Journal of the Brazilian Computer Society, 2024, 30:1, doi: 10.5753/jbcs.2024.2803

An extension to Kendall's Tau metric to evaluate dissimilarities between data series

Bruno Erbisti [**Universidade Federal Fluminense** | ***berbisti@id.uff.br***]
David Kohan Marzagão [**King's College London, University of Oxford** | ***david.kohan@kcl.ac.uk***]
Vanessa Braganholo [**Universidade Federal Fluminense** | ***vanessa@ic.uff.br***]

Universidade Federal Fluminense, Instituto de Computação, Rua Passo da Pátria, 156, São Domingos, Niterói, RJ, Brazil, 24210-240

Received:23 July 2022 • **Accepted:**31 January 2023 • **Published:** 21 September 2024

Abstract Data analysis is performed to examine, interpret, and extract information from data series, and it includes applying various methods and techniques to understand patterns and compare data. An approach to compare data is to use rank metrics that help identify how distinct two data series are when compared to each other according to patterns, formats, criteria, and dimensions in both data series. Among these metrics, Kendall's Tau metric stands out, as it is robust and inexpensive, widely used in analyzing sequences and genomes, to detect errors in flash memories, and to compare distributions and top-k ranked values. However, a challenge arises when comparing lists with different lengths or when lists do not share the same elements. This happens, for example, when lists are defined by top-k elements, commonly called k-list. In this case, there is no guarantee that two k-lists share the same set of elements. Traditional metrics like Kendall's Tau are designed to quantify differences only between shared elements in lists. Recognizing this limitation, a possible solution is to apply the metric to the shared elements of the lists. Another solution, named the generalization of Kendall's Tau, proposed by Fagin et al., considers all elements in two lists. However, this generalization of Kendall Tau is a semi-metric, as it does not satisfy the triangular inequality. To solve this problem, we propose the Extended Kendall Tau (EKT) metric that meets all the conditions of a metric and simultaneously considers the distinct elements of the compared lists. The proposed metric was evaluated by applying conventional Kendall's Tau and the extended Kendall's Tau over 40 text files divided into five different languages (eight files per language). We compared KT and EKT measures within the "same language" and across "other language" files for the two scenarios. The results revealed that both methods could accurately identify the differences between the groups of texts of the "same language" and "other language". However, the numerical results show that EKT is able to more significantly highlight the difference between groups of texts of different languages.

Keywords: Kendall's Tau, metrics, data analysis, dissimilarity

1 Introduction

With the growth in the volume and complexity of data available for research, efficient and effective methods to analyze information are permanent research subjects. For that, there are several techniques to discern patterns, trends, and correlations within data series (a sequence of ordered data)[Weiss *et al.*, 2016; Bewick *et al.*, 2003]. One related problem is data comparison, which consists of measuring the level of dissimilarity of two or more data series, considering the most diverse characteristics shared between the data compared, such as patterns, formats, dimensions, and value distributions. Using metrics to measure dissimilarity between data can be efficient and effective, transcribing complex differences between data series and converting them into a simple value that can be evaluated according to parameters and characteristics established for the problem, which makes evaluation and decision-making easier.

In data comparison, metrics can establish a relationship between two data series [Zhou and Liu, 2014]. In this work, we focus on comparing data sequences, which we refer to here as *lists*. In a list, each element is characterized by a *rank* (a position in the list), since lists have an ordered sequence between their elements. Such lists can be obtained directly from the data or by statistics aggregation [Sculley, 2007]. For example, in Bioinformatics, DNA fragments need to be compared to identify mutations or the presence of a target sequence [Berger *et al.*, 2020]; in Social Sciences, ranking metrics are used to study and compare behavioral data [Pal and Michel, 2016] which includes statistics about the population of states or regions. Several metrics can be used to measure similarity, among which rank metrics such as Cayley [1849] and Kendall [1938] stand out.

Generally, these metrics compare two lists l_1 and l_2 considering the relative position of each of their elements. However, rank metrics generally require that l_1 and l_2 have the same length and share the same set of elements (that is, the set difference between the set of elements from l_1 and the set of elements from l_2 is empty). In Kendall's Tau metric, when two lists do not share the same set of elements (that is, when the set difference is not empty), the different elements can be ignored to get around that restriction, or another treatment can be applied. Fagin *et al.* [2003] proposed a generalization to the Kendall's Tau metric to treat this case, especially in k-lists problems, which are ordered lists with the first k elements of a ranked sequence. However, the generalized Kendall's Tau proposal is a semi-metric [Wilson, 1931] because the triangular inequality (needed to characterize a metric) is not satisfied [Copson, 1988].

To address this limitation, in this work we extend the Kendall's Tau (KT) distance metric [Kendall, 1938, 1976; Abdi, 2007] to measure the differences between two lists that do not necessarily share all elements and such that the triangle inequality property is preserved. We call the proposed metric Extended Kendall's Tau (EKT). We chose Kendall's Tau metric because it has extensive applicability to k-list problems, which typically involve lists with the same length but that do not necessarily share the same elements.

In Kendall's Tau distance metric, every mismatch between pairs of elements in the compared lists is accounted for, and elements that are present in just one list can be disregarded. We propose modifications to KT to consider items that appear in just one of the lists, adapting the original definition. We show that our proposal brings advantages over the original Kendall's Tau metric in cases where one of the lists is not only a rearrangement of the other but also went through insertions and deletions.

The remainder of this article is organized as follows. In Section 2, we present several metrics used in data comparison and their applications, including the KT metric and k-list problems. In Section 3, we present our proposed extension to Kendall's Tau metric. In Section 4, we apply KT and EKT to five different language sets, each composed of 8 text documents of the same language, comprising a total of 40 text documents. We show that: i) for both KT and EKT, it is possible to identify significant differences between the values obtained between documents of "same language" and documents of "different language"; and ii) that EKT can be better than KT for the proposed problem, as it presents a significantly greater difference between the values obtained in the set of documents of "same language" and the set of documents of "different languages." In Section 5, we present the conclusions and propose future work on related questions.

2 Related Work

In this section, we review metrics that can be applied to compare two lists of elements and quantify the differences between them. For that, there are basically two types of metrics: i) correspondence metrics, which establish a correlation between two lists, l_1 and l_2, based on the comparison of each of their elements; and ii) rank metrics, which establish a hierarchy in the position occupied by the elements that are present in the compared lists.

Correspondence Metrics. Hamming [Hamming, 1950; Bookstein *et al.*, 2002] distance measures the discordance between two data strings (or lists). For this reason, this metric is often used to measure the error rate in data transfers [MacKay, 1999], machine learning [Norouzi *et al.*, 2012], and genomes [Kruskal, 1983]. The Longest Common Subsequence (LCS) algorithm [Wagner and Fischer, 1974; Hunt and MacIlroy, 1976; Bakkelund, 2009] is useful to natural language recognition [Lin and Och, 2004] and also defines a dissimilarity metric often used to measure the differences between two sequences of characters. The LCS metric is used to help distinguish differences between DNA strands [Shyu and Tsai, 2009] on genomic bases, in data compression [Cormode and Muthukrishnan, 2005], and detection of similar or plagiarized text [Elhadi and Al-Tobi, 2009].

