Technologie Recherche PDF-Box

Funktionsweise

Das Tool PDF-Box liest englische oder deutsche PDFs ein und gibt Schlüsselwörter für die jeweilige PDF aus. Die gefundenen Schlüsselwörter beziehen sich auf die gesamte PDF.

Die Analyse des Strings erfolgt mit der Open Source Software Hunspell (GPL). Diese Nutzt zur Analyse ein eigenes Wörterbuch und filtert unwichtige Wörter heraus. Jedes gefundenen Schlüsselwort bekommt ein Attribut ‚Weight' was die Relevanz des Wortes angibt. Dieser Wert liegt zwischen 0,0 und 1,0. Die Schlüsselwörter werden am Ende nach der Häufigkeit des Auftretens in der PDF sortiert und in dieser Reihenfolge in eine ArrayList geschrieben, welche dann zurückgegeben wird.

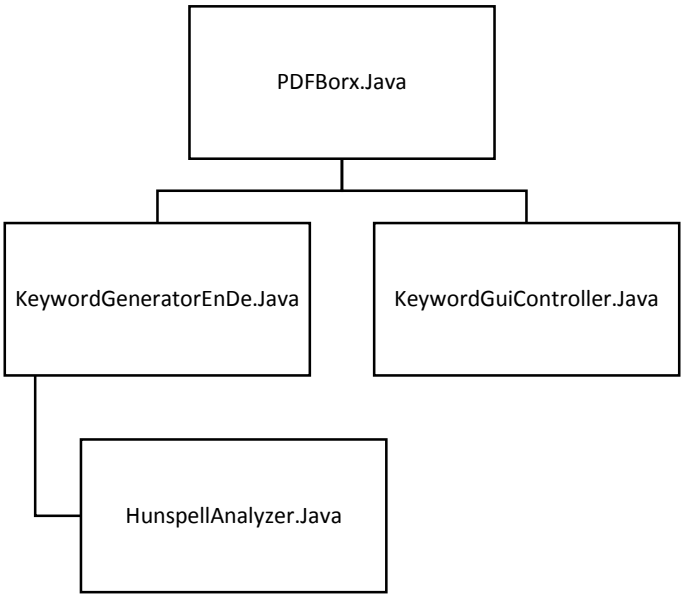
Beispiele:

- Eingabe: Technologie Recherche Themen PDF (von Mastern)
- Ausgabe: ii, git, themen, dokument, fu, projekt, recherche, technologie, gitlab, tool, tim, delle, genutzt, software, softwareprojekt, pdf, folgen

- Eingabe: Kognitive Suche Aufgabenstellung PDF (von Prof. Weicker)
- Ausgabe: suche, ko, software, pdf, einsatz, fu, ausgabe, chst, mo, ergebnis, einzeln, wahl, nnen, web

- Eingabe: Algorithmen und Datenstrukturen E-Book (von Prof. Weicker)
- Ausgabe: algorithmus, element, fur, knoten, baum, feld, schlussel, suche, wert, laufzeit, list, index, bild, link

Aufbau



In der Haupt-GUI Klasse ‚PDFBorx.Java' wird sowohl die Klasse zur Generierung der Schlüsselwörter (KeywordGeneratorEnDe.Java) als auch die Klasse zur Ausgabe der Schlüsselwörter (KeywordGuiController.Java) aufgerufen. Zur Generierung wird jedoch noch die Klasse HunspellAnalyzer.Java aufgerufen welche wiederum auf verschiedene Libaries zur Verarbeitung des Strings zurückgreift. Die KeywordGeneratorEnDe.Java Klasse hat als Rückgabewert die ArrayList mit den Schlüsselwörtern.

Mögliche Schnittstelle

Das Abfangen der Schlüsselwörter wird sich voraussichtlich am leichtesten über die Klasse KeywordGeneratorEnDe.Java erreichen lassen, da diese die Wörter in einer einfachen ArrayList zurückgibt. Zu beachten ist dabei, dass natürlich nur Schlüsselwörter von PDFs erzeugt werden können, welche zuvor von der Software eingelesen worden sind.