Name: *Caleb McWhorter— Solutions*

MAT 397— Fall 2020
Applied Problems:
Dot Product

"Plagiarize! Let no one else's work evade your eyes! Remember why the good Lord made your eyes! So don't shade your eyes, but plagiarize, plagiarize, plagiarize—only be sure always to call it please 'research.'"
–Tom Lehrer, Lobachevsky

Document Comparison

A powerful tool in Mathematics—both pure and applied—is the notion of distance. In Mathematics, a distance function is called a *metric*. A metric on a set X is a function $d : X \times X \to [0, \infty)$, i.e. the function d takes two 'points' from X and gives the distance between them (hence why the image is nonnegative), and the function d is also assumed to satisfy the following properties:

(i) $d(x, y) = 0$ if and only if $x = y$

(ii) $d(x, y) = d(y, x)$

(iii) $d(x, y) \leq d(x, z) + d(z, y)$

The first assumption says that two points are distance zero apart if and only if they are the same point. The second assumption says that the distance from x to y is the same as the distance from y to x. The final assumption is the triangle inequality. An example of a distance function is the absolute value on $\mathbb{R}$: $d(x, y) := |x - y|$. If $x = 2$ and $y = 6$, then $d(2, 6) = |-4| = 4$, because 2 and 6 are a distance of 4 apart. Of course, the reader will also know the generalization to $\mathbb{R}^2$, $d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, and its generalization to $\mathbb{R}^n$ (the Euclidean metric):[1]

$$d\left((x_1, \ldots, x_n), (y_1, \ldots, y_n)\right) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

There are many other examples of distance functions: discrete metric, taxicab metric, graph metric, hamming distance, Riemannian metric, etc. Of course, you can define other notions of distance which may not satisfy the above properties. For instance, measuring 'street distance' does not satisfy (ii) if there are one-way streets.

However, the notion of distance (or metric) applies to more than physical objects. Instead, metrics measure 'distance' between any objects (in the abstract sense) so long as one can define a function with the properties above. For example, one might want to define a distance between two documents. This is exactly what is done in some search algorithms (especially for web search engines), plagiarism detection, etc. Obviously, the notion of distance should tell how different two documents are—documents which are distance zero apart are the same, documents whose distance is 'close' are very similar, and documents whose distances are 'far' apart are very different.

Fix some document D. Denote by $D[w]$ the number of times the word w appears in $D[w]$. For example, if the 'document' D is the sentence "Haters gonna hate, hate hate," we have

$$D[\text{haters}] = 1$$
$$D[\text{hate}] = 3$$
$$D[\text{hateraide}] = 0$$

[1] In fact, a lot of metrics come from a special process. Certain functions called inner products, $\langle a, b\rangle$, give rise to functions called norms via $\|w\| := \langle w, w\rangle$, and then norms are used to give distance functions via $d(x, y) := \|x - y\|$. The dot product is an example of an inner product. If you follow this process with the dot product, namely $\langle x, y\rangle := x \cdot y$, where $x, y \in \mathbb{R}^n$, then this process gives the Euclidean metric on $\mathbb{R}^n$.

Problem:

(a) One way of using 'word' counts and $D[w]$ to give a notion of document distance is as follows: let D_1, D_2 denote two documents you wish to compare. Define a function $\tilde{d}$ by

$$\tilde{d}(D_1, D_2) := D_1 \cdot D_2 := \sum_w D_1[w] \cdot D_2[w]$$

where the sum is taken over every 'word' in document D_1 and document D_2. Compute $d(D_1, D_2)$ if D_1 is the document "I love wine. You love wine." and D_2 is the document "I like cats, and I love dogs."

(b) Is the function $\tilde{d}$ from (a) a metric? [Hint: Suppose D_1, D_2 are entirely different and consider property (i) of a metric d.]

(c) Does the function from (a) measure 'commonality', i.e. how much in common two documents have? Explain.

(d) If you have a measure of commonality of documents, should the size of the documents matter? Explain why the size of the documents as inputs to the function $\tilde{d}$ from (a) matters, i.e. explain why $\tilde{d}$ is not scale invariant.