The edit distance (ED) metric measures the number of editing operations (inclusion, deletion, and replacement) that is needed to transform a list l_1 into another list l_2. This metric is similar to LCS, and the following equation gives the relationship between them: $\text{ED}(l_1, l_2) = |l_1| + |l_2| - 2 \times \text{LCS}(l_1, l_2)$. Edit distance is often used to measure the level of similarity between words and establish a correlation between them. Its applications are extensive, ranging from temporal data analysis and data visualization [Sridharamurthy *et al.*, 2018] to spelling correction [Kukich, 1992] and neuroscience [Gillette *et al.*, 2015].

Although correspondence metrics are useful in many applications, some of the problems that require analysis of dissimilarity between different data series are better analyzed by correlation coefficients or rank metrics, in which the data source is converted into a distribution of ordinal data, formed by categories or partitions, normalized, arranged in lists, and compared using a set of rules.

Rank Metrics. Rank metrics are flexible and can be applied directly in two sequences of ordered data (in the same way as the edit distance), or to ordered lists derived from the distribution of a statistical variable. In the latter case, the elements of the lists being compared represent statistical classifications derived from the original data. At the same time, classification metrics applied to lists of ordered statistical data have the advantage of being less sensitive to outliers and errors in the source data since the comparison between elements represents the comparison of data groups and not the values of individual data items. This approach is used to compare data series when the source data maintains the same nature, type, and format and is classified according to dimensions and classifications shared between both data series.

One of the first rank-based metrics, Spearman's ρ [Spearman, 1904] is treated as a correlation coefficient (similar to Pearson coefficient [Pearson, 1895]), but, unlike it, can be associated with a normalized metric value between 0 to 1. This metric measures the dissimilarity between lists through the sum of the difference square between the ranks of each element present in two lists and helps to identify monotonical trends. Its applications include digital image processing [Nalepa and Gwiazda, 2019], machine learning and natural language processing [Galley *et al.*, 2015], and ecology [Almeida-Neto *et al.*, 2008]. In the same way, Spearman's Footrule [Spearman, 1906] measures dissimilarity by summing rank differences between elements in two lists. Its applications include journal classification [Bar-Ilan, 2010], and election bribery [Baumeister *et al.*, 2019].

Another metric that uses the ranking concept is the Cayley distance [Cayley, 1849; de Lima and Ayala-Rincón, 2012; Fligner and Verducci, 1986], which measures the minimum number of permutations between two elements to transform one sorted list into another. In this metric, the distance in rank position between the permuted elements is neglected in the calculation – only the total permutations matter. For this reason, the metric is equivalent to the minimum permutation

Table 1. Example of values, maximum values, and normalized values in several rank metrics calculated between l_1 = ['b', 'a', 'c', 'd', 'x', 'y', 'z'] and l_2 = ['b', 'c', 'z', 'd', 'x', 'y', 'a']. The normalized form to l_1 is $\hat{l_1}$=[1, 2, 3, 4, 5, 6, 7] and $\hat{l_2}$= [1, 3, 7, 4, 5, 6, 2]. "Value" is the absolute distance calculated for the presented metric; "maximum value" is the largest possible value between these two lists for the metric. The normalized value is defined by $value/maxvalue$.

rank metric	description	value	max value	norm. value
Spearman's Footrule	abs. difference between ranks	10	24	0,4167
Spearman's ρ	sqr difference between ranks	42	224	0,1875
Cayley	permutations (selection sort)	2	6	0,3333
Kendall's Tau	adjacent permutations (bubble sort)	8	21	0,3810

operations using the Selection Sort algorithm [Jadoon *et al.*, 2011]. It is used to rank modulation in flash memories [Jiang *et al.*, 2009; Chee *et al.*, 2014], to evaluate dissimilarity in genomics [Moulton and Steel, 2012] and cloud storage [Yang *et al.*, 2019].

Kendall's Tau distance [Kendall, 1938, 1976; Abdi, 2007; Fligner and Verducci, 1986], which at first glance may seem very similar to Spearman's ρ [Gilpin, 1993; Monjardet, 1998], considers the sum of discordant pairs in two normalized lists. The intuition for that is as follows. Given two lists of length n, l_1 and l_2, with l_1 = ['a', 'b', 'c', 'd'] and l_2 = ['a', 'c', 'd', 'b'], we normalize one of them and write the other as a function of the normalized form of the first one. (A list l' is called normalized when each element is replaced by its position on the corresponding original list l. For example, for l_1 = ['a', 'b', 'c', 'd'], the corresponding normalized list is $l'_1 = [0, 1, 2, 3]$.) The second list is written in terms of the normalized form of l'_1, so $l'_2 = [0, 2, 3, 1]$. In the normalized form, for each element $l'_1[i] \in l'_1$ and $l'_2[j] \in l'_2$, with $i, j = 0, 1, ...n$ and $j > i$, the discordant pairs are those where $l'_1[i] \geq l'_2[j]$ and the sum of discordant pairs is the absolute value for Kendall's Tau metric. In this case, there are two discordant pairs, which are $l'_1[1] \geq l'_2[3] \Rightarrow 1 \geq 1$ and $l'_1[2] \geq l'_2[3] \Rightarrow 2 \geq 1$.

More intuitively, we can understand this metric as the number of transpositions of adjacent elements to transform one sequence into another. This procedure results in the same number of permutations that happen in the bubble sort algorithm. It minimizes the number of adjacent movements of elements in the list since the elements, after being moved, maintain their final position until the end of the algorithm execution. Like the other rank metrics, Kendall's tau distance has applications in machine learning and natural language processing [Galley *et al.*, 2015], in flash memories and information theory [Zhang and Ge, 2015; Buzaglo and Etzion, 2015; Chee *et al.*, 2014].

To exemplify the ranked metrics we discuss in the Section, Table 1 shows the different distances for l_1= ['b', 'a', 'c', 'd', 'x', 'y', 'z'] and l_2=['b', 'c', 'z', 'd', 'x', 'y', 'a']. In the Table, "Value" is the absolute distance calculated for the presented metric, "max value" is the largest possible value calculated for a specific metric, given the length n of the lists. The normalized value is defined by $value/maxvalue$.

