**ShabbyPages:** A Robust Corpus for Training Document Image Models

Alexander Groleau, Kok Wei Chee, Stefan Larson, Jonathan Boarman

ShabbyPages is a corpus of born-digital document images with both ground truth and distorted versions appropriate for use in training and evaluating document denoising models. This new dataset with synthetically-generated real-world representations can be used to improve document layout detection, text extraction and OCR processes that depend on denoising and binarization preprocessing models. ShabbyPages was created using the Augraphy data augmentation library. The image at left shows clean and noisy document versions.

In the figure at right, we evaluate a denoising model trained on the NoisyOffice dataset on two samples from the ShabbyPages dataset (bottom two rows). We observe that the NoisyOffice model fails to denoise ShabbyPages completely.