Fundamentals of Data Science

Semester B 20-21

Tutorial 2

1.a. $SMC = \frac{1+3}{8} = \frac{1}{2}$ and $J = \frac{1}{2+2+1} = \frac{1}{5}$

b. SMC considers both 1-1 matches and 0-0 matches as equally important. On the other hand, the Jaccard coefficient disregards 0-0 matches, and only regards 1-1 matches as important.

2 .We denote article 1 as **x**, article 2 as **y**, and article 3 as **z**.

(i) Since $\mathbf{x} \cdot \mathbf{y} = 36 + 75 + 20 = 131$, $\|\mathbf{x}\| = \sqrt{3^2 + 5^2 + 2^2} = 6.1644$, and $\|\mathbf{y}\| = \sqrt{12^2 + 15^2 + 10^2 + 8^2} = 23.0868$, the cosine similarity is calculated as follows:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{131}{(6.1644)(23.0868)} = 0.9205$$

(ii) Since $\mathbf{y} \cdot \mathbf{z} = 40$, $\|\mathbf{y}\| = \sqrt{12^2 + 15^2 + 10^2 + 8^2} = 23.0868$, and $\|\mathbf{z}\| = \sqrt{5^2 + 3^2 + 5^2} = 7.6811$, the cosine similarity is calculated as follows:

$$\cos(\mathbf{y}, \mathbf{z}) = \frac{40}{(23.0868)(7.6811)} = 0.2256$$

We observe that cosine similarity is able to measure the extent to which two articles are similar to each other. For example, from the keyword counts, it can be seen that the content of article 1 is similar to that of article 2. On the other hand, the content of article 3 is quite different from that of article 2, which is also reflected in the cosine similarity value between them.

3.a(i) $t'_{ij} = 2\log_2 \frac{8}{1} = 6$

(ii) $t'_{ij} = 3\log_2 \frac{8}{5} = 2.0342$

(iii) $t'_{ij} = 5\log_2 \frac{8}{8} = 0$

A word that occurs in only one of the documents corresponds to the maximum weight value, i.e. $\log_2 m$. On the other hand, a word that occurs in more documents has a lower weight value, and a word that occurs in every document has a weight value of 0.

b. This transformation reflects the observation that words which occur in many documents are less likely to be useful for distinguishing one document from another, while those that occur in only a few documents are more likely to be useful for distinguishing among the document classes.