Seleção Não-Supervisionada de Métodos de Binarização para Documentos Históricos

MAC0499 – Trabalho de Formatura Supervisionado

Aluno: Denis T. Ikeda
Orientador: Ronaldo Fumio Hashimoto

O que é binarização?

- Binarização é o ato de transformar uma imagem em níveis de cinza em uma imagem binária, que utiliza somente preto e branco.
- Seu objetivo é separar todo o texto da imagem de entrada dos ruídos e do fundo.

Exemplo de binarização

Imagem original

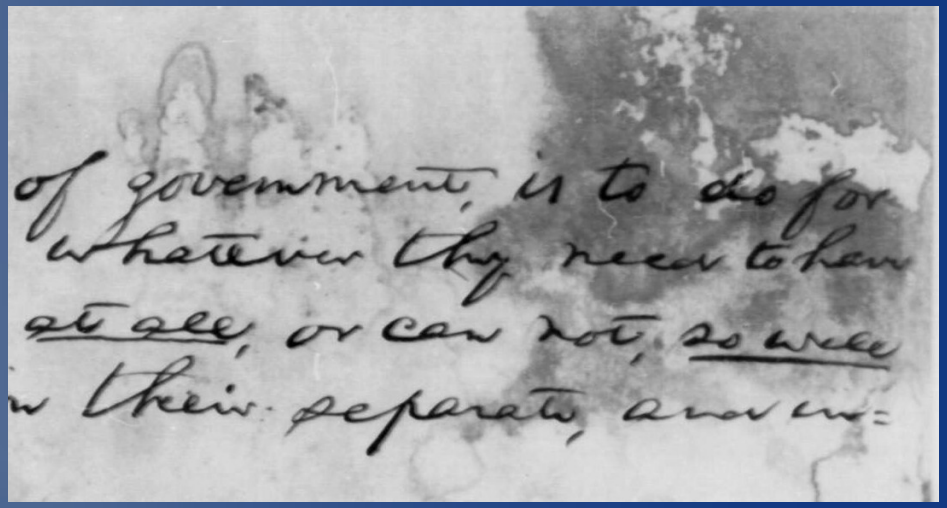
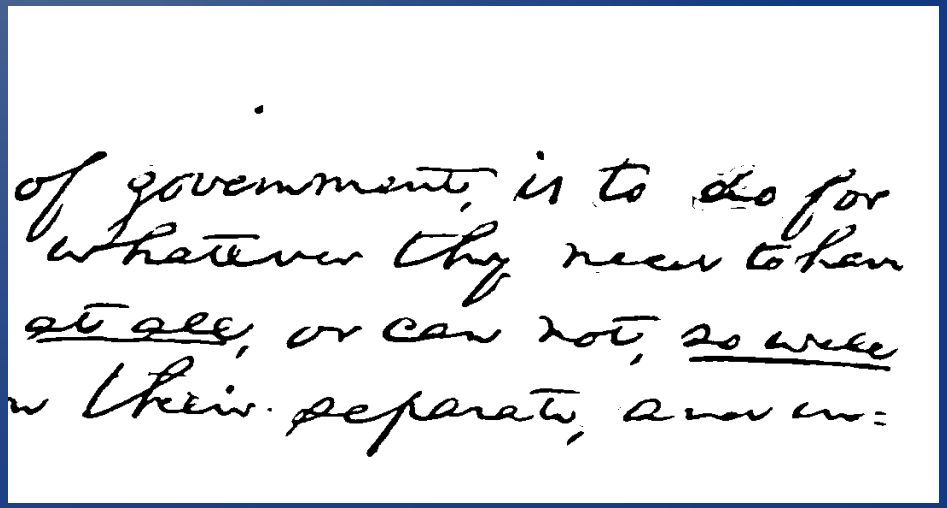


