(iii) $\mathrm{d}(\mathbf{x},\mathbf{y})\leq \mathrm{d}(\mathbf{x},\mathbf{z})+\mathrm{d}(\mathbf{z},\mathbf{y})$ (triangle inequality).

The Łukaszyk-Karmowski distance is not a metric.

Use of different similarity measures

Different types of similarity measures exist for various types of objects, depending on the objects being compared. For each type of object there are various similarity measurement formulas.

L_p (Minkowski) distance

Let $\mathbf{x} = \{x_1, \dots, x_N\}$, $\mathbf{y} = \{y_1, \dots, y_N\} \in \mathbb{R}^N$.
An L_p distance, for any parameter $p \in [1, \infty)$, is defined as

$$\mathrm{d}_p(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p = \left(\sum_{i=1}^{N} |x_i - y_i|^p \right)^{1/p}.$$

Any L_p distance is translation invariant and is also a metric.

L_1 distance

$$\mathrm{d}_1(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^{N} |x_i - y_i|.$$

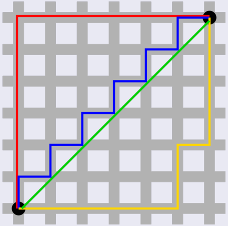
The L_1 distance between two points is the sum of the absolute differences of their Cartesian coordinates.
This is also known as the Manhattan distance since it refers to the rectilinear street layout on the island of Manhattan, where the shortest path a taxi travels between two points is the sum of the absolute values of distances that it travels on avenues and on streets; you must stay on the streets and cannot cut through buildings!
Manhattan distance depends on the rotation of the coordinate system, but does not depend on its reflection about a coordinate axis or its translation.

L_2 (Euclidean) distance

$$\mathrm{d}_2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}.$$

The L_2 distance is easy interpreted as the straight-line distance between two points or vectors, since if you draw a line segment between two points, its length measures the Euclidean distance. It is also the only L_p distance that is invariant with respect to the rotation of the coordinate system (it is also translation invariant).

Manhattan distance vs. Euclidean distance



By using the Manhattan distance, the red, yellow, blue, and green paths all have the same shortest path length of 12, whereas, by means of Euclidean distance, the green line has length $6\sqrt{2} \approx 8.49$ and is the unique shortest path, while the other paths have the longer length of 12.

L_∞ (Chebyshev) distance

$$\mathrm{d}_\infty(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_\infty = \max_{1\le i\le n} |x_i - y_i|.$$

Chebyshev distance is the distance between two vectors is the greatest of their differences along any coordinate dimension. It is

$$\mathrm{d}_\infty(\mathbf{x},\mathbf{y}) = \lim_{p\to\infty} \mathrm{d}_p(\mathbf{x},\mathbf{y}) = \lim_{p\to\infty} \left(\sum_{i=1}^{N} |x_i - y_i|^p\right)^{1/p}.$$

It is also known as chessboard distance, since in the game of chess the minimum number of moves needed by a king to go from one square on a chessboard to another equals the Chebyshev distance between the centers of the squares, if the squares have side length one, as represented in 2-D spatial coordinates with axes aligned to the edges of the board.

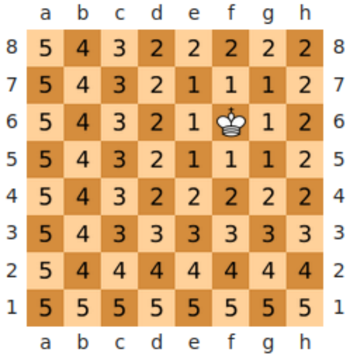
Example

In two dimensions, *i.e. plane geometry*, the Chebyshev distance betweeen two points $\mathbf{x} = (x_1, y_1)$ and $\mathbf{y} = (x_2, y_2)$ is

$$\mathrm{d}_\infty(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_\infty = \max\left(|x_2 - x_1|, |y_2 - y_1|\right).$$

Under this metric, an open ball of radius r, which is the set of points with Chebyshev distance r from a center point, is a square whose sides have the length $2r$ and are parallel to the coordinate axes.

Chebyshev distance



	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	♔	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

The king can move diagonally, so that the jumps to cover the smaller distance parallel to a rank or column is effectively absorbed into the jumps covering the larger. Above are the Chebyshev distances of each square from the square f6. For example, the Chebyshev distance between f6 and e2 equals 4.

L_0 distance

$$\mathrm{d}_0(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_0 = N - \sum_{i=1}^{N} \mathbb{I}(x_i, y_i),$$

where

$$\mathbb{I}(x_i, y_i) = \begin{cases} 1 & \text{if} \quad x_i = y_i \\ 0 & \text{if} \quad x_i \neq y_i \end{cases}$$

When each coordinate x_i or y_i is either 0 or 1, then d_0 is known as the Hamming distance.

About Hamming distance

The Hamming distance between two strings of equal length is the number of positions at which the corresponding characters are different.

