300579 Professional Experience (PX) Spring 2019
SCEM - Western Sydney University

Group Number	KS1912
Title	Extraction of structured information from dictionary PDFs
Supervisor Details	Jonathan Weightman <jono@lymlive.com.au>

Description

For my linguistics ARC project, we work a lot with pdf files containing dictionaries of foreign languages. Dictionaries tend to be structured in a very regular way (headword, translation, part of speech, sometimes also example sentences).

We usually run these through OCR software, and then search them for words we are interested in, but the process is slow and we can only search one dictionary or one word at a time.

The solution we are looking for would be a script that performs OCR on the pdf, and also then extracts the information from the output into a structured format like an xml file or an sql file. It would be ideal if we could run the script on multiple dictionaries - e.g. it could take a folder of pdfs, and return an sql or xml output with the content.

A nice-to-have feature, but not essential, would be that the user could run the script with a parameter that works as a kind of filter, e.g. I run it on five dictionaries, but I only want it to extract the content for words where the translation or description includes the word "tree". (E.g. to get all words from the five dictionaries that relate to trees in some way).

Requirements

Minimum requirements:

- script takes input of one already-OCRed pdf file (a dictionary)
- uses regex or other fuzzy search approaches to parse the structure of the file
- returns the words from the dictionary in a structured format (xml, csv or sql) structured into "headword", "part of speech", "translation", "example sentences", "other info"

Non-essential features that would be nice to have:

- script works on files that have not already been OCRed, i.e. runs the files through OCR as part of the pipeline
- script can take multiple pdf files as input
- user can pre-select certain filters so that only a subset of the total dictionary information is returned
- a simple GUI so it can be used by members of my team who are scared of the command-line

Suggested Technologies	To be identified by students
Client Details	Rachel Hendery <r.hendery@westernsydney.edu.au> Ph: 0466586774

Student Group	18667471 Roger Leo Bernardo <18667471@student.westernsydney.edu.au> 18993745 Joshua Dib <18993745@student.westernsydney.edu.au> 17593713 Josephine Paculio <17593713@student.westernsydney.edu.au> 19595913 Caleb Stephen Smith <19595913@student.westernsydney.edu.au>