Kendall's Tau for Distinct Sets. Kendall's Tau metric and other rank metrics are defined for the cases where the two compared lists (l_1 and l_2) share the same set of elements. Therefore, it becomes a challenge to deal with cases in which it is not possible to have a complete match between the elements of the lists. The naive method to deal with this problem is disregarding distinct elements, calculating the metric only for the intersection of the sets of elements of both lists. This procedure, although simple, can miss significant dissimilarity between the lists, especially when the number of different elements is considerable. To overcome this problem, [Fagin *et al.*, 2003] proposed a generalization of Kendall's Tau metric, in which the values for this metric are calculated over the entire set $\pi_{l_1} \cup \pi_{l_2}$, $\pi_{l_1} = \{l_1[i] | i = 0, 1...n\}$ and $\pi_{l_2} = \{l_2[j] | j = 0, 1...m\}$, but the proposal may violate triangular inequality and thus is a semi-metric. Despite that, it is interesting because it considers the dissimilarity introduced by elements that belong exclusively to one of the two lists. Sculley [2007] proposed establishing a similarity identification and a correspondence between distinct elements to overcome this problem. By using similarity to match elements, the value of the metric may be influenced by incorrect matches. Also, this does not solve the problem of comparing lists with different sizes.

k-lists. Problems involving k-lists (ordered lists with the top k elements according to any arbitrary criteria) are types of problems in which Kendall's Tau metric is widely used, and the presence of distinct elements between lists cannot be simply ignored. Problems involving k-lists consist of comparing the first k elements in two or more lists. As the top k elements of a list are unlikely to be the same as those of another list, dealing with distinct elements can be imperative. Kendall's Tau metric is widely used in k-lists: in computing, k-list problems are found in page ranking systems [DeWitt, 2004], in data mining [Xiong *et al.*, 2006], in machine learning [Collas and Irurozki, 2021], in recommendation systems [Schröder *et al.*, 2011], locality-sensitive hashing (LSH) [Pal and Michel, 2016] and search engines [Hong *et al.*, 2011].

3 Extended Kendall's Tau

The previous section presented several metrics that can be applied to compare data sequences and pointed out their main differences. Although these metrics can be useful in many cases [Schober *et al.*, 2018], rank metrics can only be properly applied when all elements of the ordered sequences are present in both lists so that $l_2 \setminus l_1 = l_1 \setminus l_2 = \varnothing$. In real situations, this condition may not be satisfied, and to work around this problem, Cicirello [2020] proposes that lists with different elements should be treated by removing these elements, leaving only the common elements. This approach avoids the problem of calculating rank metrics for different-sized lists and does not interfere with how the metric's value is calculated. However, differences introduced by the removed elements will not be measured, so the metric's value may not be adequate to compare such lists.

To solve this problem, we propose an extension that considers transforming one compared list into the other, that is, transforming l_1 into l_2 by: i) removing exclusive elements from l_1; ii) transforming the new l_1 into a sub-sequence of l_2; iii) and finally, inserting the missing elements to complete l_2 in the proper order. This extension can be generalized to different order metrics, but we chose Kendall's Tau because this metric considers each movement between adjacent elements to transform one list into the other, thus establishing a more robust relationship between the positions of elements in the lists and the value of the metric. Furthermore, Kendall's Tau minimizes the number of adjacent moves, as its calculation is identical to the number of sorting moves that are done by the bubble sort algorithm [Astrachan, 2003]. In this way, we can describe Kendall's Tau as a count of the bubble sort operations to turn $l_1 \rightarrow l_2$ when the lists are formed by the same elements, as mentioned in item ii).

Kendall's Tau [Kendall, 1938, 1976] can be mathematically described as follows. Let $l_1 = \{a_1, a_2, ..., a_n\}$ and $l_2 = \{b_1, b_2, ..., b_n\}$ be two lists whose elements have an ordering relation. For each pair of elements (a_i, a_j) in l_1 and (b_i, b_j) in l_2 where $i < j$, we calculate a discordant pair if: $(a_i > a_j$ and $b_i < b_j)$ or $(a_i < a_j$ and $b_i > b_j)$. We write this as $\tau((a_i - a_j) \cdot (b_i - b_j))$ where τ is an indicator function that is 1 if (a_i, a_j) and (b_i, b_j) are discordant pairs or 0 otherwise. The Kendall Tau distance $d_{kt}(l_1, l_2)$ between the two lists l_1 and l_2 with length n is then the total number of discordant pairs. Mathematically, this can be expressed as:

$$d_{kt}(l_1, l_2) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \tau((a_i - a_j) \cdot (b_i - b_j)) \qquad (1)$$

To add steps i) and iii) into the original Kendall's Tau algorithm, we must also consider that the insertion/removal method of elements must preserve a metric model under the space of ordered lists, complying with the following conditions [Choudhary, 1993]:

C1. coincidence: $m(l_1, l_2) = 0 \Leftrightarrow l_1 = l_2$, for every l_1, l_2 ;
C2. symmetry: $m(l_1, l_2) = m(l_2, l_1)$, for every l_1, l_2 ;
C3. triangle inequality: $m(l_1, l_3) \leq m(l_1, l_2) + m(l_2, l_3)$, for every l_1, l_2, l_3 .

To meet the symmetry criteria (C2), the removal and insertion of distinct elements from the lists must be performed in the inverse order; that is, elements should be removed from list l_1 (item i) from lowest to highest rank position, while the insertion order of elements (item iii) must be from highest to the lowest rank. The total removal distance d_{rem} and total insertions distance d_{ins} are represented in Equations 2 and 3 below:

$$d_{rem} = \sum_{i=1}^{n_{l_1}} \delta(l_1[i] \in l_1 \setminus l_2) \cdot i \qquad (2)$$

$$d_{ins} = \sum_{i=1}^{n_{l_2}} \delta(l_2[i] \in l_2 \setminus l_1) \cdot i \qquad (3)$$

where n_{l_1} and n_{l_2} are the length of list l_1 and l_2, respectively, and δ is an indicator function that equals 1 when the condition is true and 0 otherwise. Thus, Kendall's Tau extended distance d_{ekt} can be written in the form:

$$d_{ekt} = d_{kt} + d_{rem} + d_{ins} \qquad (4)$$

For normalization, we must consider the maximum value z that can be obtained in the disarrangement between three sets: the set of size n_1 containing the common elements of l_1 and l_2, the set of size n_2 containing the inserted elements, and the set of size n_3 containing the removed elements:

$$z = \frac{n_1 \cdot (n_1 - 1)}{2} + n_1 \cdot n_t + \frac{n_t}{2} + \frac{n_t^2}{2} - n_2 \cdot n_3 \qquad (5)$$

For ease of writing, we represent $n_t = n_2 + n_3$, and the extended Kendall's Tau distance in normalized form is given by:

$$\hat{d}_{ekt}(l_1, l_2) = \frac{1}{z} \cdot (d_{kt} + d_{ins} + d_{rem}) \qquad (6)$$

Kendall's Tau metric between two lists l_1 and l_2 establishes a measure directly linked to the concept of the smallest number of adjacent movements between elements to transform a list l_1 into l_2. This property is closely related to the idea of this metric itself and its definition in Equation 1 and guarantees that this transformation occurs in the same way when translated into an algorithm (in this case, bubble sort). To ensure the same condition using the presented removal and insertion criteria, the removal must be carried out before the insertion of elements to transform one list into another, always from the highest to the lowest rank position, and in a symmetrical way, the insertion must be carried out after the removal, from the lowest to the highest rank position. The procedure prevents redundant moves, resulting in fewer adjacent insertions, removals, and moves.