Imagem binarizada



Aplicações da binarização para documentos

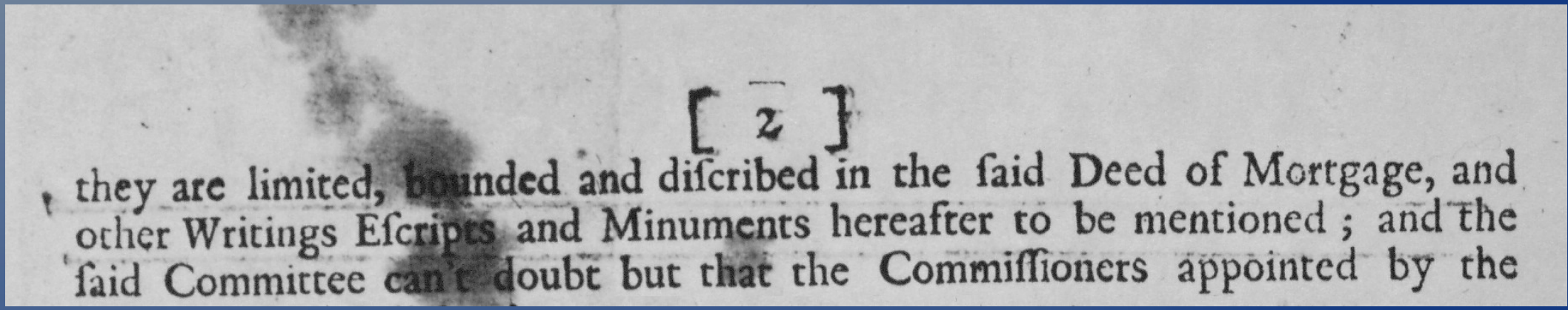
- Pré-processamento do OCR
- Diminui tamanho dos arquivos de imagem
- Melhora legibilidade dos documentos

Documentos históricos

- Problema resolvido para documentos em bom estado
- Documentos históricos possuem vários defeitos

