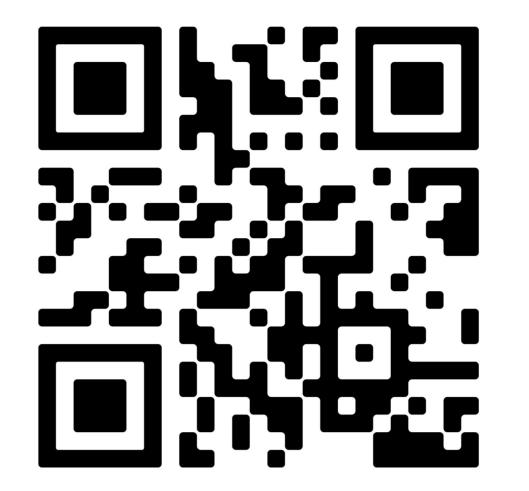
ShabbyPages: A Robust Corpus for **Training Document Image Models**

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

X⇒ Sparkfish

<u>github.com/sparkfish/shabby-pages</u>



favouritism can secure substantial prerogatives and profit or some social sub-groups.

What Gauses Corruption? 3.

Many plausible theories on corruption have been derive characteristics of individual societies¹². It has for insta salience of corruption is the carry-over into present-day po values inherited form a patrimonial past, like negunconditional solidarity with extended families, clans and Sardan 1999:25) This may explain the contrast between differences between the catholic Western European count and the Nordic, protestant countries.

Besides, in some countries, private-regarding behav agents who act for the benefit of his family and friends, i furthermore considered a moral duty. From the culturally even been argued that corruption is not a crime whenev culture. The one and same act may therefore be judicially accepted. Furthermore, the illegality of corruption varies ac

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 syntheticallygenerated 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 created was 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.

Spanish is a widespread major language. A Spanish is a widespread major language. reason was to create a corpus from semant, reason was to create a corpus from semant. These tasks are commonly used in practice These tasks are commonly used in practice of linguistic knowledge beyond the lexicon of linguistic knowledge beyond the lexicon nition process.

most databases do not contain spanish sente most databases do not contain spanish sente nition process.

	A CARLES AND AND ADDRESS AND ADDRESS AND ADDRESS AND ADDRESS AD
WDTH >/= 4X THK Innished 7224900065 OTHER SEMIFINISHED OTH ALLOY STL, NOT TOOL STL, CIRCULAR CROSS SECTION Blooms, billets and slabs, semi- finished. 7224900075 OTHER SEMIFINISHED OTH ALLOY STL, NOT TOOL STL, CROSS SECTION OTHER THAN RECT/SO/CIRC Blooms, billets - and slabs, semi- finished. 7225110000 FLAT-ROLLED OTH ALLOY STL, WDTH >/= 600MM, SILICON ELECTRICAL STL, GRAIN-ORIENTED Silicon electrical steel sheets and strip 7225190000 FLAT-ROLLED OTH ALLOY STL, WDTH >/= 600MM, SILICON ELECTRICAL STL, GRAIN-ORIENTED Silicon electrical steel sheets and strip	7153000001 OTHER SEMIFONEMED OTH ALLOY STL, NOT TOOL STL, CECURAR CROSS SECTION Binears, bullers Intended 723000001 OTHER SEMIFONEMED OTH ALLOY STL, NOT TOOL STL, CECURAR CROSS SECTION Intense, bullers Intended 723000001 OTHER SEMIFONEMED OTH ALLOY STL, NOT TOOL STL, CECURAR CROSS SECTION Intense, bullers Intended 723000001 OTHER SEMIFONEMED OTH ALLOY STL, NOT TOOL NEWOR Intense, bullers 725010000 FLAT, EOLIED OTH ALLOY STL, WOTH 24- 6006104, SELEON FLECTRICAL STL, OR NOTHER THED Silicon electrical steal sheets and to steal sheets and sheets and to steal sheets and sheets and sheets and sheets and sheets and sheets and sh
Ar least in the normative ideal of political representation, the presence of programmatic protected normality grout for the quality of descendacy, but is also an efficient purp building armest for office-seeking politicians. Accordingly, in the medium to long term—and in the context of stable party systems—ambitions politicians would be better off investing their symbolic and material resources in crafting programmatic organizations. Programmatic parties can contribute to realizing that ideal using a number of methods.	Loting, & TOLO do do do a final and an and the second down of the s