1. The difference between sampling scheme (a) and (b) is in how representative each scheme is. In (a), for each group we take an amount that is proportional with the group size compared to m, that is $\frac{m_i}{m}$. This means that the final n elements that are chosen will include a number of elements from each group proportional to the group size relative to all objects. The bigger a group is, the more elements from that group will be present in the final n elements chosen. On the other hand, sampling scheme (b) will be a simple sampling with replacement without taking in consideration group size.

2. The purpose of the inverse document frequency is to moderate the attention given to the common words such as “the”, “and”, “is”, etc. This allows for more unique words to be weighted higher across the corpus. The more documents that contain the word, the lower the value of $\frac{m}{df_i}$ and thus, the lower the importance of the word in a particular document.

3. Some of the issues that might be encountered is in how the data is represented. It is more difficult to calculate the similarity of vehicles when their attributes are represented as strings. To solve this, we can instead have integers represent each of the string values in the attributes. For example, for vehicle type, the number 1 can represent a sedan while number 2 can represent an SUV. This way, we don’t deal with string characters and will be able to compare vehicles more accurately.

4. The vectors [1 2 3 4] and [8 -15 2 4] are orthogonal. When we take the dot product between them, we obtain 8 – 30 + 6 + 16, which equals 0. Whenever the dot product is equal to 0, then the angle between them is 90 degrees.
   In their current form, the vectors are not orthonormal. However, we can make them orthonormal by multiplying the vectors by the inverse of their norm. The

norm of A is $\sqrt{1^2 + 2^2 + 3^2 + 4^2} = \sqrt{1 + 4 + 9 + 16} = \sqrt{30}$ and the norm of B is $\sqrt{8^2 + -15^2 + 2^2 + 4^2} = \sqrt{64 + 225 + 4 + 16} = \sqrt{309}$. The normalized versions of the vectors are now:

$$A = \begin{bmatrix} \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{30}} \\ \frac{3}{\sqrt{30}} \\ \frac{4}{\sqrt{30}} \end{bmatrix}, \qquad B = \begin{bmatrix} \frac{8}{\sqrt{309}} \\ \frac{-15}{\sqrt{309}} \\ \frac{2}{\sqrt{309}} \\ \frac{4}{\sqrt{309}} \end{bmatrix}$$