Machine Learning
Exercises: vector-based document characterizations

Laura Kallmeyer

Summer 2016, Heinrich-Heine-Universität Düsseldorf

Exercise 1 *Consider the tf-idf weighting scheme from slide 9.*

Explain why we obtain $tf_{td}idf_t$ values 0 for terms occurring in all documents.

Solution:

We define the weight as

$$w_{td} = tf_{td}idf_t = tf_{td}\log\left(\frac{|D|}{df_t}\right)$$

For terms occurring in all documents, we have $df_t = |D|$ and therefore $idf_t = \log\frac{|D|}{|D|} = \log 1 = 0$, consequently the entire product is 0.

Exercise 2 *Consider the following 2-dimensional vectors $\vec{v}_1 = \langle 1, 2\rangle, \vec{v_2} = \langle 3, 6\rangle, \vec{v_3} = \langle 2, -1\rangle$.*

Calculate

1. *the normalized vectors (length 1) for $\vec{v}_1$, $\vec{v}_2$ and $\vec{v}_3$;*
2. *the Euclidian distances between $\vec{v}_1$ and $\vec{v}_2$ and between $\vec{v}_1$ and $\vec{v}_3$ without normalization;*
3. *the pairwise Euclidian distance of the corresponding normalized vectors (again between $\vec{v}_1$ and $\vec{v}_2$ and between $\vec{v}_1$ and $\vec{v}_3$); and*
4. *the cosine similarity, again for $\vec{v}_1$ and $\vec{v}_2$ and for $\vec{v}_1$ and $\vec{v}_3$.*

Solution:

1. $\frac{\vec{v}_1}{|\vec{v}_1|} = \langle\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\rangle = \langle 0.445, 0.89\rangle$

   $\frac{\vec{v}_2}{|\vec{v}_2|} = \langle\frac{3}{\sqrt{45}}, \frac{6}{\sqrt{45}}\rangle = \langle 0.445, 0.89\rangle$

   $\frac{\vec{v}_3}{|\vec{v}_3|} = \langle\frac{2}{\sqrt{5}}, \frac{-1}{\sqrt{5}}\rangle = \langle 0.89, -0.445\rangle$

2. $\vec{v}_1, \vec{v}_2$: $\sqrt{(3-1)^2 + (6-2)^2} = \sqrt{4+16} = \sqrt{20}$

   $\vec{v}_1, \vec{v}_3$: $\sqrt{(2-1)^2 + (-1-2)^2} = \sqrt{1+9} = \sqrt{10}$

3. $\frac{\vec{v}_1}{|\vec{v}_1|}, \frac{\vec{v}_2}{|\vec{v}_2|}$: $\sqrt{(0)^2 + (0)^2} = 0$

   $\frac{\vec{v}_1}{|\vec{v}_1|}, \frac{\vec{v}_3}{|\vec{v}_3|}$: $\sqrt{(\frac{2}{\sqrt{5}} - \frac{1}{\sqrt{5}})^2 + (-\frac{1}{\sqrt{5}} - \frac{2}{\sqrt{5}})^2} = \sqrt{\frac{1}{5} + \frac{9}{5}} = \sqrt{2} = 1.41$

4. $\frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1| \cdot |\vec{v}_2|} = \frac{3+12}{\sqrt{5}\sqrt{45}} = \frac{15}{\sqrt{5 \cdot 45}} = 1$

   $\frac{\vec{v}_1 \cdot \vec{v}_3}{|\vec{v}_1| \cdot |\vec{v}_3|} = \frac{2-2}{\sqrt{5}\sqrt{5}} = 0$