World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:14, No:11, 2020

Machine Learning Strategies for Data Extraction from Unstructured Documents in Financial Services

Authors : Delphine Vendryes, Dushyanth Sekhar, Baojia Tong, Matthew Theisen, Chester Curme

Abstract : Much of the data that inform the decisions of governments, corporations and individuals are harvested from unstructured documents. Data extraction is defined here as a process that turns non-machine-readable information into a machine-readable format that can be stored, for instance, in a database. In financial services, introducing more automation in data extraction pipelines is a major challenge. Information sought by financial data consumers is often buried within vast bodies of unstructured documents, which have historically required thorough manual extraction. Automated solutions provide faster access to non-machine-readable datasets, in a context where untimely information quickly becomes irrelevant. Data quality standards cannot be compromised, so automation requires high data integrity. This multifaceted task is broken down into smaller steps: ingestion, table parsing (detection and structure recognition), text analysis (entity detection and disambiguation), schema-based record extraction, user feedback incorporation. Selected intermediary steps are phrased as machine learning problems. Solutions leveraging cutting-edge approaches from the fields of computer vision (e.g. table detection) and natural language processing (e.g. entity detection and disambiguation) are proposed.

Keywords : computer vision, entity recognition, finance, information retrieval, machine learning, natural language processing
Conference Title : ICDAR 2020 : International Conference on Document Analysis and Recognition
Conference Location : Jerusalem, Israel
Conference Dates : November 26-27, 2020