Example 1

Consider the vectors $\mathbf{x} = \{1, 0, 1, 1, 1, 0, 1\}$ and $\mathbf{y} = \{1, 0, 0, 1, 0, 0, 1\}$, then:

$$\mathrm{d}_0(\mathbf{x},\mathbf{y}) = 7 - (1 + 1 + 0 + 1 + 0 + 1 + 1) = 2.$$

Example 2

Consider the vectors $\mathbf{x} = \{2, 1, 4, 3, 8, 9, 6\}$ and $\mathbf{y} = \{2, 2, 3, 3, 7, 9, 6\}$, then:

$$\mathrm{d}_0(\mathbf{x},\mathbf{y}) = 7 - (1 + 0 + 0 + 1 + 0 + 1 + 1) = 3.$$

Mahalanobis distance

An extension of the L_2 distance is the Mahalanobis distance defined for two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ and a $N \times N$ matrix M as

$$\mathrm{d}_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T M(\mathbf{x} - \mathbf{y})}.$$

Properties

- If $M = I$ (the identity matrix) $\Rightarrow \mathrm{d}_M = \mathrm{d}_2$.
- If M is a diagonal matrix $\Rightarrow \mathrm{d}_M$ can be interpreted as skewing the Euclidean space (shrink some coordinates, and expanding others) based on the matrix M.
- If M is not a diagonal matrix $\Rightarrow$ the skewing of the Euclidean space still holds, but the skew is not aligned with the coordinate axis.
- The Mahalanobis distance is a measure of the distance between a point and a probability distribution. This distance is zero when the point is at the mean of the distribution, and grows as the point moves away from the mean along each principal component axis. If the axes are scaled to have unit variance, then the Mahalanobis distance corresponds to standard Euclidean distance. Thus, the Mahalanobis distance is unitless, scale-invariant, and takes into account the correlations of the data set.

Cosine distance

In an Euclidean space with an orthonormal basis, the cosine distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ is defined as

$$\mathrm{d}_{\cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|} = 1 - \frac{\sum_{i=1}^{N} x_i y_i}{\|\mathbf{x}\| \, \|\mathbf{y}\|} = 1 - \cos(\theta),$$

where θ is the angle between vectors $\mathbf{x}$ and $\mathbf{y}$, along with the relation

$$\mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \, \|\mathbf{y}\| \cos(\theta).$$

Properties

- $\mathrm{d}_{\cos}(\mathbf{x}, \mathbf{y}) \in [0, 2]$.
- The cosine distance is not a metric since the identity axiom and the triangle inequality don't hold.
- The cosine distance depends only on direction of vectors. This is useful when a vector of objects represent datasets of different sizes and we want to compare how similar are those distributions, but not their size.
- If the cosine distance is defined only with respect to normalized vectors, *i.e.* $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, then

$$\mathrm{d}_{\cos}(\mathbf{x}, \mathbf{y}) = 1 - \mathbf{x} \cdot \mathbf{y}.$$

Angular distance

In an Euclidean space with an orthonormal basis, the angular distance between two normalized vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, *i.e.* $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, is defined as

$$\mathrm{d}_{\mathrm{ang}}(\mathbf{x}, \mathbf{y}) = \arccos(\mathbf{x} \cdot \mathbf{y}) = \arccos(\cos(\theta)) = \theta.$$

Properties

- $\mathrm{d}_{\mathrm{ang}}(\mathbf{x}, \mathbf{y}) \in [0, \pi]$.
- The angular distance is a metric.
- For all normalized vectors $\mathbf{u}, \mathbf{v}, \mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ if $\mathrm{d}_{\mathrm{cos}}(\mathbf{u}, \mathbf{v}) < \mathrm{d}_{\mathrm{cos}}(\mathbf{x}, \mathbf{y}) \Rightarrow \mathrm{d}_{\mathrm{ang}}(\mathbf{u}, \mathbf{v}) < \mathrm{d}_{\mathrm{ang}}(\mathbf{x}, \mathbf{y})$.
- The angular distance measures only the angle between two vectors.

Distances for Sets and Strings

Now we introduce distances used to understand the relationship between text data: strings of words or characters. There are other techniques taking into account the semantic and the structural properties of text. We focus here on the distances owning simple mathematical connections and as a result are often more scalable and flexible.

Sets: notation

A set A is a collection of elements $\{a_1, \dots, a_N\}$. We have that:

- the cardinality (counts how many elements are in A) of A is denoted by $|A|$;
- the ordering of elements does not matter, *i.e.*, $\{a_1, a_2\} = \{a_2, a_1\}$;
- the multiplicity is ignored, *i.e.*, $\{a_1, a_1, a_2\} = \{a_1, a_2\}$; otherwise, we have a multiset.

Operation between sets: notation

- The intersection between two sets A and B is denoted by $A \cap B$.
- The union between two sets A and B is denoted by $A \cup B$.
- The difference of two sets A and B is denoted by $A \backslash B$.
- The symmetric distance between two sets A and B is denoted by $A \Delta B$ and it is $A \Delta B = (A \backslash B) \cup (B \backslash A) = (A \cup B) \backslash (A \cap B)$.
- Given a domain Ω, the complement of a set A is defined as $\overline{A} = \Omega \backslash A$.

Multiset

A multiset (or bag) is a generalization of the concept of a set that, unlike a set, allows for multiple occurrences for each one of its elements.

Definition

- The number of occurrences given for each element is called the multiplicity of that element in the multiset.
- The cardinality of a multiset is the sum of the multiplicities of all its elements.
- The ordering of elements does not matter in discriminating multisets.

Example

- The set $X = \{a, b\}$ is a multiset where the elements a and b both have multiplicity equal to 1, and cardinality $|X| = 2$.
- The set $X = \{a, a, b, b, b, c\}$ is a multiset where the multiplicity of a is 2, the multiplicity of b is 3, and the multiplicity of c is 1; the cardinality is $X = 2 + 3 + 1 = 6$.
- $\{a, a, b\}$ and $\{a, b, a\}$ denote the same multiset.

Remark

The usual operations between sets may be extended to multisets.

