Description

A system that takes for input a list of URLs to photos containing text and executes OCR (Tesseract) on them.

The system is distributed and contains 3 parts:

A *Manager* (server) and a *Worker*, designed to work on AWS EC2.

A *Client* (Local), designed to work on a local computer.

Client sends a message to *Manager*, asking it to execute OCR on a list of URLs (that *Client* uploaded to AWS S3), *Manager* distributes the work to several *Workers*, who execute the OCR. *Manager* returns the results to *Client* via AWS S3.

Manager is a server based on the Reactor server with few adjustments, and can serve several *Clients*. The scalability of *Manager* was the most important property in its implementation.

Manual

Client requires 3 input values:

1) The path to the input file, containing the list of URLs
2) The wanted path to the output file
3) The maximum number of URLs to be handled at once by a *Worker*

Optional: write the word "terminate" in order to terminate *Manager* after the execution of current *Client*.

At the end of the process the output will appear as a html file in the current directory.

Important: *Client* requires the following files in current directory in order to operate:

1) Manager.jar
2) Worker.jar
3) Tessdata.zip