Spá: a web-based viewer for PDF text mining

Kuiper, J. Marshall I.J. Wallace, B.C. Swertz M.A.

April 11, 2014

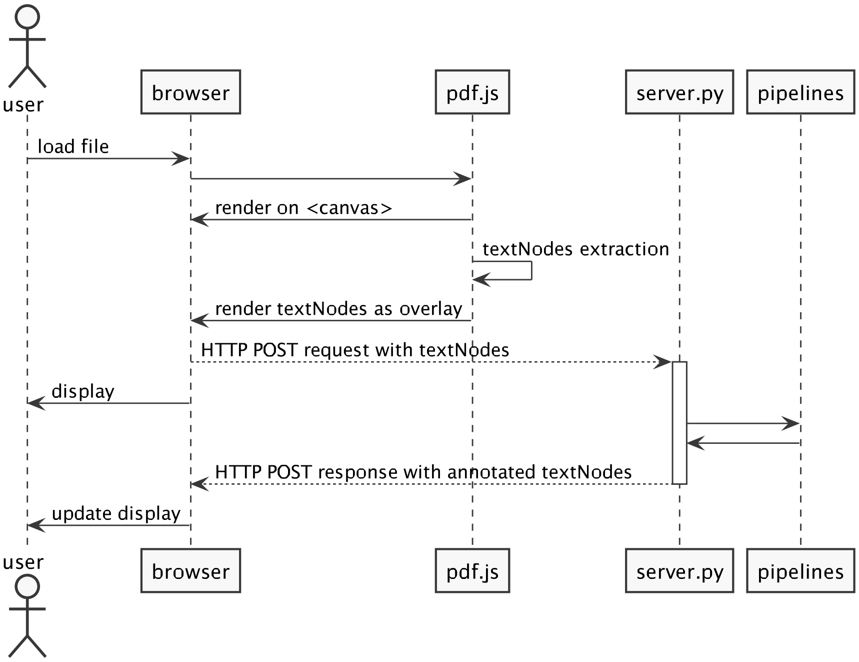
Abstract

When doing sentence extraction or document level predictions on unstructured text the results of trained models are often hard to interpret within the context of the documents at hand. Furthermore presenting results to end users can be a considerable user interface challenge. To this end we present Spá, a generic web-based visualizer for sentence and document level classifiers on unstructured PDF documents. Spá allows the results of sentence extractions to be visualized within the PDF document itself, and allows other results to be presented alongside it.

1 Introduction

Finding sentences or words with particular characteristics within a larger document is an important task in natural language processing and machine learning. For example, one may wish to identify the most important sentences in a document to automatically generate a summary, or extract words that match a certain ontology.

2 Architecture



3 Case Study

4 Discussion & Future work