Diameter of a set

Let A be a subset of a metric space (M, d). The diameter of A, denoted by $\operatorname{diam}(A)$, of a metric space (M, d) is the upper bound of the set of all distances between pairs of points in X, *i.e.*,

$$\operatorname{diam}(A) = \sup_{\mathbf{x},\mathbf{y}\in A} \mathrm{d}(\mathbf{x}, \mathbf{y}).$$

Distance of a point from a set

Let A be a subset of a metric space (M, d). The distance of a point $\mathbf{x}$ from A is defined as the infimum distance between the point and the various points of A, *i.e.*,

$$\mathrm{d}(\mathbf{x}, A) = \inf_{\mathbf{y}\in A} \mathrm{d}(\mathbf{x}, \mathbf{y}).$$

It is the distance between the point $\mathbf{x}$ and the closest point $\mathbf{y}$ in the set A.

Example

Let $\mathrm{d}(x, y) = |x - y|$. Compute the distance between the point $x = 1$ and the set $A = \{3, 6\}$.
Solution.

$$\mathrm{d}(1, \{3, 6\}) = \inf_{y \in A} \mathrm{d}(1, y) = \inf \{\mathrm{d}(1, 3), \mathrm{d}(1, 6)\} = \inf \{2, 5\} = 2.$$

Example

Let $\mathrm{d}(x, y) = |x - y|$. Compute the distance between the point $x = 7$ and the set $A = \{3, 6\}$.
Solution.

$$\mathrm{d}(7, \{3, 6\}) = \inf_{y \in A} \mathrm{d}(7, y) = \inf \{\mathrm{d}(7, 3), \mathrm{d}(7, 6)\} = \inf \{4, 1\} = 1.$$

Distances between sets

There are different methods to measure the distance between sets that consist of more than one point:

- measure the distance between representative points of two objects, such as the center of mass;
- measure the distance between the closest points of the two objects; this idea can be used to define the distance between two subsets of a metric space. The distance between sets A and B is the infimum of the distances between any two of their respective points:

  $$\mathrm{d}(A, B) = \inf_{\mathbf{x}\in A,\, \mathbf{y}\in B} \mathrm{d}(\mathbf{x}, \mathbf{y}).$$

  This does not define a metric on the set of such subsets: the distance between overlapping sets is zero, and this distance does not satisfy the triangle inequality for any metric space with two or more points (consider the triple of sets consisting of two distinct singletons and their union);
- use the Hausdorff distance between two subsets of a metric space that measures how far they are from perfectly overlapping, *i.e.*, how far are they from each other.

Example

Let $\mathrm{d}(x, y) = |x - y|$. Compute the distance between the sets $A = \{1, 7\}$ and $B = \{3, 6\}$.
Solution.

$$\mathrm{d}(\{1, 7\}, \{3, 6\}) = \inf\{\mathrm{d}(1, 3), \mathrm{d}(1, 6), \mathrm{d}(7, 3), \mathrm{d}(7, 6)\} = \inf\{2, 5, 4, 1\} = 1.$$

Hausdorff distance

The Hausdorff distance between two non-empty subsets X and Y of a metric space (M, d) is either the distance from X to the farthest point of Y, or the distance from Y to the farthest point of X, whichever is larger. Informally, two sets are close in the Hausdorff distance if every point of either set is close to some point of the other set. In other words, it is the greatest of all the distances from a point in one set to the closest point in the other set.

We define their Hausdorff distance $\mathrm{d}_{\mathrm{H}}(X, Y)$ by

$$\mathrm{d}_{\mathrm{H}}(X, Y) = \max\{\mathrm{d}(X, Y), \mathrm{d}(Y, X)\} = \max\left\{ \sup_{\mathbf{x}\in X} \mathrm{d}(\mathbf{x}, Y), \sup_{\mathbf{y}\in Y} \mathrm{d}(\mathbf{y}, X) \right\},$$

where sup represents the supremum, inf the infimum, $\mathrm{d}(\mathbf{x}, Y) = \inf_{\mathbf{y}\in Y} \mathrm{d}(\mathbf{x}, \mathbf{y})$ quantifies the distance from a point $\mathbf{x} \in X$ to the subset Y, and $\mathrm{d}(X, Y) = \sup_{\mathbf{x}\in X} \mathrm{d}(\mathbf{x}, Y)$.

Properties

- $\mathrm{d}_{\mathrm{H}}(X, Y)$ may be infinite. If both X and Y are bounded, then $\mathrm{d}_{\mathrm{H}}(X, Y)$ is guaranteed to be finite.
- $\mathrm{d}_{\mathrm{H}}(X, Y) = 0$ if and only if X and Y have the same closure.
- $\mathrm{d}(\mathbf{x}, Y) \leq \mathrm{d}(\mathbf{x}, Z) + \mathrm{d}_{\mathrm{H}}(Y, Z)$, $\forall \mathbf{x} \in M$ and for any non-empty subsets $Y, Z \subseteq M$.
- $|\mathrm{diam}(Y) - \mathrm{diam}(X)| \leq 2\mathrm{d}_{\mathrm{H}}(X, Y)$.

