Big Data Analytics Laboratory

In big data applications, the input to e.g. search problems often does not fit in a single file, or even on the hard disk of a single node in your network. In those cases you will have to involve all nodes in the search algorithm. They can each come to their own preliminary conclusions, but this needs to be followed by a synchronisation or reduction step where different nodes put together the results they obtained individually.

In this assignment you will need to write a data parallel programme in matlab that counts the number of occurrences of your University account name (don't worry, we left out the passwords) over a set of large files.

The starting point is the following sequential programme:

```
id = labindex;
fid = fopen(['input' 1 '.txt']);
A = textscan(fid,'%q');
numtimes = sum(ismember(A{1},'aa'));

totaltimes = numtimes;
fprintf('Total number of occurences: %d\n', totaltimes);
```

Generalise this to a parallel programme. To get you started:

- A cluster of parallel pool can be started by issuing the `parpool` command.
- In matlab `spmd` blocks are executed on all nodes in a parallel pool. Variables populated in an spmd block outlive that block as composite values that can be accessed as cells (e.g. `numtimes{3}`).

Programming is 90% reading the manual! Read the docs for: `parpool, spmd, Composite`! Refer to the Parallel Toolbox Documentation for examples.