10 11 2016 **SEWN 2016 labsheet 4 MSc version (assessed coursework)**
Calculating PageRank

This labsheet follows logically from labsheet 3 where you are required to use the adjacency matrix you produced in that assignment to calculate PageRank for each page in the matrix.

Any MSc student attempting the BSc version of this labsheet may do so only at the discretion of the module's lecturer or TA.

The maximum grade achievable for this labsheet is 100%.

1. Familiarise yourself with the following formula for PageRank, and how it is iteratively calculated:

$$PR(W) = \frac{T}{N} + (1-T)\left(\frac{PR(W_1)}{O(W_1)} + \frac{PR(W_2)}{O(W_2)} + \ldots + \frac{PR(W_n)}{O(W_n)}\right)$$

   You may want to check that you understand the (slightly simplified) example PageRank calculation from the *Link Analysis* lecture slides. Further information concerning PageRank is available at:

   - http://www.webworkshop.net/pagerank.html.
   - http://pr.efactory.de/e-pagerank-algorithm.shtml

2. You are required to use your representation of the adjacency matrix from labsheet 3 to compute the PageRank for each page included in the matrix.

   For any page in the matrix, in order to do this, you will need to determine the set of pages that link to it, i.e. *inlinks*, and the set of pages that it links to, i.e. o*utlinks*.

   Your solution should calculate and output the PageRank for each page in the matrix, using the teleportation constant T = 0.15, the value allegedly used in Google's own calculations.

   You will need to decide on a suitable number of iterations to run the calculations in each case: think carefully about your criteria for convergence before you terminate your calculations.

   Your solution's output should be to a file called `pagerank015.txt`, of the form:

```
Iteration <iteration number>
   <URL1> <PageRank1>
   <URL2> <PageRank2>
   ...
```

   You are free to use any demonstrable means to calculate PageRank.

3. Re-run your solution, this time with T = 1.0 and output the results to a file called `pagerank100.txt`. What do you notice about the time to convergence and the PageRank values in this case?

4. This final part of the lab is **optional** but carries extra credit.

   Run your PageRank calculations again using different teleportation values:

   T = 0.00 T = 0.25 T = 0.50 T = 0.75

   Analyse your results (including those for T=0.15 and T=1.00):

   a. Do the results converge to the same value of PageRank no matter what the teleportation probability is?
   b. What difference does the teleportation probability make to the number of iterations necessary?
   c. Can you see any reason to choose 0.15 as a *standard* value for teleportation?
   d. Is there anything else interesting you can see in the pattern of results?

What to hand in:
Submit a single `.zip` file via the **Labsheet 4 Calculating PageRank** drop box in Moodle containing:

1. The representation of your adjacency matrix from labsheet 3 in order to run (2) below.
2. Any program code, or equivalent implementation, and any special instructions for compiling and running your solution (if there are several code files, please include the folder structure in your `.zip` file), should be supplied. Depending upon your solution, you may be required to demonstrate your solution in one of the PC laboratories at the college.
3. The outputs of your program, i.e. (i.e. the files named `pagerankXXX.txt`), showing the results of your PageRank calculations for (3).
4. A short A4 pages (`.doc(x)` or `.pdf` format) report of no more than two pages explaining how your solution works, and outlining your observations from (4) (if you do this part of the labsheet).
5. Any references to books or on-line material for this labsheet should be included in your report.

Submission deadline: 01 12 2016

Late assignments: No extensions are available for this lab and any late submissions will be graded as per the guidelines of the relevant course being studied.