vector.ai

Overview

“Computer! Sort Out my PDF”: Build an ML model that identifies the structure of a pdf containing multiple documents

Project Description

Business documents are central to the operation of business. Such documents include sales agreements, vendor contracts, mortgage terms, loan applications, purchase orders, invoices, financial statements, employment agreements and a wide many more. The information in such business documents is presented in natural language, and can be organised in a variety of ways from straight text, multi-column formats, and a wide variety of tables. Understanding these documents is made challenging due to inconsistent formats, poor quality scans and OCR, internal cross references, and complex document structure. Furthermore, these documents often reflect complex legal agreements and reference, explicitly or implicitly, regulations, legislation, case law and standard business practices. At Vector AI we have developed an ML pipeline to read, understand and interpret business documents. However, before being able to do that, documents come in bulk in a pdf and we need to sort them out. This project involves classifying pdf pages into type of documents and understand how documents are sorted in the pdf.

You will need to extend our ML service which aims to classify and paginate documents in a PDF using computer vision and natural language processing techniques.

Who we are looking for

For these projects you should be able to program in Python and have an interest in developing your ML skills.

Supervisor Profile

Name: Dr Aldo Lipani
Current Position: Senior Machine Learning Researcher
Linkedin: https://uk.linkedin.com/in/aldo-lipani

Reading Material

LayoutLM: Pre-training of Text and Layout for Document Image Understanding:
https://arxiv.org/abs/1912.13318