There are 4 questions in total; you have to answer 3 of them.

Each question covers one broad topic, and typically asks several sub-questions.

The exam tests two types of knowledge:

Factual knowledge as covered in class / on the slides

Procedural knowledge, i.e., understanding and knowing the algorithms and evaluation techniques seen in class and being able to apply them, as seen in class and in the additional exercises on learning central

Below are some examples of the general kinds of sub-questions you should expect. This list is not exhaustive.

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

Week3\_2 44 45

1. broad folksonomy: the result of many people tagging the same item, e.g. del.icio.us

2. narrow folksonomy: the result of a smaller number of people tagging items for later personal retrieval or for their own convenience, e.g. Flickr

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

Define the Jaccard similarity and illustrate it with an example.

Week3\_3 19

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

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

Week4\_1 15

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

Week4\_1 38

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

Hierarchical:

clusters are permitted to have subclusters

a set of nested clusters that are organised in a tree

each node (cluster) in the tree (except the leaves) is the union of its children

the root of the tree contains all objects

Partitional:

It is a simply division of the set of objects into non-overlapping subsets.

Hierarchical clusters can be view as a sequence of partitional clusters.

It could be obtained by cutting the tree at specific level

Week4\_2 12

What is a decision tree?

Week5 37

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

Week5 38-60

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.

P = TP/(TP+FP) = 6/(6+4) = 0.6

R = TP/(TP+FN) = 6/(6+1) = 0.8751

F = 2PR/(P+R) = 0.7059

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

RDF Model

Represent infoinformation as directed graphygraphs with labellabelled nodes and arcs

Is designed for information integration

Multiple sources

Heterogeneous structure

Different schemata

RDF statement

Triples: subject predicate and object

Mirror the structure of a simple sentence

Using URIs

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