As an example, suppose two lists: l_1=[2, 65, 21, 42, 6, 5, 12, 35] and l_2 =[2, 65, 32, 21, 12, 5, 6, 35]. Note that $\{l_2\} \setminus \{l_1\}$ = {32}, and $\{l_1\} \setminus \{l_2\}$ = {42}. To turn l_1 into l_2, we remove the 4th element (42) in l_1 from right to left, with four movements (and shift elements from right to left to occupy the position of the removed element). The resulting list is [2, 65, 21, 6, 5, 12, 35] and 4 movements in the removal step. Now, we insert the 3rd element of l_2 (32) into l_1, from left to right, in the third position, calculating the number of adjacent movements that are needed for this operation. In this case, starting before the 1st position, the total number of movements is $1 + 1 + 1 = 3$ in the insertion step, and the resulting list is [2, 65, 32, 21, 6, 5, 12, 35]. Finally, we calculate Kendall's Tau (Eq. 1) - which is equivalent to calculating the total number of bubble sort moves, transforming l_1 into l_2. In this procedure, the 7th element (12) changes to the 5th position with two movements, and the 6th element (5) changes to the 5th position, with one movement. The total of adjacent movements is the not-normalized distance from l_1 to l_2, which in this example is $4 + 3 + 2 + 1 = 10$.

Normalization ensures that the metric value is within the range $[0, 1]$. The normalization factor z for two lists l_1 and l_2, with sizes m and n respectively, and one set of common elements n_c is given in equation 5, with z = 37. Then $\hat{d_{ekt}} = \frac{10}{37} = 0.27027$.

Our proposed extension does not consider absolute values of the ranks but the relative ranks between elements. It doesn't matter whether the elements in both lists are sorted (ascending or descending) by relevance, value, or any other feature. As long as the same criteria are used to sort both of the lists, the metric value is preserved. However, when introducing the concept of insertion and removal of elements, we establish a direction to apply the procedure, considering the importance level and hierarchical order. To avoid problems in applying the method, by default, we consider insertions and removals from the left, and we consider that the lists are sorted in an ascending way in terms of pertinence.

4 Results

This section compares results in applying the original Kendall's Tau metric and the extended proposal in a language dissimilarity identification problem. We show that by applying a rank metric, we can (i) effectively identify dissimilarities between languages and (ii) increase the efficiency when we use the extended Kendall's Tau metric to consider lists containing distinct elements. To meet this goal, our work was directed towards a simple problem that implies reduced list lengths since the complexity and volume of information are out of the scope of this work.

Table 2. Group statistics for the eight documents we selected for each language.

Idiom	letters	words	phrases	pages
ge	898214	163979	8016	534
es	708431	117281	7208	480
en	3209103	677922	36830	2455
it	452194	79848	4380	292
pt	3600734	559119	36790	2453
Total	8868676	1598149	93224	6215

In this evaluation, we search for 40 documents written in 5 different languages, forming a set of 8 text documents for each language. The chosen languages were Portuguese (pt), English (en), Italian (it), German (get), and Spanish (es), and the documents were selected in an ad-hoc way, without intention as to content, date, or author, in digital public libraries and public repositories in different countries. All documents were available in PDF format, converted to plain text using an OCR algorithm, and checked during conversion.

To prepare the texts for subsequent analysis, we removed characters that were not relevant and standardized the text to avoid the distinction between uppercase and lowercase to make the analysis simpler. Characters removed comprise control codes, spaces, unusual symbols, accents, and other non-alphabetic characters. Table 2 shows some statistics about the collected documents. Even with great variability in the volumes of data collected per language, the smallest number of letters per language is close to half a million (Italian), and individual samples are well distributed per document. The original documents and the source code we used in our analysis are available at `https://github.com/dew-uff/EKT/`.

Table 3. There are five "same language" sets (pt-pt, sp-sp, ge-ge, it-it, en-en) and ten "different languages" sets (ge-pt, sp-pt, en-pt, it-pt, ge-it, sp-it, en-it, ge-en, sp-en, ge-sp). In "same language," we combine 8 lists 2 by 2. In "different language," eight lists in one language combine with eight lists in another language.

	sets	combinations	total
same language	5	$\binom{8}{2} = 28$	140
different languages	$\binom{5}{2} = 10$	$8 \cdot 8 = 64$	640
total			780

An associated list was generated for each document: the elements in these lists are the text characters ordered according to their frequency in the respective document. To calculate all KT and EKT distances, firstly, we form all pairs of lists out of the 40 lists without caring about the order, according to condition C2 in Section 3, which represents a total of $\binom{40}{2} = 780$ pairs of lists. These pairs of lists can be categorized as: "same language" (group 1) and "different language" (group 2).

The total number of pairs of lists of group 1 is obtained by choosing a pair of lists among the eight in the same language, forming 28 possible combinations for each language. Since there are five language sets, the total number of list pairs is $28 \cdot 5 = 140$.

The sets of group 2 are formed by choosing two lists corresponding to different languages, among 5, in a total of ten different-language combinations. In each of the ten combinations, eight lists of one language are compared with the other eight lists of another language, resulting in $10 \cdot 8 \cdot 8 = 640$ pairs. In Table 3, we see the description of the formed pairs, the number of combinations, and the pairs formed in each combination. For each pair, we calculate KT and EKT distances.

Table 4 and Table 5 show examples of calculated KT and EKT values for list pairs. In the 4 table, we show an example of a "same language" set, in this case, with list pairs of texts in Portuguese. For each KT and EKT metric, 28 values are generated among the eight lists of "same language." In Table 5, we show an example of the "different languages" category, in this case, English and Portuguese. 64 KT and EKT values are generated for the 16 lists of these two languages.

To show that the metric effectively detects differences between languages, we compare the KT and EKT values obtained in group 1 and those obtained in group 2. To make this analysis more robust, we perform this procedure for each language separately, as shown in Figure 1. The Figure shows the groups formed by group 1 and group 2. For every group from the same language, there are four others from different languages to compare to. In this way, $5 \cdot 4 = 20$ comparisons are performed comparing these two groups, according to Figure 1. The complete set of values can be accessed in `https://github.com/dew-uff/EKT/`.

We compared groups of values using the Mann-Whitney test [Mann and Whitney, 1947] both for KT and EKT measurements. The test is justified because we have sets of different sizes, with 28 and 64 unpaired values for each metric.

