OCRBench - Benchmark Framework for OCR Engines.

build unknown release no releases or repo not found release date no releases or repo not found github repo or version not found

---

Table of Contents

---

1. Introduction

`OCRBench` can be used to determine the performance of different OCR engines under identical conditions.

The currently supported OCR engines are:

Product	Company	Remarks
Document AI	Google Cloud	Cloud, ML
Tesseract OCR	Ray Smith	Local, Open Source
Textract	Amazon Web services, Inc.	Cloud, ML

---

2. Resources

2.1 Papers

Hegghammer, T. (2021). OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment.

Journal of Computational Social Science, 2021, pp. 2432-2725 [Online] Available at https://doi.org/10.1007/s42001-021-00149-1 (Accessed 04 January 2022).

Optical Character Recognition (OCR) can open up understudied historical documents to computational analysis, but the accuracy of OCR software varies. This article reports a benchmarking experiment comparing the performance of Tesseract, Amazon Textract, and Google Document AI on images of English and Arabic text. English-language book scans (n=322) and Arabic-language article scans (n=100) were replicated 43 times with different types of artificial noise for a corpus of 18,568 documents, generating 51,304 process requests. Document AI delivered the best results, and the server-based processors (Textract and Document AI) performed substantially better than Tesseract, especially on noisy documents. Accuracy for English was considerably higher than for Arabic. Specifying the relative performance of three leading OCR products and the differential effects of commonly found noise types can help scholars identify better OCR solutions for their research needs. The test materials have been preserved in the openly available “Noisy OCR Dataset” (NOD) for reuse in future benchmarking studies.

Minaee, S. et al. (2021). Deep Learning Based Text Classification: A Comprehensive Review.

arXiv [Online] Available at https://arxiv.org/abs/2004.03705 (Accessed 04 January 2022).

Deep learning based models have surpassed classical machine learning based approaches in various text classification tasks, including sentiment analysis, news categorization, question answering, and natural language inference. In this paper, we provide a comprehensive review of more than 150 deep learning based models for text classification developed in recent years, and discuss their technical contributions, similarities, and strengths. We also provide a summary of more than 40 popular datasets widely used for text classification. Finally, we provide a quantitative analysis of the performance of different deep learning models on popular benchmarks, and discuss future research directions.

Paaß G., Konya I. (2011) Machine Learning for Document Structure Recognition.

In: Mehler A., Kühnberger KU., Lobin H., Lüngen H., Storrer A., Witt A. (eds) Modeling, Learning, and Processing of Text Technological Data Structures. Studies in Computational Intelligence, vol 370. Springer, Berlin, Heidelberg. Available at https://www.researchgate.net/publication/265487498_Machine_Learning_for_Document_Structure_Recognition (Accessed 04 January 2022).

The backbone of the information age is digital information which may be searched, accessed, and transferred instantaneously. Therefore the digitization of paper documents is extremely interesting. This chapter describes approaches for document structure recognition detecting the hierarchy of physical components in images of documents, such as pages, paragraphs, and figures, and transforms this into a hierarchy of logical components, such as titles, authors, and sections. This structural information improves readability and is useful for indexing and retrieving information contained in documents. First we present a rule-based system segmenting the document image and estimating the logical role of these zones. It is extensively used for processing newspaper collections showing world-class performance. In the second part we introduce several machine learning approaches exploring large numbers of interrelated features. They can be adapted to geometrical models of the document structure, which may be set up as a linear sequence or a general graph. These advanced models require far more computational resources but show a better performance than simpler alternatives and might be used in future.