Cosine similarity - case study:

Task:

- find similarity of three sentences

Input data:

- tokenize the sentences

$$\begin{aligned} &sentence\_1 = [I\ love\ apples] \\ &sentence\_2 = [I\ hate\ apples] \\ &sentence\_3 = [I\ love\ oranges] \end{aligned} \tag{1}$$

Tokenization:

- tokenize the sentences

$$\begin{aligned} &set1 = [I, \quad love, \quad apples] \\ &set2 = [I, \quad love, \quad apples, \quad lemons] \\ &set3 = [I, \quad hate, \quad oranges] \end{aligned} \tag{2}$$

Vocabulary Corpus:

- put all unique tokens together in a corpus

$$set = [I, \quad love, \quad apples, \quad lemons, \quad hate, \quad oranges] \tag{3}$$

BOW - Bag of Words vectors

- create vectors

$$\begin{aligned} &bow_{vect_1} = [1, \quad 1, \quad 1, \quad 0, \quad 0 \quad 0] = A \\ &bow_{vect_2} = [1, \quad 1, \quad 1, \quad 1, \quad 0 \quad 0] = B \\ &bow_{vect_3} = [1, \quad 0, \quad 0, \quad 0, \quad 1 \quad 1] = C \end{aligned} \tag{4}$$

Cosine similarity

$$Cosine\ Similarity = \frac{\boldsymbol{A}.\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} \tag{5}$$

- dot product of **vectors**

$$\boldsymbol{A}.\boldsymbol{B} = a_0.b_0 + a_1.b_1 \ldots. + a_4.b_4 = \mathbf{1.1} + \mathbf{1.1} + \mathbf{1.1} + 0.1 + 0.0 + 0.0 = 3 \tag{6.1}$$

$$\boldsymbol{A}.\boldsymbol{C} = a_0.c_0 + a_1.c_1 \ldots. + a_4.c_4 = \mathbf{1.1} + 1.0 + 1.0 + 0.0 + 0.1 + 0.1 = 1 \tag{6.2}$$

$$\boldsymbol{B}.\boldsymbol{C} = b_0.c_0 + b_1.c_1 \ldots. + b_4.c_4 = \mathbf{1.1} + 1.0 + 1.0 + 1.0 + 0.1 + 0.1 = 1 \tag{6.3}$$

- magnitudes of **vectors**

$$\|\boldsymbol{A}\| = \sqrt{\boldsymbol{a_0^2} + \boldsymbol{a_1^2} + \cdots + \boldsymbol{a_4^2}} = \sqrt{\mathbf{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2}} = \sqrt{\mathbf{3}} \tag{7.1}$$

$$\|\boldsymbol{B}\| = \sqrt{\boldsymbol{b_0^2} + \boldsymbol{b_1^2} + \cdots + \boldsymbol{b_4^2}} = \sqrt{\mathbf{1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2}} = \sqrt{\mathbf{4}} = \mathbf{2} \tag{7.2}$$

$$\|\boldsymbol{C}\| = \sqrt{\boldsymbol{c_0^2} + \boldsymbol{c_1^2} + \cdots + \boldsymbol{c_4^2}} = \sqrt{\mathbf{1^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2}} = \sqrt{\mathbf{3}} \tag{7.3}$$

$$cos.sim.(AB) = \frac{\boldsymbol{A}.\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} = \frac{3}{\sqrt{3}.2} = \frac{3}{\sqrt{3}.2}.\frac{\sqrt{3}}{\sqrt{3}} = \frac{3.\sqrt{3}}{2.\sqrt{9}} = \frac{\cancel{3}.\sqrt{3}}{2.\cancel{3}} = \frac{\sqrt{3}}{2} = 0.86 \tag{8.1}$$

$$cos.sim.(AC) = \frac{\boldsymbol{A}.\boldsymbol{C}}{\|\boldsymbol{A}\|\|\boldsymbol{C}\|} = \frac{1}{\sqrt{3}.\sqrt{3}} = \frac{1}{\sqrt{9}} = \frac{1}{3} = 0.33 \tag{8.2}$$

$$cos.sim.(BC) = \frac{\boldsymbol{B}.\boldsymbol{C}}{\|\boldsymbol{B}\|\|\boldsymbol{C}\|} = \frac{1}{2.\sqrt{3}} = \frac{1}{2.\sqrt{3}}.\frac{\sqrt{3}}{\sqrt{3}} = \frac{\sqrt{3}}{6} = 0.289 \tag{8.3}$$

Angles between vectors:

$$\begin{aligned}
\theta_{AB} &= arccos\big(cos.sim.(AB)\big) = arccos(0.86) = 30.68° \\
\theta_{AC} &= arccos\big(cos.sim.(AC)\big) = arccos(0.33) = 70.73° \\
\theta_{BC} &= arccos\big(cos.sim.(BC)\big) = arccos(0.289) = 73.20°
\end{aligned} \tag{9}$$

Euclidean distance

- is another way how to measure the similarity

$$Distance = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \tag{10}$$