Automated Decision Making System for Processing Large PDF Files

Degree programme: Master of Science in Engineering | Specialisation: Information and Communications Technologies
Thesis advisor: Prof. Andreas Habegger
Expert: Dr. Souhir Ben Souissi (Berner Fachhochschule)

Preparing a patient history for a medical expertise involves the analysis of huge amounts of records of various origins. To date, the entire case file - up to 5‘000 pages - is printed, then analyzed and sorted by hand to create a chronological dossier containing only the relevant pages which amount to about 10% of the entire case file. An Automated Decision Making (ADM) System should now be developed to assist in compiling patient‘s medical history.

Approach

A selection of common case files are analyzed by hand to determine the structure and available information within the PDF files. Based on this analysis, following Metrics characterizing a page are defined and the approach for extracting them is established: Relevance (is the page relevant, should it be included?), Keywords (keywords summarizing the page), Coherence (which pages form an in of itself coherent document), Pagination (page number and number of pages), Date (when was the page created?), Duplicates (which pages are duplicates of one another?) Persona (who was involved in the correspondence?).

Forensic Image Hashes

A prediction on the Relevance of a page is made based solely on the visual aspect using gray scale images of the PDF page. However, the image is reduced to a forensic image hash to reduce the complexity of the classifiers and to maintain the privacy of patient records in case external infrastructure is used. A hand full of common classifiers are compared to evaluate the expedience of forensic image hashes in a classification environment using supervised learning.

In addition, an analysis of the similarity of the image hashes for pages of the same class and pages of different classes is performed to gain insights into the data set and to provide context for the resulting model.

Name	Recall	Percision	Accuracy	MCC	F1
RandomForest	0.820	0.871	0.833	0.689	0.823
Lin. SVC	0.818	0.817	0.818	0.635	0.817
AdaBoost	0.835	0.836	0.837	0.671	0.836
NNSimple	0.865	0.871	0.868	0.735	0.866
NNComplex	0.874	0.877	0.877	0.751	0.875

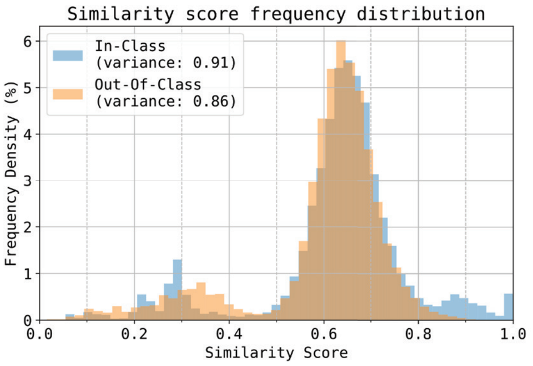
Performance values for relevance prediction of different classifiers based on 1024 bit average hashes

Forensic image hashes are further used to detect Duplicates, as is the common use case for forensic image hashes. For this purpose, each page is compared with one another and assessed based on the similarity of the forensic image hash. Only in cases where the hash similarity is ambiguous is an additional comparison of the text performed.

Lukas Alexander Studer

Results

The extraction of the Coherence, as well as the characteristic Date and the Pagination of pages was possible through the use of embedded text and meta information without the need for machine learning. Preliminary results of the detection of Duplicates and suggestions on the Relevance of pages using forensic image hashes proved to be highly successful with accuracies of up to 99% and 87% respectively. This proves that forensic image hashes can be used to detect page duplicates and classify pages using supervised learning. Further, with the overlapping frequency distribution of the similarity score, it can be said that the model is not trivial and is not based on a threshold. The forensic image hash therefore preserves the essence of the page, which can be used as a basis for predictions. However, the extraction of Keywords and correspondences between Personas was found to be generally unsuccessful.



Frequency distribution of the similarity score in-class and out-of-class samples based on 1024 bit average hashes