1 ex 1

1.1 a

The transformation in the case where a term appears in only one document the tf is multiplied by a large factor, where as in the case of all documents it is multiplied by log 1, which actually ends up taking down the value of tf, definitely not increasing it like in the case of a lower df.

1.2 b

The overall effect is that a term frequency of a word in a document is increased based on the inverse document frequency, the bigger the increase the less documents contain this term. The increase is logarithmic so tf still has the main role. The purpose of using this transformation is based on the fact that many words are common in various languages, so finding relevant or interesting words is made easier by this transformation.

1.3 c

If we for example want to look at the frequency a single customer has bought some product from a supermarket, we might want to do a similar transformation to take into account some very ordinarily bought products that can be found commonly from a majority of customers. An idf transformation here would favor those purchases that are more unique for the user in question, and we would get a better idea of what kind of ”user-specific” products are interesting to this person.

2 ex 2

2.1 a

Cosine similarity has a range of [-1, 1]. The values of $cos(x,y)$ will range from 0 towards 1 the larger the dot product of x and y is if nonnegative, and from 0 to -1 the more negative the product is.

2.2 b

No. Since the cosine similarity measures the cosine of the angle between the two vectors, x and y, the vectors need not be exactly the same for the cosine similarity $cos(x,y)$ to be equal to 1. Instead they just need to be parallel (can even point in opposite direction). However, if the cosine similarity $cos(x,y) = 1$, then we know that vectors x and y differ from each other only by a constant factor.

2.3 c

If we look at how cosine similarity is defined, it is basically the dot product of two given vectors, divided by the the product of the norms of the two vectors: $\frac{x^T y}{||x||||y||}$.

Correlation is defined as $\frac{cov(x,y)}{\sigma_x \sigma_y}$, where σ stands for standard deviation, and *cov* is the covariance between the given vectors. We can see by the definition of σ and the definition of *cov* that correlation is exactly the same as cosine similarity, but the data is ”centered”, which means that the mean of each column/feature is is deducted from each data point. If we do this as a preprocessing step to our data, the cosine similarity of this preprocessed data is exactly the same as the correlation of the original data.

2.4 d

For these types of vectors their variance is n times the sum of squared attribute values, and the correlation is dot product over n.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{\sum_{i=1}^{n} x_i^2 - 2x_i y_i + y_i^2} = \sqrt{1 - 2cos(x,y) + 1} = \sqrt{2(1 - cos(x,y))} \tag{1}$$

2.5 e

If all the values are standardized for both variables x and y, we now have a situation where both vectors have a mean of 0 and a standard deviation of 1. That means that the variance of these vectors is just n times the sum of squared attribute values. Also in this case the correlation between them is only their dot product over n: $d(x,y) = \sqrt{n - 2ncor(x,y) + n} = \sqrt{2n(1 - cor(x,y))}$.

3 ex 3

3.1 a

Let $\mathscr{A}$ be a set of more than one objects. The proximity among the objects in this set can be defined in many different ways. One of these is the maximum of the distance between two points in the set. Another might be the mean distance between any two points in $\mathscr{A}$. Here distance refers to some measure of distance for example euclidean distance.

3.2 b

The distance between two sets of points in euclidean space could be intuitively defined as the distance between the medoids (centers of mass) of these two sets. Another approach would be to define the distance as the minimum distance between any two points a and b, where $a \in A$ and $b \in B$.

3.3 c

As the minimum of all the proximity measures between points $a \in A$ and $b \in B$, or once again we could define this for the medoid of the sets. The first one seems more logical to me though.