Table 4. An example of a set of KT and EKT distance values obtained within the same language (Portuguese). All groups of the same language are of the form $l_i - l_i$ (pt-pt, sp-sp, ge-ge, it-it, en-en) and have 28 calculated values for each group.

n	pair $(l_i - list_a, l_i - list_b)$	KT	EKT
1	pt-list1,pt-list2	0.0683	0.0716
2	pt-list1, pt-list3	0.037	0.037
3	pt-list1, pt-list4	0.0512	0.0555
4	pt-list1, pt-list5	0.0725	0.0716
5	pt-list1, pt-list6	0.0654	0.0671
.	.	.	.
.	.	.	.
.	.	.	.
25	pt-list5, pt-list8	0.0397	0.0397
26	pt-list7, pt-list6	0.0564	0.0545
27	pt-list8, pt-list6	0.0436	0.0429
28	pt-list7, pt-list8	0.0718	0.0718

Table 5. An example of a set of KT and EKT distance values obtained in different languages, in this case, English and Portuguese (en-pt). All groups of the same language are of the form $l_i - l_j, i \neq j$ (ge-pt,sp-pt, en-pt, it-pt, ge-it, sp-it, en-it, ge-en, sp-en, ge-sp), and have 64 calculated values for each group.

n	pair$(l_i - list_a, l_j - list_b)$	KT	EKT
1	en-list1,pt-list1	0.173333	0.271782
2	en-list1,pt-list2	0.156695	0.245902
3	en-list1,pt-list3	0.196667	0.286086
4	en-list1,pt-list4	0.176638	0.268265
5	en-list1,pt-list5	0.162393	0.264817
.	.	.	.
.	.	.	.
.		.	.
61	en-list8, pt-list5	0.202279	0.282472
62	en-list8, pt-list6	0.196581	0.285388
63	en-list8, pt-list7	0.196581	0.272383
64	en-list8, pt-list8	0.202279	0.277427

Also, as it is a non-parametric test, it admits the possibility that the distributions of the distance values are not necessarily normal, which makes our result more robust.

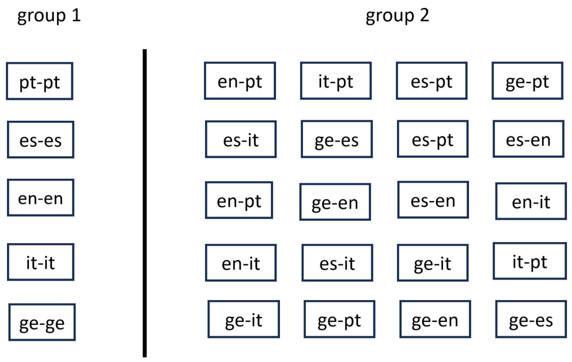

Figure 1. For each language, KT and EKT group values calculated across lists of the same language (group 1) are compared with the group values from the other four different languages (group 2). Each language generates four comparisons between 2 groups of the same language and different languages, amounting to 20 comparisons. Comparisons are made using the non-parametric Mann-Whitney test, ensuring that the comparison can be performed regardless of normality and variance between groups.

Table 6. Mann-Whitney test for two set values (group 1 vs group 2); measure = KT distance; H_0: group 1 (same languages) = group 2 (different languages): all p-values $< 10^{-13}$.

group 1	group 2	p-value
ge-ge	sp-ge	2.9429E-14
ge-ge	en-ge	2.8130E-14
ge-ge	it-ge	2.9462E-14
ge-ge	pt-ge	2.9309E-14
sp-sp	sp-ge	2.9389E-14
sp-sp	sp-en	2.9356E-14
sp-sp	sp-it	3.8185E-14
sp-sp	sp-pt	2.9369E-14
en-en	sp-en	2.9469E-14
en-en	en-ge	3.5467E-14
en-en	en-it	5.7920E-13
en-en	en-pt	2.9476E-14
it-it	sp-it	7.3390E-14
it-it	en-it	4.1146E-14
it-it	en-pt	2.9476E-14
it-it	it-ge	2.9536E-14
pt-pt	sp-pt	4.6571E-14
pt-pt	en-pt	2.9469E-14
pt-pt	pt-ge	2.9376E-14
pt-pt	pt-it	2.9630E-14

table[t]

group 1	group 2	U-value	p-value
ge-ge	sp-ge	0,0	2,9757E-14
ge-ge	en-ge	41.0	4,1195E-13
ge-ge	it-ge	0.0	2,9838E-14
ge-ge	pt-ge	0.0	2,9804E-14
sp-sp	sp-ge	0.0	2,9710E-14
sp-sp	sp-en	0.0	2,9623E-14
sp-sp	sp-it	0.0	2,9751E-14
sp-sp	sp-pt	0.0	2,9751E-14
en-en	sp-en	0.0	2,9677E-14
en-en	en-ge	0.0	2,9730E-14
en-en	en-it	0.0	2,9858E-14
en-en	en-pt	0.0	2,9818E-14
it-it	sp-it	0.0	2,9811E-14
it-it	en-it	0.0	2,9865E-14
it-it	en-pt	0.0	2,9825E-14
it-it	it-ge	0.0	2,9852E-14
pt-pt	sp-pt	0.0	2,9764E-14
pt-pt	en-pt	0.0	2,9778E-14
pt-pt	pt-ge	0.0	2,9771E-14
pt-pt	pt-it	0.0	2,9804E-14

Table 7. Mann-Whitney test for two set values (group 1 vs group 2); measure = KT distance; H_0: group 1 (same languages) = group 2 (different languages): all p-values $< 10^{-13}$.

group 1	group 2	U-value	p-value
ge-ge	sp-ge	0.0	2.9429E-14
ge-ge	en-ge	0.0	2.8130E-14
ge-ge	it-ge	0.0	2.9462E-14
ge-ge	pt-ge	0.0	2.9309E-14
sp-sp	sp-ge	0.0	2.9389E-14
sp-sp	sp-en	0.0	2.9356E-14
sp-sp	sp-it	4.0	3.8185E-14
sp-sp	sp-pt	0.0	2.9369E-14
en-en	sp-en	0.0	2.9469E-14
en-en	en-ge	3.5	3.5467E-14
en-en	en-it	46.5	5.7920E-13
en-en	en-pt	0.0	2.9476E-14
it-it	sp-it	14.0	7.3390E-14
it-it	en-it	5.0	4.1146E-14
it-it	en-pt	0.0	2.9476E-14
it-it	it-ge	0.0	2.9536E-14
pt-pt	sp-pt	7.0	4.6571E-14
pt-pt	en-pt	0.0	2.9469E-14
pt-pt	pt-ge	0.0	2.9376E-14
pt-pt	pt-it	0.0	2.9630E-14

