Lab 3: Content based analysis of data

Two tasks involving different content analysis and comparison methods.

Part1: Image comparison

Take the zipped data file and extract the 12 images it contains.

Write a program to load each image (suitable image loader libraries can easily be found on the web) and analyse it to produce a feature vector. Suitable features might include:

1. Colour content
2. Colour distribution around the central point
3. Colour distribution around several points
4. Luminance distributions around one or several points
5. Edge positions and orientations
6. anything else that you can think of

Pick one image from the set and use comparison of the feature vector to rank the other 11 in order of similarity to your chosen image.

Things of interest for your report: how did you make the feature vector comparison? Did you apply any weighting of features over one another? Why?

Part 2: Text comparison

Take the zipped file and extract the 10 text files it contains.

Two of the ten files contain plagiarised content from one of the other files. In one it is a single sentence, in the other an entire paragraph.

Write a program to analyse the 11 texts for textual content and then use comparison to identify the plagiarised sections.

Tips:

1. Remove the punctuation using a linear replacement and converting the sentences into single lines of text. Also convert everything to upper or lower case.
2. Analyse the files to create a word list (a hashed dictionary) replacing the words with numerical values of the word positions in the list.
3. Linked lists are the ideal mechanism for creating this dictionary
4. Compare the numerical sentence sequences to identify the copied text and the source and destination files for each copied element.

The final step may be quite a lengthy process. I'd be interested to know the computation time for your particular hardware.