vector.ai

Overview
Perform unsupervised clustering of documents to detect similarities.

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, when a new customer comes, before being able to read, understand and interpret their documents, we need to understand how similar their documents are with respect to those used for training by our ML pipeline.

You will need to explore datasets of pdf documents (images) using unsupervised learning techniques like clustering and anomaly detection methods.

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
t-SNE:
- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
- https://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf