Peer review for “California Police Records Access Project.”

Hyeong Seok Oh

Summary

This paper presents a novel data pipeline to process police records from the California Police Records Access Project (PRAP) that are stored in various types of documents. First, the authors extract text using a combination of direct PDF text extraction (PyMuPDF) and OCR (PyTesseract), and extract 113 out of 184 documents that had sufficient text (>50 characters). Then they employ LayoutLMv3 to create document embeddings, combining the text and bounding boxes, and average token embeddings to get one vector per document.
They also try and evaluate several text-only sentence embedding models. On top of these embeddings, they experiment with different unsupervised clustering methods (k-means, Gaussian mixtures, DBSCAN, HBDSCAN) to group documents by type or content. They evaluate clustering with silhouette score, and Calinski–Harabasz score, and discuss the results.

Claims and Evidence

Overall, the authors’ claims and the evidence match well.

1. It is possible to build a practical pipeline for processing and clustering unstructured police-record PDFs using OCR, LayoutLMv3, and clustering.
    a. Evidence: The authors provide all the steps clearly, and show that they can run these steps on 113 documents with sufficient text.
2. LayoutLMv3 produces useful multimodal document embeddings for clustering document types.
    a. Evidence: The authors show that clustering produces some structures, and the table 1 shows the summary of it.
3. A document QA model (based on LayoutLM) can automatically suggest document type labels.
    a. Evidence: They query a Q&A model with “What kind of document is this?” and obtain answers with mostly logical type descriptions, and also odd type estimates. They then filter answers above a confidence threshold.
4. SBERT-style text embeddings yield moderate but non-trivial clustering structure.

    a. Evidence: They show an inertia and silhouette score plot for K-means over different k values and report maximum silhouette scores around 0.16–0.22, with no clear elbow in the inertia curve.
5. The data pipeline can be scaled to large unstructured datasets of PDFs.
    a. Evidence: X (I don't see any good supporting evidence for this claim.)

Relation to Prior Works

Overall, the authors properly mention the prior works.

Legal document embeddings and clusterings: The paper discusses TF–IDF, Doc2Vec, and LSA, citing prior work that evaluated these for legal documents and found TF–IDF to perform best with k-means and LDA. This motivates their own TF–IDF baseline and clustering work.

Multimodal document models:LayoutLMv3 is appropriately cited and described as a state-of-the-art model combining text and image patches with pretraining tasks (MLM, MIM, WPA, etc.). This motivates using LayoutLMv3 for the multimodal embedding step.

Other Aspects

- Clarity and writing: Overall quite clear and logically structured. The introduction is particularly strong, clearly explaining the social context.

- Figures and tables: The cluster plots are helpful to understand, and table 1 summarizes the cluster semantics nicely. However, I wasn't able to understand Figure 2. It would be great if the authors can clarify what the x-axis and y-axis are about.

- Typo: I found some typos
    - 3.1 Dataset: Some are document are … → Some documents are
    - 3.4 a threshold (10 → a threshold (10)

Questions for Authors

1. The abstract says "Large database", and introduction says "25 TB", but I think the analysis is only performed for 184 documents. In that case, how did the authors sample the 184 documents out of that large database?

2. The authors claim that the data pipeline can be scaled to large unstructured datasets of PDFs. Could the authors explain this in more detail?
3. Based on my understanding, you created a manually labeled set of 10 document types. Did you compute any label-based clustering metrics for any of your embeddings and clustering algorithms?
4. In the Q&A-based document typing experiment, how exactly did you pick the confidence threshold, and how often did the model give obviously wrong labels versus partially correct ones?
5. Could the authors clarify the x-axis and y-axis in Figure 2? It is not labeled.