Table 8. Mann-Whitney test for two set values (group 1 vs group 2); measure = EKT distance; H_0: group 1 (same languages) = group 2 (different languages): all p-values$< 10^{-13}$.

group 1	group 2	p-value
ge-ge	sp-ge	2.9757E-14
ge-ge	en-ge	4.1195E-13
ge-ge	it-ge	2.9838E-14
ge-ge	pt-ge	2.9804E-14
sp-sp	sp-ge	2.9710E-14
sp-sp	sp-en	2.9623E-14
sp-sp	sp-it	2.9751E-14
sp-sp	sp-pt	2.9751E-14
en-en	sp-en	2.9677E-14
en-en	en-ge	2.9730E-14
en-en	en-it	2.9858E-14
en-en	en-pt	2.9818E-14
it-it	sp-it	2.9811E-14
it-it	en-it	2.9865E-14
it-it	en-pt	2.9825E-14
it-it	it-ge	2.9852E-14
pt-pt	sp-pt	2.9764E-14
pt-pt	en-pt	2.9778E-14
pt-pt	pt-ge	2.9771E-14
pt-pt	pt-it	2.9804E-14

The p-values for the test performed between group 1 and group 2 for the KT and EKT values are presented in Tables 7 and 8, respectively. As we can see, the tests conducted on pairs of groups generate very small p-values for both KT and EKT, and therefore, it is irrelevant to apply the Bonferroni correction [Bonferroni, 1936] in this case. These values show that we completely reject the hypothesis that group 1 and group 2 distributions could behave similarly in both metrics.

There is no closeness between groups 1 and 2 at any individualized experiment values for either KT or EKT. As the p-values of KT and EKT are extremely small and oscillate in a minimal range $(10^-13to10^-14)$, we assess that both already have very high efficiency in differentiating lists from group 1 and group 2, and any gain in this sense would not be significant for the proposed problem. However, we can measure d differences between the measures of group 1 and group 2 and observe whether they significantly differ between the KT and EKT values. This statistic is called DiD (differences in differences) and is generally used in longitudinal analyses and econometrics [Wing *et al.*, 2018] or when evaluating paired groups under two different conditions. Our distinct states are represented by KT and EKT metrics, and paired differences are obtained between group 1 and group 2.

We calculated the differences in absolute and standardized values using Cohen's d [Cohen, 2013] to make the analysis more thorough. As a standardized measure, Cohen's d removes any interpretation related to specific orders of magnitude of the experiment and minimizes the effects of possible data outliers on the groups. The absolute differences and Cohen's d are presented in Table 9, where we observe the differences between groups 1 and 2, representing 20 observations for each metric. Although the number of observations is relatively low, the differences between groups are significant and can be observed in a paired way, making it possible to use the Wilcoxon statistical test [Wilcoxon, 1945], specific for paired observations. In Table 10, we present the results of the one-tailed Wilcoxon test for H_0: $EKT <= KT$. For absolute d values, $EKT > KT$ with p-value 0.004 (< 0.01). Even when considering Cohen's d, we confirm the same result, with a p-value of 0.02 (< 0.05). Table 10 ensures that the difference between the measures of equal groups and different language groups (group 1 and group 2) are more significant when we use the EKT metric.

table[t]

Wilcoxon, one-tailored, d-value (KT and EKT)

measure	z-value	w-value	p-value	result
Cohen's d	-1,9386	223	0,0262	approximately normal, significantly different groups
absolute d	-2,6506	34	0,0040	n=20, w < 60, significantly different groups

5 Conclusions

In this article, we propose an extension to the Kendall's tau metric. The extension accommodates handling lists with different elements while still preserving the metric properties, unlike previous extensions proposed in the literature [Fagin *et al.*, 2003; Sculley, 2007].

Comparison of Kendall's tau and modified Kendall's tau metrics applied to forty documents divided into groups of eight documents for five different languages proved to be effective in differentiating documents of the same language and documents of different languages. In both metrics, the values obtained in the group of pairs of documents with dif-

Table 9. Absolute and standardized (Cohen´s d) difference between mean values on group 1 (same language) and group 2 (different language).

compared groups		absolute d values		Cohen's d values	
group 1	group 2	KT	EKT	KT	EKT
ge-ge	sp-ge	0.1701	0.1920	13.3087	5.5082
ge-ge	en-ge	0.1114	0.1811	9.2520	3.0360
ge-ge	it-ge	0.1468	0.1177	11.7628	5.1245
ge-ge	pt-ge	0.1739	0.2311	12.1702	4.9695
sp-sp	sp-ge	0.1508	0.2035	9.8763	11.8260
sp-sp	sp-en	0.1035	0.1438	5.7589	8.7214
sp-sp	sp-it	0.1845	0.1682	3.4274	8.3757
sp-sp	sp-pt	0.1846	0.1247	3.5886	8.3890
en-en	sp-en	0.2117	0.1351	3.9122	6.2865
en-en	en-ge	0.1210	0.1149	4.0150	4.7303
en-en	en-it	0.0810	0.1204	2.6527	4.4222
en-en	en-pt	0.1170	0.2088	4.9538	10.9579
it-it	sp-it	0.0558	0.1601	3.5771	6.3979
it-it	en-it	0.0978	0.1211	3.7763	4.4547
it-it	en-pt	0.1339	0.2094	7.0455	11.0249
it-it	it-ge	0.1340	0.1845	8.2820	8.1525
pt-pt	sp-pt	0.0454	0.1215	3.1657	7.6649
pt-pt	en-pt	0.1239	0.2144	6.6945	14.2463
pt-pt	pt-ge	0.1511	0.2394	8.6392	8.0217
pt-pt	pt-it	0.0604	0.2107	3.2196	13.8630

Table 10. Wilcoxon test for 20 paired groups of d-values for KT and EKT as presented in Table 9. The tests were conducted in two cases: absolute d-values and Cohen's d-values. Tests present $EKT > KT$ in d-values in both absolute and Cohen's d values.

Wilcoxon, one-tailored, d-value (KT and EKT)		
measure	p-value	result
Cohen's d	0.0262	significantly different groups
absolute d	0.0040	significantly different groups

ferent languages and in the group of pairs of documents with the same language are significantly different, with a p-value of less than 10^13 for all languages analyzed, which guarantees a robust differentiating power for the methods applied to the proposed problem. All pairs can be classified with absolute precision as "same" or "different", and although it is not possible to assign a priori an absolute value that identifies the dissimilarities between any pairs, but limits for significant EKT values can be previously obtained according to a set of previously analyzed data.

The difference between the p-values in both metrics oscillated between 10^{13} and 10^{14}, which are very small values, even when considering the Bonferroni [Bonferroni, 1936] correction. Therefore, both metrics are effective in solving the proposed problem, and, in this case, there is no significant improvement in the use of the proposed method.

