UCT CSC303 2005 :: XML/IR + ToA :: Test [35 marks]

Answer Questions 1, 2 AND 3.

Question 1: XML / XSLT [10]

Answer the following questions based on this piece of XML:

```
<mdata>
   <name>CSC303 Test</name>
   <author>
      <first_name>hussein</first_name>
      <last_name>suleman</last_name>
   </author>
</mdata>
```

Assume that the **name** and **author** elements are both infinitely repeatable and optional and that **first_name** and **last_name** must both appear exactly once each.

1. Write an XML Schema complexType type definition **mdataType** corresponding to the content of the **mdata** element and its descendents. [5]

2. Write an XPath expression that locates the **first_name** node corresponding to the first **author** node (assuming there could be multiple authors). The current context node is the root node **mdata**. [1]

3. The given XML document is well-formed - what 2 properties make it well-formed? [2]

4. If exactly one bit of a well-formed XML document is corrupted, how does the self-segregrating nature of UTF-8 prevent cascading errors. [2]

Question 2: Information Retrieval [10]

1. Inverted files can be used to optimise filtering. What are inverted files and what is filtering? [2]

2. In a typical inverted file, how can we reduce the number of bytes required without reducing the information content? [1]

3. Discuss 2 different storage approaches for inverted files. [2]

4. Inverted files only produce a list of exact matches, which is not always enough – discuss one technique to increase recall. [2]

5. After filtering documents, we can rank them on the basis of links. Describe the simple PageRank algorithm used for this purpose. [3]