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 XSLT template to transform the XML fragment into:

```
<dublin_core>
   <title>CSC303 Test</title>
   <creator>hussein suleman</creator>
</dublin_core>
```

Assume that the values of the content of the **name** and **author** nodes may differ from one document to another. Assume the source namespace prefix is *source* and the destination prefix is *dest*. Use the following as a starting point. [5]

```
<xslt:template match=”source:mdata“>
. . .
</xslt:template>
```

```
<xslt:template match=”source:mdata”>
  <dest:dublin_core>
  <xslt:for-each select=”source:name”>
    <dest:title>
      <xslt:value-of select=”.”/>
    </dest:title>
  </xslt:for-each>
  <xslt:for-each select=”source:author”>
    <dest:creator>
      <xslt:value-of select=”first_name”/>
      <xslt:text> </xslt:text>
      <xslt:value-of select=”last_name”/>
    </dest:creator>
  </xslt:for-each>
```

</dest:dublin_core>

</xslt:template>

Minus one mark for each major error.

Note that the for-each statements could be replaced by apply-templates – then additional templates need to be defined for name and author.

2. Write code that uses the DOM API to access the contents of the **first_name** node within the first **author** node and store it into the *fname* (or *$fname*) variable, given that the document has been parsed and assigned to the *top* (or *$top*) variable.

Note: The sequence of commands is important, not the programming language. [2]

$fname = top->getDocumentElement->getElementsByTagName ('author')->item(0)
->getElementsByTagName ('first_name')->item(0)->getFirstChild->getData

-1/2 for each error in the method call sequence.

3. Why do we need namespaces in XML documents, such as the XSLT template created in the previous question? [1]

to disambiguate tags in different contexts

4. It is possible to create XML documents with multiple namespaces but no defined namespace prefixes. How? [2]

by defining a default namespace at every tag that is not in the same namespace as its parent [2]

Question 2: Information Retrieval [10]

1. If we always look at only the first 20 documents in a result set, ranking of documents serves to increases precision. What is ranking and what is precision? [2]

Ranking is the ordering of result sets according to the estimated relevance.

Precision is the number of documents in the result set that are relevant.

2. Discuss 2 techniques to increase the precision of an IR system. [2]

use LSI to eliminate noise and find stronger correlations among documents [1]

use stopping to remove very common words that do not add meaning [1]

use only the metadata and not the full text when searching [1]

etc.

3. When ranking documents, it is common to use logarithmic functions in the ranking formulae – what is the purpose of this non-uniform scaling of values? [2]

to ensure that the initial occurrences of a term impact on the rank, but subsequent occurrences do not have as much impact as the first ones [2]

4. Describe the simple HITS algorithm. [4]

step1: Assign arbitrary hub and authority values to all nodes such that each sum=1 [1]

step2: calculate new authority for each node as the sum of hub values from all incoming links. calculate new hub for each node as the sum of authority values from all outgoing links. [1]

step3: normalise hubs and authorities so each sums to 1. [1]

step4: iterate step2/3 until a steady state is reached [1]