However, when we consider the relative distances between the values x' obtained for the set of pairs of a language and the set of values x obtained for the pairs formed between this language and a different language, we observe that the differences $(x' - x)$ are significantly larger when using Kendall's tau extension, even when using these distances in the standardized form [Cohen, 2013].

We can observe that extended Kendall's tau considers an additional component in the calculation of Kendall's tau distance, and therefore it is natural that all distances calculated by the method are greater than or equal to the distances calculated by Kendall's tau. However, as the difference between the two distances is also significantly greater, we claim that the application of the extended method can better differentiate data samples in terms of their distribution properties. In practice, the application of Kendall's tau extension considers factors such as the occurrence of a letter in only one document. This type of phenomenon is also correlated with the language in which a document is written, as some letters are particularly used in a given language and not on another.

The study suggests that the application of EKT can be useful in problems where lists are not guaranteed to have the same elements without compromising their effectiveness. Furthermore, this study suggests that the application of EKT may represent increased value differences between similar and dissimilar values, which may facilitate a binary recognition pattern.

Another important aspect to consider is that applications of KT and EKT metrics can be done to more complex problems than the one we show in this work, and that can generate significant computational cost. While the purpose of this work was to show a practical application in which EKT metric may be more appropriate, we can also optimize the metric calculation on much larger lists. This problem can be solved by applying a more efficient algorithm adapted to measure the neighboring displacements (bubble sort movements) using a merge function. This method can reduce the complexity of the problem from O(n^2) to O(n.log n). As the proposed solution in EKT metric also assumes a rank-ordered hierarchy (from left to right or from right to left), it may be useful to apply a cut or limit on the length of the lists since the variation in the order of the elements with little frequency can produce undesirable random variations in distance values. We plane to address these issues in future work.

Declarations

Acknowledgements

We would like to thank the reviewers for their valuable suggestions. We would also like to thank IBGE, FAPERJ, and CNPq for funding this research.

Funding

CNPq grants 305020/2019-6 and 309410/2023-1, FAPERJ grant E-26/210.478/2024, and IBGE grant (PhD license 0001788.00000332/2018-23).

Authors' Contributions

Bruno Erbisti proposed the extension to Kendall's Tau metric. He also conducted the experimental evaluation and wrote the first draft of the paper. David contributed to the formalism and polishing of the final version of the paper. Vanessa worked on the text, polishing and rewriting it, and supervised the research.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

To ensure reproducibility, Python codes, input data, and results presented in the paper are available at: https://github.com/dew-uff/EKT/

References

Abdi, H. (2007). The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510. Available at:`https://www.cs.uoi.gr/~pvassil/supervision/diplomatikes/extraMaterial/Correlation/Kendall/2007_KendallCorrelation_shortExplanation.pdf`.

Almeida-Neto, M., Guimaraes, P., Guimaraes Jr, P. R., Loyola, R. D., and Ulrich, W. (2008). A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, 117(8):1227–1239. DOI: 10.1111/j.0030-1299.2008.16644.x.

Astrachan, O. (2003). Bubble sort: an archaeological algorithmic analysis. *ACM Sigcse Bulletin*, 35(1):1–5. DOI: 10.1145/792548.611918.

Bakkelund, D. (2009). An lcs-based string metric. *Olso, Norway: University of Oslo*. Available at:`https://www.researchgate.net/profile/Daniel-Bakkelund/publication/242065008_An_LCS-based_string_metric/links/57ff543b08ae56fae5f55421/An-LCS-based-string-metric.pdf`.

Bar-Ilan, J. (2010). Rankings of information and library science journals by jif and by h-type indices. *Journal of informetrics*, 4(2):141–147. DOI: 10.1016/j.joi.2009.11.006.

Baumeister, D., Hogrebe, T., and Rey, L. (2019). Generalized distance bribery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1764–1771. DOI: 10.1609/aaai.v33i01.33011764.

Berger, B., Waterman, M. S., and Yu, Y. W. (2020). Levenshtein distance, sequence comparison and biological database search. *IEEE transactions on information theory*, 67(6):3287–3294. DOI: 10.1109/TIT.2020.2996543.

Bewick, V., Cheek, L., and Ball, J. (2003). Statistics review 7: Correlation and regression. *Critical care*, 7:1–9. DOI: 10.1186/cc2401.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62. Available at:`https://cir.nii.ac.jp/crid/1570009749360424576`.

Bookstein, A., Kulyukin, V. A., and Raita, T. (2002). Generalized hamming distance. *Information Retrieval*, 5:353–375. DOI: 10.1023/A:1020499411651.

Buzaglo, S. and Etzion, T. (2015). Bounds on the size of permutation codes with the kendall τ-metric. *IEEE Transactions on Information Theory*, 61(6):3241–3250. DOI: 10.1109/TIT.2015.2424701.

Cayley, A. (1849). Lxxvii. note on the theory of permutations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 34(232):527–529. DOI: 10.1080/14786444908646287.

Chee, Y. M. *et al*. (2014). Breakpoint analysis and permutation codes in generalized kendall tau and cayley metrics. In *2014 IEEE International Symposium on Information Theory*, pages 2959–2963. IEEE. DOI: 10.1109/ISIT.2014.6875376.

Choudhary, B. (1993). *The elements of complex analysis*. New Age International. Book.

Cicirello, V. A. (2020). Kendall tau sequence distance: Extending kendall tau from ranks to sequences. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, 7(23). DOI: 10.4108/eai.13-7-2018.163925.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press. DOI: 10.4324/9780203771587.

Collas, F. and Irurozki, E. (2021). Concentric mixtures of mallows models for top-k rankings: sampling and identifiability. In *International Conference on Machine Learning*, pages 2079–2088. PMLR. Available at:`https://proceedings.mlr.press/v139/collas21a.html`.

Copson, E. T. (1988). *Metric spaces*. Number 57. CUP Archive. Book.

Cormode, G. and Muthukrishnan, S. (2005). Substring compression problems. In *SODA*, volume 5, pages 321–330. Citeseer. Available at:`https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=38591684ad4c2362b7ca3de3401e9bae3ce3b66d`.

de Lima, T. A. and Ayala-Rincón, M. (2012). Complexity of cayley distance and other general metrics on permutation groups. In *2012 7th Colombian Computing Congress (CCC)*, pages 1–6. IEEE. DOI: 10.1109/ColombianCC.2012.6398020.

DeWitt, Y. W. D. J. (2004). Computing pagerank in a distributed internet search system. In *Proceedings of the 30th VLDB Conference, Toronto, Canada*, volume 30, pages 420–431. Book.

Elhadi, M. and Al-Tobi, A. (2009). Webpage duplicate detection using combined pos and sequence alignment algorithm. In *2009 WRI world congress on computer science and information engineering*, volume 1, pages 630–634. IEEE. DOI: 10.1109/CSIE.2009.771.

Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160. DOI: 10.1137/S0895480102412856.

Fligner, M. A. and Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):359–369. DOI: 10.1111/j.2517-6161.1986.tb01420.x.

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., and Dolan, B. (2015). deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*. DOI: 10.48550/arXiv.1506.06863.

Gillette, T. A., Hosseini, P., and Ascoli, G. A. (2015). Topological characterization of neuronal arbor morphology via sequence representation: Ii-global alignment. *BMC bioinformatics*, 16(1):1–17. DOI: 10.1186/s12859-015-0605-1.

Gilpin, A. R. (1993). Table for conversion of kendall's tau to spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and psychological measurement*, 53(1):87–92. DOI: 10.1177/0013164493053001007.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160. DOI: 10.1002/j.1538-7305.1950.tb00463.x.

Hong, Y., Vaidya, J., and Lu, H. (2011). Search engine query clustering using top-k search results. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 112–119. IEEE. DOI: 10.1109/WI-IAT.2011.224.

Hunt, J. W. and MacIlroy, M. D. (1976). *An algorithm for differential file comparison*. Bell Laboratories Murray Hill. Available at:https://www.cs.dartmouth.edu/doug/diff.pdf.

Jadoon, S., Solehria, S. F., Rehman, S. u., and Jan, H. (2011). Design and analysis of optimized selection sort algorithm. *International Journal of Electric & Computer Sciences (IJECS-IJENS)*, 11(01):16–22. Available at:https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0047d34b344c05ffa4c0e94ab134d58546a7c6c2.

Jiang, A., Mateescu, R., Schwartz, M., and Bruck, J. (2009). Rank modulation for flash memories. *IEEE Transactions on Information Theory*, 55(6):2659–2673. DOI: 10.1109/TIT.2009.2018336.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93. DOI: 10.1093/biomet/30.1-2.81.

Kendall, M. G. (1976). *Rank correlation methods*. Charles Griffin & Co., London. Book.

Kruskal, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, 25(2):201–237. DOI: 10.1137/1025045.

Kukich, K. (1992). Techniques for automatically correcting words in text. *ACM computing surveys (CSUR)*, 24(4):377–439. DOI: 10.1145/146370.146380.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612. Available at:https://aclanthology.org/P04-1077.pdf.

MacKay, D. J. (1999). Good error-correcting codes based on very sparse matrices. *IEEE transactions on Information Theory*, 45(2):399–431. DOI: 10.1109/18.748992.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60. DOI: 10.1214/aoms/1177730491.

Monjardet, B. (1998). On the comparison of the spearman and kendall metrics between linear orders. *Discrete mathematics*, 192(1-3):281–292. DOI: 10.1016/S0012-365X(98)00076-4.

Moulton, V. and Steel, M. (2012). The 'butterfly effect'in cayley graphs with applications to genomics. *Journal of mathematical biology*, 65(6-7):1267–1284. DOI: 10.1007/s00285-011-0498-1.

Nalepa, B. and Gwiazda, A. (2019). Spearman's rho modification in digital image processing. In *IOP Conference Series: Materials Science and Engineering*, volume 591, page 012060. IOP Publishing. DOI: 10.1088/1757-899X/591/1/012060.

Norouzi, M., Fleet, D. J., and Salakhutdinov, R. R. (2012). Hamming distance metric learning. *Advances in neural information processing systems*, 25. Available at:https://proceedings.neurips.cc/paper/2012/hash/59b90e1005a220e2ebc542eb9d950b1e-Abstract.html.

Pal, K. and Michel, S. (2016). Efficient similarity search across top-k lists under the kendall's tau distance. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management*, pages 1–12. DOI: 10.1145/2949689.2949709.

Pearson, K. (1895). Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242. DOI: 10.1098/rspl.1895.0041.

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768. DOI: 10.1213/ANE.0000000000002864.

Schröder, G., Thiele, M., and Lehner, W. (2011). Setting goals and choosing metrics for recommender system evaluations. In *UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA*, volume 23, page 53. Available at:https://www.researchgate.net/profile/Maik-Thiele-2/publication/268381252_Setting_Goals_and_Choosing_Metrics_for_Recommender_System_Evaluations/links/595e42e1a6fdccc9b17fcffe/Setting-Goals-and-Choosing-Metrics-for-Recommender-System-Evaluations.pdf.

Sculley, D. (2007). Rank aggregation for similar items. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 587–592. SIAM. DOI: 10.1137/1.9781611972771.66.

Shyu, S. J. and Tsai, C.-Y. (2009). Finding the longest common subsequence for multiple biological sequences by ant colony optimization. *Computers & Operations Research*, 36(1):73–91. DOI: 10.1016/j.cor.2007.07.006.

Spearman, C. (1904). The proof and measurement of association between two things.. DOI: 10.1037/11491-005.

Spearman, C. (1906). Footrule for measuring correlation. *British Journal of Psychology*, 2(1):89. Available at:https://www.proquest.com/openview/8e47927e86a14cc15df27ede8444480f/1?pq-origsite=gscholar&cbl=1818401.

Sridharamurthy, R., Masood, T. B., Kamakshidasan, A., and Natarajan, V. (2018). Edit distance between merge trees. *IEEE transactions on visualization and computer graphics*, 26(3):1518–1531. DOI:

10.1109/TVCG.2018.2873612.

Wagner, R. A. and Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173. DOI: 10.1145/321796.321811.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., *et al*. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 10(7):1669–1681. DOI: 10.1038/ismej.2015.235.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83. DOI: 10.1007/978-1-4612-4380-$9_1$6.

Wilson, W. A. (1931). On semi-metric spaces. *American Journal of Mathematics*, 53(2):361–373. DOI: 10.2307/2370790.

Wing, C., Simon, K., and Bello-Gomez, R. A. (2018). Designing difference in difference studies: best practices for public health policy research. *Annual review of public health*, 39:453–469. DOI: 10.1146/annurev-publhealth-040617-013507.

Xiong, H., Brodie, M., and Ma, S. (2006). Top-cop: Mining top-k strongly correlated pairs in large databases. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 1162–1166. IEEE. DOI: 10.1109/ICDM.2006.161.

Yang, S., Schoeny, C., and Dolecek, L. (2019). Theoretical bounds and constructions of codes in the generalized cayley metric. *IEEE Transactions on Information Theory*, 65(8):4746–4763.

Zhang, Y. and Ge, G. (2015). Snake-in-the-box codes for rank modulation under kendall's τ -metric. *IEEE Transactions on Information Theory*, 62(1):151–158. DOI: 10.1109/TIT.2015.2502602.

Zhou, Y. and Liu, Y. (2014). Correlation analysis of performance metrics for classifier. In *Decision Making and Soft Computing: Proceedings of the 11th International FLINS Conference*, pages 487–492. World Scientific. DOI: 10.1142/$9789814619998_0$081.