Example

Let $\mathrm{d}(x,y) = |x-y|$. Compute Hausdorff distance $\mathrm{d}_{\mathrm{H}}(X,Y)$ between the sets $X = \{1,7\}$ and $Y = \{3,6\}$. Remember that

$$\mathrm{d}(X,Y) = \sup_{\mathbf{x}\in X} \mathrm{d}(\mathbf{x},Y) = \sup_{\mathbf{x}\in X} \left\{ \inf_{\mathbf{y}\in Y} \mathrm{d}(\mathbf{x},\mathbf{y}) \right\}.$$

Solution.

$$\begin{aligned}
\mathrm{d}(X,Y) = \mathrm{d}(\{1,7\},\{3,6\}) &= \sup_{x\in\{1,7\}} \mathrm{d}(x,\{3,6\}) = \sup\{\mathrm{d}(1,\{3,6\}),\mathrm{d}(7,\{3,6\})\} = \\
&= \sup\{\inf\{\mathrm{d}(1,3),\mathrm{d}(1,6)\},\inf\{\mathrm{d}(7,3),\mathrm{d}(7,6)\}\} = \\
&= \sup\{\inf\{2,5\},\inf\{4,1\}\} = \sup\{2,1\} = 2. \\
\mathrm{d}(Y,X) = \mathrm{d}(\{3,6\},\{1,7\}) &= \sup_{x\in\{3,6\}} \mathrm{d}(x,\{1,7\}) = \sup\{\mathrm{d}(3,\{1,7\}),\mathrm{d}(6,\{1,7\})\} = \\
&= \sup\{\inf\{\mathrm{d}(3,1),\mathrm{d}(3,7)\},\inf\{\mathrm{d}(6,1),\mathrm{d}(6,7)\}\} = \\
&= \sup\{\inf\{2,4\},\inf\{5,1\}\} = \sup\{2,1\} = 2.
\end{aligned}$$

Then:

$$\mathrm{d}_{\mathrm{H}}(X,Y) = \max\{\mathrm{d}(X,Y),\mathrm{d}(Y,X)\} = \max\left\{ \sup_{\mathbf{x}\in X} \mathrm{d}(\mathbf{x},Y),\ \sup_{\mathbf{y}\in Y} \mathrm{d}(\mathbf{y},X) \right\} = \max\{2,2\} = 2.$$

Example

Let $\mathrm{d}(x,y)=|x-y|$. Compute Hausdorff distance $\mathrm{d_H}(X,Y)$ between the sets $X=\{1,3,6,7\}$ and $Y=\{3,6\}$. Remember that

$$\mathrm{d}(X,Y)=\sup_{\mathbf{x}\in X}\mathrm{d}(\mathbf{x},Y)=\sup_{\mathbf{x}\in X}\left\{\inf_{\mathbf{y}\in Y}\mathrm{d}(\mathbf{x},\mathbf{y})\right\}.$$

Solution.

$$\mathrm{d}(\{1,3,6,7\},\{3,6\})=\sup_{x\in\{1,3,6,7\}}\mathrm{d}(x,\{3,6\})=\sup\{\mathrm{d}(1,\{3,6\}),\mathrm{d}(3,\{3,6\}),\mathrm{d}(6,\{3,6\}),\mathrm{d}(7,\{3,6\})\}=$$

$$=\sup\{\inf\{\mathrm{d}(1,3),\mathrm{d}(1,6)\},\inf\{\mathrm{d}(3,3),\mathrm{d}(3,6)\},\inf\{\mathrm{d}(6,3),\mathrm{d}(6,6)\},\inf\{\mathrm{d}(7,3),\mathrm{d}(7,6)\}\}=$$

$$=\sup\{\inf\{2,5\},\inf\{0,3\},\inf\{3,0\},\inf\{4,1\}\}=\sup\{2,0,0,1\}=2.$$

$$\mathrm{d}(\{3,6\},\{1,3,6,7\})=\sup_{x\in\{3,6\}}\mathrm{d}(x,\{1,3,6,7\})=\sup\{\mathrm{d}(3,\{1,3,6,7\}),\mathrm{d}(6,\{1,3,6,7\})\}=$$

$$=\sup\{\inf\{\mathrm{d}(3,1),\mathrm{d}(3,3),\mathrm{d}(3,6),\mathrm{d}(3,7)\},\inf\{\mathrm{d}(6,1),\mathrm{d}(6,3),\mathrm{d}(6,6),\mathrm{d}(6,7)\}\}=$$

$$=\sup\{\inf\{2,0,3,4\},\inf\{5,3,0,1\}\}=\sup\{0,0\}=0.$$

Then:

$$\mathrm{d_H}(X,Y)=\max\{\mathrm{d}(X,Y),\mathrm{d}(Y,X)\}=\max\left\{\sup_{\mathbf{x}\in X}\mathrm{d}(\mathbf{x},Y),\ \sup_{\mathbf{y}\in Y}\mathrm{d}(\mathbf{y},X)\right\}=\max\{2,0\}=2.$$

Jaccard distance

The Jaccard distance between two sets A and B is defined as

$$\mathrm{d}_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

The Jaccard distance measures the dissimilarity between two sets based on the number of items that are present in both sets relative to the total number of items.

Properties

- $\mathrm{d}_J(A, B) \in [0, 1]$;
- $\mathrm{d}_J(A, B)$ is a metric;
- $\mathrm{d}_J(A, B)$ is invariant with respect to the size of the sets;
- $\mathrm{d}_J(A, B)$ depends only on the sets A and B, not on the domain Ω they come from.

