text-extraction-service

TES is a simple Go service for extracting and storing textual content from PDF and office documents.

Status

This started as an exercise in using Golang and cgo. But it is about to be used in production (at least for PDFs). The use case is the fast processing of binary documents for repeated search machine indexation (see blow for details).

The RegEx-based RTF parser is rather inefficient.

The parser for XML-based office formats is not very sophisticated and might need more testing.

Apache Tika is definitively a more versatile and mature solution to be considered.