(e) To make the function $\tilde{d}$ size invariant, we divide by the 'lengths' of the documents

$$\frac{D_1 \cdot D_2}{|D_1||D_2|}$$

where by $|D_i|$ we mean $|D_i| := \sqrt{\sum_w D_i[w]^2}$, i.e. $|D_i|^2 = D_1 \cdot D_1$, and by $D_1 \cdot D_2$ we mean $D_1 \cdot D_2 := \tilde{d}(D_1, D_2)$. But then we recognize this as very nearly a specific geometric object. So instead, we define our final notion of document distance

$$d(D_1, D_2) := f^{-1}\left(\frac{D_1 \cdot D_2}{|D_1||D_2|}\right)$$

for some function f. Based on your knowledge of Calculus, find an appropriate function f. Explain what d is then measuring geometrically, then find $d(D_1, D_2)$ for the documents in (a).

(f) Based on the previous part, complete the following sentences: "If d is 'close' to (or exactly) _____, then the documents are nearly identical (or identical). If d is close to _____, then the documents are 'very' different."

Solution.

(a) There are a total of 8 different words to sum over. We have

$$\begin{aligned} D_1[\text{I}] &= 1 & D_2[\text{I}] &= 2 \\ D_1[\text{love}] &= 2 & D_2[\text{like}] &= 2 \\ D_1[\text{wine}] &= 2 & D_2[\text{cats}] &= 1 \\ D_1[\text{you}] &= 1 & D_2[\text{and}] &= 1 \\ & & D_2[\text{love}] &= 1 \\ & & D_2[\text{dogs}] &= 1 \end{aligned}$$

Then we have (working down the above columns left to right)

$$\tilde{d}(D_1, D_2) = \sum_w D_1[w] \cdot D_2[w] = 1(2) + 2(1) + 2(0) + 1(0) + 0(2) + 0(1) + 0(1) + 0(1) = 4$$

(b) The function $\tilde{d}$ from (a) cannot be a metric. Let D_1 be the document "A" and D_2 be the document "B". Then $\tilde{d}(D_1, D_2) = 1(0) + 0(1) = 0$ but D_1 and D_2 are not identical.

(c) We know that $\tilde{d}$ from (a) is not a metric, so it does not measure distance (as we have defined it). But it does give some measure of commonality. If two documents are entirely different, then $\tilde{d}(D_1, D_2) = 0$. But the more words the documents have in common, the larger $\tilde{d}(D_1, D_2)$ will be because it will involve less and less sums of products of a nonzero number with zero.

(d) This depends on the measure you wish to create. Perhaps, you want to consider five words in common for a ten word document to be a high degree of similarity (it is half the document) or a low one (its easier for this to be 'coincidence'). You may want to consider a single sentence in a large document in common with a sentence in another document to 'mean less' because of the size of the documents.

But certainly, the function $\tilde{d}$ is size dependent. Imagine two documents with all the same words (each occurring two times) and 25 words long. In this case, with the documents 100% similar, $\tilde{d}(D_1, D_2) = 100$ (because you have $2(2)+2(2)+\cdots+2(2)$). But imagine two documents that are 2% similar, each with 5,000 words. If each of the words in common to both documents occur once, then in this case we have $\tilde{d}(D_1, D_2) = 100$. In our two examples, the documents received the same score despite having very different levels of similarity. Therefore, the function $\tilde{d}$ is not scale invariant.

(e) If $\mathbf{v}, \mathbf{w}$ are vectors, then $\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}||\mathbf{w}| \cos\theta$. Hence, $\cos\theta = \dfrac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}||\mathbf{w}|}$. Notice this is very nearly identical to D_1, D_2. So we define

$$d(D_1, D_2) = \arccos\left(\frac{D_1 \cdot D_2}{|D_1||D_2|}\right)$$

Then d is measuring the angle between D_1, D_2 in 'word space.' For the documents D_1, D_2 in (a), we have

$$\begin{aligned} D_1 \cdot D_2 &= \tilde{d}(D_1, D_2) = 8 \\ |D_1| &= \sqrt{1^2 + 2^2 + 2^2 + 1^2} = \sqrt{10} \\ |D_1| &= \sqrt{2^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{12} \\ d(D_1, D_2) &= \arccos\left(\frac{D_1 \cdot D_2}{|D_1||D_2|}\right) = \arccos\left(\frac{8}{\sqrt{10}\sqrt{12}}\right) \approx 0.75204 \text{ radians} = 43.0887^\circ \end{aligned}$$

Generally, you would want to use this metric on both single words and sequences of words (taken two, three, or more at a time), and create a weighted average of the distances to measure document similarity. Of course, document comparison is still more complicated than this.

(f) If d is 'close' to (or exactly) $\underline{0}$, then the documents are nearly identical (or identical). If d is close to $\underline{\pi/2 \text{ radians}}$, then the documents are 'very' different.