Jaccard distance: example 1

Consider the sets $A = \{0, 1, 2, 5, 6\}$ and $B = \{0, 2, 3, 5, 7, 9\}$, then:

$$\mathrm{d}_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{|\{0, 2, 5\}|}{|\{0, 1, 2, 3, 5, 6, 7, 9\}|} = 1 - \frac{3}{8} = 0.625$$

Jaccard distance: example 2

Let $A = \{0, 1, 2, 5, 6, 9\}$ and $B = \{0, 2, 3, 5, 7, 9\}$, then:

$$\mathrm{d}_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{|\{0, 2, 5, 9\}|}{|\{0, 1, 2, 3, 5, 6, 7, 9\}|} = 1 - \frac{4}{8} = 0.5$$

Jaccard distance: example 3

Let $A = \{0, 1, 2, 5, 6, 8\}$ and $B = \{0, 2, 3, 5, 7, 9\}$, then:

$$\mathrm{d}_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{|\{0, 2, 5\}|}{|\{0, 1, 2, 3, 5, 6, 7, 8, 9\}|} = 1 - \frac{3}{9} = 0.666$$

Lee distance

The Lee distance is a distance between two strings $\mathbf{x} = \{x_1, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, \ldots, y_n\}$ of equal length n over the q-ary alphabet $\{0, 1, \ldots, q-1\}$ of size $q \geq 2$.
It is a metric defined as

$$\mathrm{d}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \min\left(|x_i - y_i|,\, q - |x_i - y_i|\right).$$

If $q = 2$ or $q = 3$ the Lee distance coincides with the Hamming distance, because both distances are 0 for two single equal components and 1 for two single non-equal components.

Example

If $q = 6$, the Lee distance between the strings $x = 3140$ and $y = 2543$ is

$$\mathrm{d}(x, y) = \sum_{i=1}^{4} \min\left(|x_i - y_i|,\, 6 - |x_i - y_i|\right) = \min\left(|3-2|,\, 6 - |3-2|\right) + \min\left(|1-5|,\, 6 - |1-5|\right) +$$
$$+ \min\left(|4-4|,\, 6 - |4-4|\right) + \min\left(|0-3|,\, 6 - |0-3|\right) = 1 + 2 + 0 + 3 = 6.$$

Edit distance

Edit distance is a way of quantifying how dissimilar two strings are by counting the minimum number of operations required to transform one string into another one. Edit distances find applications in natural language processing to correct spelling mistakes, in bioinformatics to quantify the similarity of DNA sequences, in computational biology, in coding theory, and approximate string matching, where the objective is to find matches for short strings in many longer texts.

Edit distance: definition

Let Σ be a set, that is an alphabet of possible characters, bytes, etc. The edit distance between two strings $a, b \in \Sigma$ is defined as

$$\mathrm{d}_{\mathrm{ed}}(a, b) = \text{minimum number of edit operations to make } a = b,$$

where different types of edit operations can be used on a single character:

- insertion;
- deletion;
- substitution;
- transposition.

There are different types of edit distance which allow different sets of string operations.

Edit distance: properties

- Edit distance with non-negative cost satisfies the axioms of a metric, when the following conditions hold:
    - every edit operation has positive cost;
    - for every operation, there is an inverse operation with equal cost.

  With these properties, the metric axioms are satisfied as follows:

  1. $\mathrm{d}_{\mathrm{ed}}(a, b) \geq 0$, *i.e.*, the number of operations is always non negative;
  2. $\mathrm{d}_{\mathrm{ed}}(a, b) = 0$ if and only if $a = b$, *i.e.*, there are no operations if two strings are equal;
  3. $\mathrm{d}_{\mathrm{ed}}(a, b) = \mathrm{d}_{\mathrm{ed}}(b, a)$, *i.e.*, the operations can be reversed;
  4. if $c \in \Sigma$, it is $\mathrm{d}_{\mathrm{ed}}(a, b) \leq \mathrm{d}_{\mathrm{ed}}(a, c) + \mathrm{d}_{\mathrm{ed}}(c, b)$;
- edit distance is not good for large documents;
- edit distance requires quadratic time dynamic programming in the worst case to find the smallest set of edits;
- it is possible that removing one sentence a large edit distance without changing meaning is obtained.

Longest Common Subsequence (LCS) distance

The Longest Common Subsequence (LCS) distance allows insertion and deletion.

LCS distance: example

Consider two strings $a = \texttt{kitten}$ and $b = \texttt{sitting}$.
We have:

1. $\texttt{kitten} \longrightarrow \texttt{itten}$ (delete "k");
2. $\texttt{itten} \longrightarrow \texttt{sitten}$ (insert "s");
3. $\texttt{sitten} \longrightarrow \texttt{sittn}$ (delete "e");
4. $\texttt{sittn} \longrightarrow \texttt{sittin}$ (insert "i");
5. $\texttt{sittin} \longrightarrow \texttt{sitting}$ (insert "g").

Then,

$$\mathrm{d}_{\mathrm{ed}}(a, b) = 5.$$

Property

An upper bound for LCS distance is the sum of lengths of a pair of strings.

Levenshtein distance

The Levenshtein distance allows deletion, insertion and substitution.

Levenshtein distance: example

Consider two strings $a = \texttt{kitten}$ and $b = \texttt{sitting}$.
We have:

1. $\texttt{kitten} \longrightarrow \texttt{sitten}$ (substitute “k” by “s”);
2. $\texttt{sitten} \longrightarrow \texttt{sittin}$ (substitute “e” by “i”);
3. $\texttt{sittin} \longrightarrow \texttt{sitting}$ (insert “g”).

Then,

$$\mathrm{d}_{\mathrm{ed}}(a, b) = 3.$$

Properties

- LCS distance is an upper bound on Levenshtein distance.
- A lower bound for Levenshtein distance is the difference of the lengths of the two strings.
- An upper bound for Levenshtein distance is the length of the longer string.
- The Levenshtein distance between two strings is not greater than the sum of their Levenshtein distances from a third string (triangle inequality).

Hamming distance

The Hamming distance allows only substitution, hence, it only applies to strings of the equal length. The Hamming distance between two strings of equal length is the minimum number of substitutions required to change a string into the other one, or the minimum number of errors that could have transformed one string into the other.

