CMT-209 Informatics

1. What types of folksonomies are there? Give an example of each type and discuss the commonalities and differences of the different types.

a. Broad folksonomy: a large number of people tag the same item. For example, del.icio.us
b. Narrow folksonomy: a smaller number of people tag the same item. Usually, they use for later personal retrieval. For instance, Flickr.

Similarity: all of their users can use their own words to describe the tag.
Difference: BF could be used for pick the prefer term or extract the controlled vocabulary, NF normally would be used for later personal retrieval.

2. Which relationships are used in a thesaurus? Explain and illustrate with an example.

a. Equivalence: synonymies with ability to suggest which term is the prefer term
b. Hierarchy: a broad class for term; broad term and narrow term
c. Associative: relationship across hierarchies (related terms)
d. Scope note: defines terms or breadth of term and its usage

Example:
Legal service centres
SN: a place serving the public which may provide a variety of services including ……
BT: legal service providers
NT: clinics
NT: legal information centre
RT: legal services

3. Define the Jaccard similarity and illustrate it with an example.

Jaccard similarity of sets A and B is defined as the ratio of the size of the intersection of A and B to the size of their union.

sim(A, B) = |A ∩ B| / |A ∪ B|

4. Formally define distance and give an example of a distance measure.

distance is also a measure of how close to each other two instances are. The closer the instances are to each other, the smaller is the distance value.
sim(A, B) = 1 / d(A, B) or sim(A, B) = 1 / (d(A, B) + 0.5)
d(A, B) = 1 – sim(A, B)
d(A, B) = $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$

5. Use the Wagner-Fischer algorithm to compute the edit distance of [two words given].

6. What are the two basic types of clustering? Discuss their commonalities and differences.

7. What is a decision tree?

8. Use Hunt's algorithm with classification error to construct a decision tree for the data in the table below. [table of data given]

9. Figure 1 shows a decision tree, Table 2 lists test data. Define the three evaluation measures most commonly used in classification and use them to evaluate the decision tree on the test data.

10. What is the RDF data model? What is an RDF statement? Illustrate with an example.

11. What is stemming? What is lemmatisation? Give one example where they produce the same result, and one example where they produce different results.