World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:16, No:12, 2022

The Platform for Digitization of Georgian Documents

Authors : Erekle Magradze, Davit Soselia, Levan Shughliashvili, Irakli Koberidze, Shota Tsiskaridze, Victor Kakhniashvili, Tamar Chaghiashvili

Abstract : Since the beginning of active publishing activity in Georgia, voluminous printed material has been accumulated, the digitization of which is an important task. Digitized materials will be available to the audience, and it will be possible to find text in them and conduct various factual research. Digitizing scanned documents means scanning documents, extracting text from the scanned documents, and processing the text into a corresponding language model to detect inaccuracies and grammatical errors. Implementing these stages requires a unified, scalable, and automated platform, where the digital service developed for each stage will perform the task assigned to it; at the same time, it will be possible to develop these services dynamically so that there is no interruption in the work of the platform.

Keywords : NLP, OCR, BERT, Kubernetes, transformers
Conference Title : ICNLP 2022 : International Conference on Natural Language Processing
Conference Location : Sydney, Australia
Conference Dates : December 02-03, 2022