Hamming weight

The Hamming weight of a string of length k is its Hamming distance from the string consisting of k zeros, *i.e.*, it is the number non-zero elements ina string. For example, the Hamming weight of the string 11101 is 4, *i.e.*, for a binary strings the Hamming weight is just the number of 1.

Hamming distance: properties

- For a fixed length n, the Hamming distance is a metric on the set of the words of length n.
- For strings of the equal length, Hamming distance is an upper bound on Levenshtein distance.
- For binary strings, the Hamming distance is equal to the number of ones in a XOR.
- Hamming distance is used in telecommunications to count the number of bad bits in a fixed-length binary word, in order to estimate errors. Hamming weight analysis of bits is used in information theory, coding theory and cryptography.

Hamming distance: example 1

Consider two strings $a =$ `kathrin` and $b =$ `kerstin`.

We have:

1. `kathrin` ⟶ `kethrin` (substitute "a" by "e");
2. `kethrin` ⟶ `kerhrin` (substitute "t" by "r");
3. `kerhrin` ⟶ `kersrin` (substitute "h" by "s");
4. `kersrin` ⟶ `kerstin` (substitute "r" by "t").

Then,

$$\mathrm{d}_{\mathrm{ed}}(a, b) = 4.$$

Hamming distance: example 2

Consider two strings $a = 2173895$ and $b = 2233796$.
We have:

1. $2173895 \longrightarrow 2273895$ (substitute “1” by “2”);
2. $2273895 \longrightarrow 2233895$ (substitute “7” by “3”);
3. $2233895 \longrightarrow 2233795$ (substitute “8” by “7”);
4. $2233895 \longrightarrow 2233796$ (substitute “5” by “6”).

Then,

$$\mathrm{d}_{\mathrm{ed}}(a, b) = 4.$$

Hamming distance: example 3

Consider two strings $a = 10100$ and $b = 11011$.
We have:

1. $10100 \longrightarrow 11100$ (substitute “0” by “1”);
2. $11100 \longrightarrow 11000$ (substitute “1” by “0”);
3. $11000 \longrightarrow 11010$ (substitute “0” by “1”);
4. $11010 \longrightarrow 11011$ (substitute “0” by “1”).

Then,

$$\mathrm{d}_{\mathrm{ed}}(a, b) = 4.$$

Hamming distance vs Levenshtein distance: example

Consider two strings $a = \texttt{flaw}$ and $b = \texttt{lawn}$. Compute the Hamming distance.
We have:

1. $\texttt{flaw} \longrightarrow \texttt{llaw}$ (substitute "f" by "l");
2. $\texttt{llaw} \longrightarrow \texttt{laaw}$ (substitute "l" by "a");
3. $\texttt{laaw} \longrightarrow \texttt{laww}$ (substitute "a" by "w");
4. $\texttt{laww} \longrightarrow \texttt{lawn}$ (substitute "w" by "n").

Then,

$$\mathrm{d_{ed}}(a, b) = 4.$$

If we compute the Levenshtein distance:

1. $\texttt{flaw} \longrightarrow \texttt{law}$ (delete "f");
2. $\texttt{law} \longrightarrow \texttt{lawn}$ (insert "n").

Hence,

$$\mathrm{d_{ed}}(a, b) = 2.$$

Damerau-Levenshtein distance

The Damerau-Levenshtein distance allows deletion, insertion, substitution and transposition of two adjacent characters.

Motivation

Damerau stated that the four operations correspond to more than 80% of all spelling errors.

Damerau-Levenshtein distance: example

Consider two strings $a = \{\texttt{an act}\}$ and $b = \{\texttt{a cat}\}$. We have:

1. $\texttt{an act} \longrightarrow \texttt{a act}$ (delete "n");
2. $\texttt{a act} \longrightarrow \texttt{a cat}$ (swap "a" and "c").

Then,

$$\mathrm{d}_{\mathrm{ed}}(a, b) = 2.$$

If we compute the Levenshtein distance:

1. $\texttt{an act} \longrightarrow \texttt{a act}$ (delete "n");
2. $\texttt{a act} \longrightarrow \texttt{a cct}$ (substitute "a" by "c");
3. $\texttt{a cct} \longrightarrow \texttt{a cat}$ (substitute "c" by "a").

Hence,

$$\mathrm{d}_{\mathrm{ed}}(a, b) = 3.$$

Modeling text with distances

There are many choices of distances. Which one to choose is based on:

- computational efficiency;
- modeling effectiveness;
- mathematical properties;
- empirical measures.

The Euclidean distance and the Jaccard distance are often chosen because various algorithmic and computational benefits area available. The edit distance is useful for shorter strings.
There are other variants more useful when dealing with larger texts!

The problem

Consider the text from the following 4 short documents (this problem can be generalized for larger documents).

$$D_1 : \texttt{I am Sam.}$$

$$D_2 : \texttt{Sam I am.}$$

$$D_3 : \texttt{I do not like jelly and ham.}$$

$$D_4 : \texttt{I do not, do not, like them, Sam I am.}$$

How can we measure the distance among these 4 documents?

Bag-of-words approach

The simplest model for converting text into an abstract representation to applying a distance is the bag-of-words approach. Each document creates a "bag" and throws each word in that bag, and maintains only the count of each word. This transforms each document into a multiset. However, it is convenient to think this multiset as a (very sparse, meaning mostly 0s) vector.

Bag-of-words vectors

Consider a vector $\mathbf{x} \in \mathbb{R}^N$, for N very large, where N is the number of all possible words. Each coordinate of the vector $\mathbf{x}$ corresponds to a word, and records the number of occurrences of that word.

