Aaron Morgenegg A02072659

CS 5665 Homework 2

Tan 2.2

(a) Continuous(Time is continuous and can be measured in infinitely smaller increments), quantitative, interval
(b) Continuous, quantitative, ratio
(c) Discrete, qualitative, ordinal
(d) Continuous,quantitative, ratio
(e) Discrete, qualitative, ordinal
(f) Continuous, quantitative, ratio
(g) Discrete, quantitative, interval
(h) Discrete, qualitative, ordinal
(i) Discrete, qualitative, ordinal
(j) Discrete, qualitative, ordinal
(k) Continuous, quantitative, ratio
(l) Continuous, quantitative, ratio
(m) Discrete, qualitative, nominal

Tan 2.8

A document-term matrix is asymmetric, because only non-zero values matter, and discrete, because the values are either 0 or non-zero. In a sense, it is also binary, because you can represent any non-zero value as a 1.

Tan 2.10

Precision of a measurement is the closeness of repeated measurements to each other.

Single precision means that the floating point data is represented as 32 bits in the computer, while double precision means that the floating point data is 64 bits. The details vary based on the encoding system, but for IEEE,
single precision -> 23 bits of significand, 8 bits of exponent, and 1 sign bit.
double precision ->52 bits of significand, 11 bits of exponent, and 1 sign bit.

Tan 2.11

1. Text data is much easier to work with programatically.
2. Text data is can be read by humans more easily.

Tan 2.12

(a) Noise is undesirable, and is caused by measurement error. Outliers may be useful or interesting, such as credit card fraud.
(b) Noise may resemble outlier data, but it is caused by measurement error and should not be analyzed.
(c) Noise objects are not always outliers.
(d) Outliers may not always be noise objects, if a dataset has no noise for example.
(e) Yes, noise may change the shape of the data, and make it seem as if an outlier fits into a pattern.

Aaron Morgenegg A02072659

Tan 2.14

You could use the simple matching coefficient(SMC) to group the elephants attributes by whichever value you want, such as weight and tusk length.

Tan 2.19

(a) cos(x, y) = 1, corr(x, y) = null, d(x, y) = 2

(b) cos(x, y) = 0, corr(x, y) = -1, d(x, y) = 2, j(x, y) = 0

(c) cos(x, y) = 0, corr(x, y) = 0, d(x, y) = 2

(d) cos(x, y) = .75, corr(x, y) = 0.25, j(x, y) = j(x, y) = 1

(e) cos(x, y) = 0, corr(x, y) = 0

Tan 2.23

You could take the inverse of the similarity measure.

Tan 2.24

(a) Simple matching coefficient, Jaccard coefficient.

(b) Calculate the distance using the distance formula.

(c) Use one of the proximity measures, such as the Jaccard coefficient.