Tutorial 1

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. a. Since $\mathbf{x}\cdot\mathbf{y}=6+30+2=38$, $\|\mathbf{x}\| = \sqrt{3^2+5^2+1^2+1^2} = 6$, and $\|\mathbf{y}\| = \sqrt{2^2+6^2+2^2+3^2} = \sqrt{53}$, the cosine similarity is calculated as follows:

   $$\cos(\mathbf{x},\mathbf{y}) = \frac{38}{6\sqrt{53}} = 0.87$$

   b. They are not necessarily identical. In this case, the two vectors point in the same direction, but they may have different lengths.

   c. The cosine similarity corresponds to the cosine of the angle between **x** and **y**.

3. a. A term that occurs in every document has a weight of 0. On the other hand, a term that occurs in only one of the documents corresponds to the maximum weight value, i.e. log *m*.

   b. This transformation reflects the observation that terms which occur in every document do not have any power to distinguish one document from another, while those that occur in only a few documents are more useful for distinguishing one document from another.