Welcome to DSH

Marcelo Riss

2017-12-05

1 Document Smart Hightlights

Document Smart Hightlights aims to provide a set of web services to allow uploading of PDF and HTML files and return a list of keywords and most relevant sentences. This system applies state of the art algorithms, using NLP techniques to produce both the list of keywords and most relevant sentences. This last one, using automatic document summarization techniques.

1.1 Overall Process Description

The overal process of file submission and keywords/relevant sentece extraction is described as follows:

1.1.1 File Submission

A web service is established in a way that:

- User uses a web service call to submit/upload a file.
    - System sumit the file to process assynchronously and return a token to the user
- User uses a web service call to keep probing the status of file processing, using the token previsously returned.
- When the file processin is done, the previous call return success and user uses another call to retreive the results having two parts:
    - A list of keywords
    - A list of most relevant sentences

1.1.2 File Indexing

- The first processing task is to submit files to a database where it will be indexed and processed to be able to identify individual terms or tokens inside the file context. This is used submiting the files to be stored at a database provided by Apache SOLR.

1.1.3 File Keyword Extraction

- The keyword extraction is executed by setting scores to each term in the document's text. This is using a TF/IDF approach through several different implementations of TF and IDF, as shown at this paper.
- Those terms having the best scores obtained by a meta-algorithmic combination of all TF/IDF combination, will then be returned as the best ranked keywords.

1.1.4 Most Relevant Sentences Extraction

- This is achieved appling typical automatic summarization techniques like extractive summarization or keyphrase extraction.

- The sentence relevance is calculated taking into account two main criteria:
    - Similarity of the sentence with the title: using the algorithm provided from this paper
    - Multiple combinations of the TF/IDF of each term of the sentence, using the keyword extraction mechanism described above.
- The sum of two criteria will define the score. The best scored sentences are returned aside with their respective paragraph numbers.