Remark

These vector representation suggests to use an L_p distance, most commonly L_2, to measure the distance. However, it is more common to use the cosine distance. This has the advantage of not penalizing documents for their length, in principle focusing more on their content. For istance, by using the cosine distance, a document with a simple phrase would be identical to another one that repeates the same phrase many times.

Example: bag-of-words

Starting from the previous example, *i.e.*, let's try to measure the distance among these 4 documents:

$$\begin{aligned} D_1 &: \texttt{I am Sam.} \\ D_2 &: \texttt{Sam I am.} \\ D_3 &: \texttt{I do not like jelly and ham.} \\ D_4 &: \texttt{I do not, do not, like them, Sam I am.} \end{aligned}$$

Consider a N-dimensional space with $N = 11$; it could be much higher.
For each coordinate, we list the corresponding word into a bag as:

$$(\texttt{am, and, do, ham, I, jelly, like, not, Sam, them, zebra}).$$

Now each of the documents D_i $(i = 1, \ldots, 4)$ has the following representative vectors:

$$\begin{aligned} \mathbf{x}_1 &= (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0), \\ \mathbf{x}_2 &= (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0), \\ \mathbf{x}_3 &= (0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0), \\ \mathbf{x}_4 &= (1, 0, 2, 0, 2, 0, 1, 2, 1, 1, 0). \end{aligned}$$

Example: bag-of-words

By using Euclidean and cosine distances, we have:

	$(\mathbf{x}_1, \mathbf{x}_2)$	$(\mathbf{x}_1, \mathbf{x}_3)$	$(\mathbf{x}_1, \mathbf{x}_4)$	$(\mathbf{x}_2, \mathbf{x}_3)$	$(\mathbf{x}_2, \mathbf{x}_4)$	$(\mathbf{x}_3, \mathbf{x}_4)$
d_2	0	2.828	3.317	2.828	3.317	3
$d_{\cos}$	0	0.782	0.423	0.782	0.423	0.339

Remark

- Since the bag-of-words representation measures which words are present, not how they are used, it cannot distinguish the semantic of documents.
- We see that the 11-th coordinate for the word "zebra" is never used, and is 0 in all coordinates of the vectors. If this coordinate was omitted and only a 10-dimensional representation was used, the distances would remain the same. On the other hand, these vectors can be much larger and represent many other words, and this will not affect the distance.

k-Grams

As an alternative to bag-of-word vectors, *k*-Grams can provide a richer context for text. These convert a document into a set (not a multiset), but does not use just the words. There are a few variants of how to apply *k*-Grams. We will consider only Word *k*-Grams and Character *k*-Grams.

Word *k*-Grams

The Word *k*-Grams over a document is the set of all witnessed consecutive sets of k words.

Character *k*-Grams

The Character *k*-Grams over a document is the set of all witnessed consecutive sets of k characters.

Example: Word 2-Grams

Consider the previous 4 documents:

$$D_1 : \texttt{I am Sam.} \qquad\qquad D_2 : \texttt{Sam I am.}$$
$$D_3 : \texttt{I do not like jelly and ham.}$$
$$D_4 : \texttt{I do not, do not, like them, Sam I am.}$$

Let's show the Word 2-Grams of each of the documents. Each Gram G_i is shown as a set of two words in square brackets, and is associated to the document D_i.

$$\begin{aligned}
G_1 &= \{\texttt{[I am], [am Sam]}\},\\
G_2 &= \{\texttt{[Sam I], [I am]}\},\\
G_3 &= \{\texttt{[I do], [do not], [not like], [like jelly], [jelly and], [and ham]}\},\\
G_4 &= \{\texttt{[I do], [do not], [not do], [not like], [like them], [them Sam],}\\
&\qquad \texttt{[Sam I], [I am]}\}.
\end{aligned}$$

We can then compute the Jaccard distances between the sets G_i as representations of the distances between the documents D_i:

	(G_1, G_2)	(G_1, G_3)	(G_1, G_4)	(G_2, G_3)	(G_2, G_4)	(G_3, G_4)
$\mathrm{d_J}$	$1 - \frac{1}{3} = 0.667$	$1 - \frac{0}{8} = 1$	$1 - \frac{1}{9} = 0.889$	$1 - \frac{0}{8} = 1$	$1 - \frac{2}{8} = 0.75$	$1 - \frac{3}{11} = 0.727$

Example: Word 3-Grams

Consider the previous 4 documents:

$$D_1 : \texttt{I am Sam.} \qquad\qquad D_2 : \texttt{Sam I am.}$$
$$D_3 : \texttt{I do not like jelly and ham.}$$
$$D_4 : \texttt{I do not, do not, like them, Sam I am.}$$

Let's show the Word 3-Grams of each of the documents. Each Gram G_i is shown as a set of three words in square brackets, and is associated to the document D_i.

$$G_1 = \{\texttt{[I am Sam]}\}, \qquad\qquad G_2 = \{\texttt{[Sam I am]}\},$$
$$G_3 = \{\texttt{[I do not], [do not like], [not like jelly], [like jelly and], [jelly and ham]}\},$$
$$G_4 = \{\texttt{[I do not], [do not do], [not do not], [do not like], [not like them],}$$
$$\texttt{[like them Sam], [them Sam I], [Sam I am]}\}.$$

We can then compute the Jaccard distances between the sets G_i as representations of the distances between the documents D_i:

	(G_1, G_2)	(G_1, G_3)	(G_1, G_4)	(G_2, G_3)	(G_2, G_4)	(G_3, G_4)
$\mathrm{d_J}$	$1 - \frac{0}{2} = 1$	$1 - \frac{0}{6} = 1$	$1 - \frac{0}{9} = 1$	$1 - \frac{0}{6} = 1$	$1 - \frac{1}{8} = 0.88$	$1 - \frac{2}{11} = 0.818$

Remark

Many other modeling decisions go into constructing a k-Gram. Should punctuation be included? Should whitespace be used as a character, or should sentence breaks be used as a word in Word k-Grams? Should differently capitalization characters or words represent distinct objects? And most notoriously, how large should k be?

Rules for k-Grams

- The more expressive the k-Grams (*e.g.*, keeping track of punctuation, capitalization, and whitespace), the large the quantity of data that is required to get meaningful results; otherwise, most documents will have a Jaccard distance 1 or very close to it unless large blocks of text are verbatim repeated (for instance, plagiarism).
- With longer documents it is better to use Word k-Grams and larger values of k, whereas with shorter documents (like tweets) it is better to use Character k-Grams and smaller values of k. The values used for k are often surprisingly short, such as $k = 3$ or $k = 4$ for both characters and words.
- It can be useful to keep track of starts of sentences, capitalized words, or Word k-Grams which start with "stop words" (that is, very common words like $\{$a, for, the, to, and, that, it,...$\}$ that often signal starts of expressions).

Example: Character 3-Grams

Consider the previous 4 documents:

$$D_1 : \texttt{I am Sam.} \qquad D_2 : \texttt{Sam I am.}$$
$$D_3 : \texttt{I do not like jelly and ham.}$$
$$D_4 : \texttt{I do not, do not, like them, Sam I am.}$$

Let's show the Character 3-Grams of each of the documents, ignoring whitespaces, punctuation, and capitalization. Each Gram G_i is shown as a set of three characters in square brackets.

$$G_1 = \{\texttt{[iam], [ams], [msa], [sam]}\}, \qquad G_2 = \{\texttt{[sam], [ami], [mia], [iam]}\},$$
$$G_3 = \{\texttt{[ido], [don], [ono], [not], [otl], [tli], [lik], [ike], [kej], [eje],}$$
$$\texttt{[jel], [ell], [lly], [lya], [yan], [and], [ndh], [dha], [ham]}\},$$
$$G_4 = \{\texttt{[ido], [don], [ono], [not], [otd], [tdo], [otl], [tli], [lik], [ike],}$$
$$\texttt{[ket], [eth], [the], [hem], [ems], [msa], [sam], [ami], [mia], [iam]}\}.$$

We can then compute the Jaccard distance between the sets G_i as representations of the distances between the documents D_i:

	(G_1, G_2)	(G_1, G_3)	(G_1, G_4)	(G_2, G_3)	(G_2, G_4)	(G_3, G_4)
$\mathrm{d_J}$	$1 - \frac{2}{6} = 0.667$	$1 - \frac{0}{23} = 1$	$1 - \frac{3}{21} = 0.857$	$1 - \frac{0}{23} = 1$	$1 - \frac{4}{20} = 0.8$	$1 - \frac{8}{31} = 0.742$

Example: Character 4-Grams

Consider the previous 4 documents:

$$D_1 : \texttt{I am Sam.} \qquad D_2 : \texttt{Sam I am.}$$
$$D_3 : \texttt{I do not like jelly and ham.}$$
$$D_4 : \texttt{I do not, do not, like them, Sam I am.}$$

Let's show the Character 4-Grams of each of the documents, ignoring whitespaces, punctuation, and capitalization. Each Gram G_i is shown as a set of four characters in square brackets.

$$G_1 = \{\texttt{[iams]},\ \texttt{[amsa]},\ \texttt{[msam]}\}, \qquad G_2 = \{\texttt{[sami]},\ \texttt{[amia]},\ \texttt{[miam]}\},$$
$$G_3 = \{\texttt{[idon]},\ \texttt{[dono]},\ \texttt{[onot]},\ \texttt{[notl]},\ \texttt{[otli]},\ \texttt{[tlik]},\ \texttt{[like]},\ \texttt{[ikej]},\ \texttt{[keje]},\ \texttt{[ejel]},$$
$$\texttt{[jell]},\ \texttt{[elly]},\ \texttt{[llya]},\ \texttt{[lyan]},\ \texttt{[yand]},\ \texttt{[andh]},\ \texttt{[ndha]},\ \texttt{[dham]}\},$$
$$G_4 = \{\texttt{[idon]},\ \texttt{[dono]},\ \texttt{[onot]},\ \texttt{[notd]},\ \texttt{[otdo]},\ \texttt{[tdon]},\ \texttt{[notl]},\ \texttt{[otli]},\ \texttt{[tlik]},\ \texttt{[like]},$$
$$\texttt{[iket]},\ \texttt{[keth]},\ \texttt{[ethe]},\ \texttt{[them]},\ \texttt{[hems]},\ \texttt{[emsa]},\ \texttt{[msam]},\ \texttt{[sami]},\ \texttt{[amia]},\ \texttt{[miam]}\}.$$

We can then compute the Jaccard distance between the sets G_i as representations of the distances between the documents D_i:

	(G_1, G_2)	(G_1, G_3)	(G_1, G_4)	(G_2, G_3)	(G_2, G_4)	(G_3, G_4)
$\mathrm{d_J}$	$1 - \frac{0}{6} = 1$	$1 - \frac{0}{21} = 1$	$1 - \frac{1}{22} = 0.95$	$1 - \frac{0}{21} = 1$	$1 - \frac{3}{20} = 0.85$	$1 - \frac{7}{31} = 0.774$