Documentos históricos - Exemplos

- Manchas

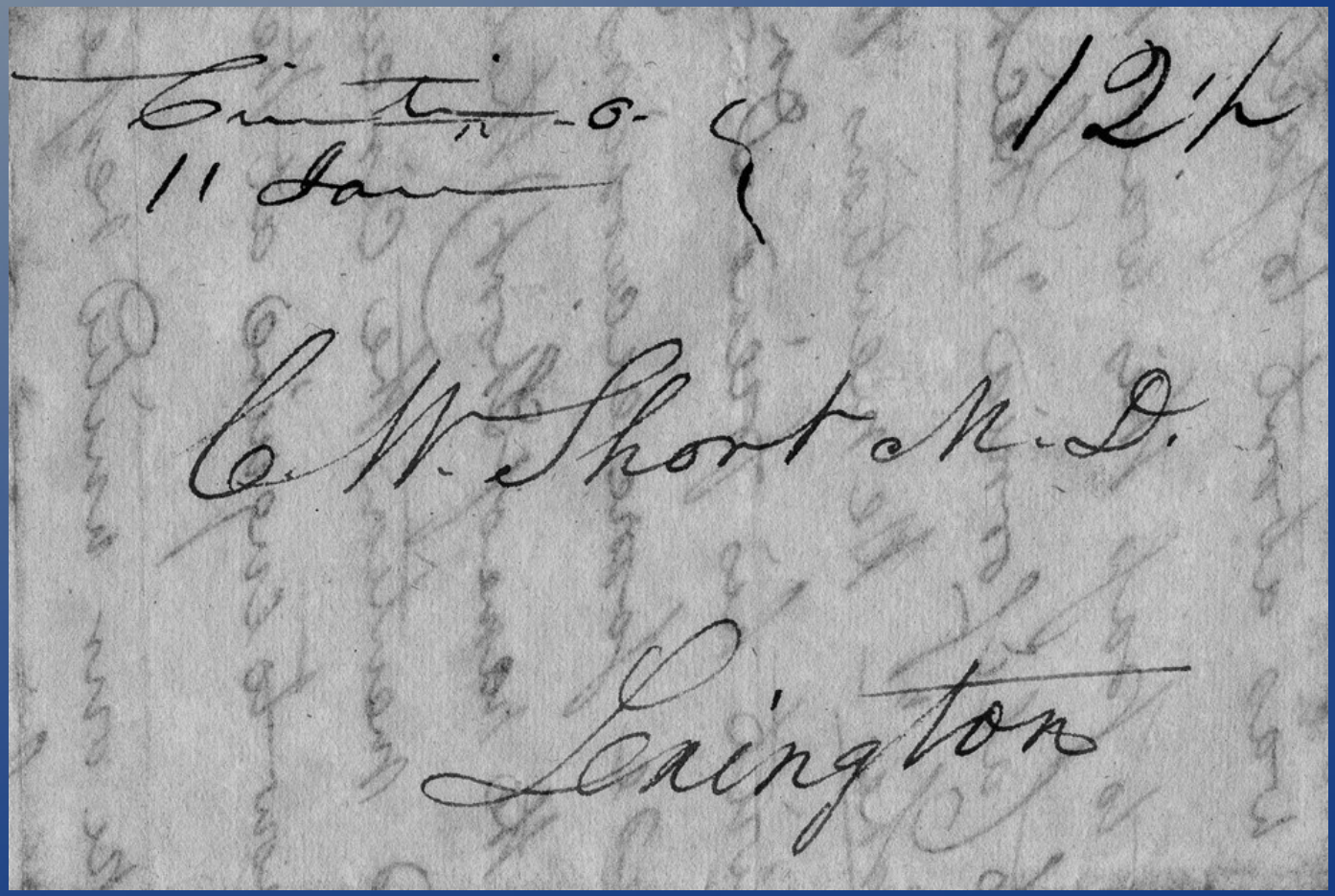


[z]

they are limited, bounded and diſcribed in the ſaid Deed of Mortgage, and other Writings Eſcripts and Minuments hereafter to be mentioned ; and the ſaid Committee can't doubt but that the Commiſſioners appointed by the

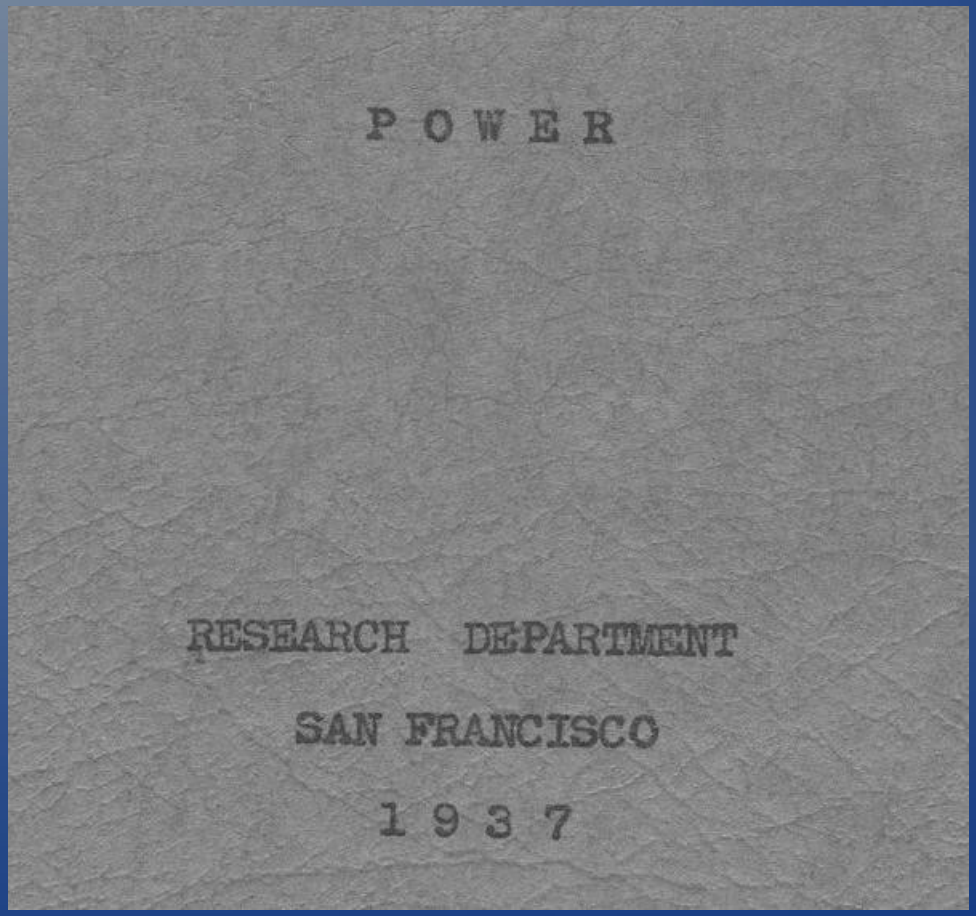
Documentos históricos - Exemplos

- Verso aparecendo na frente



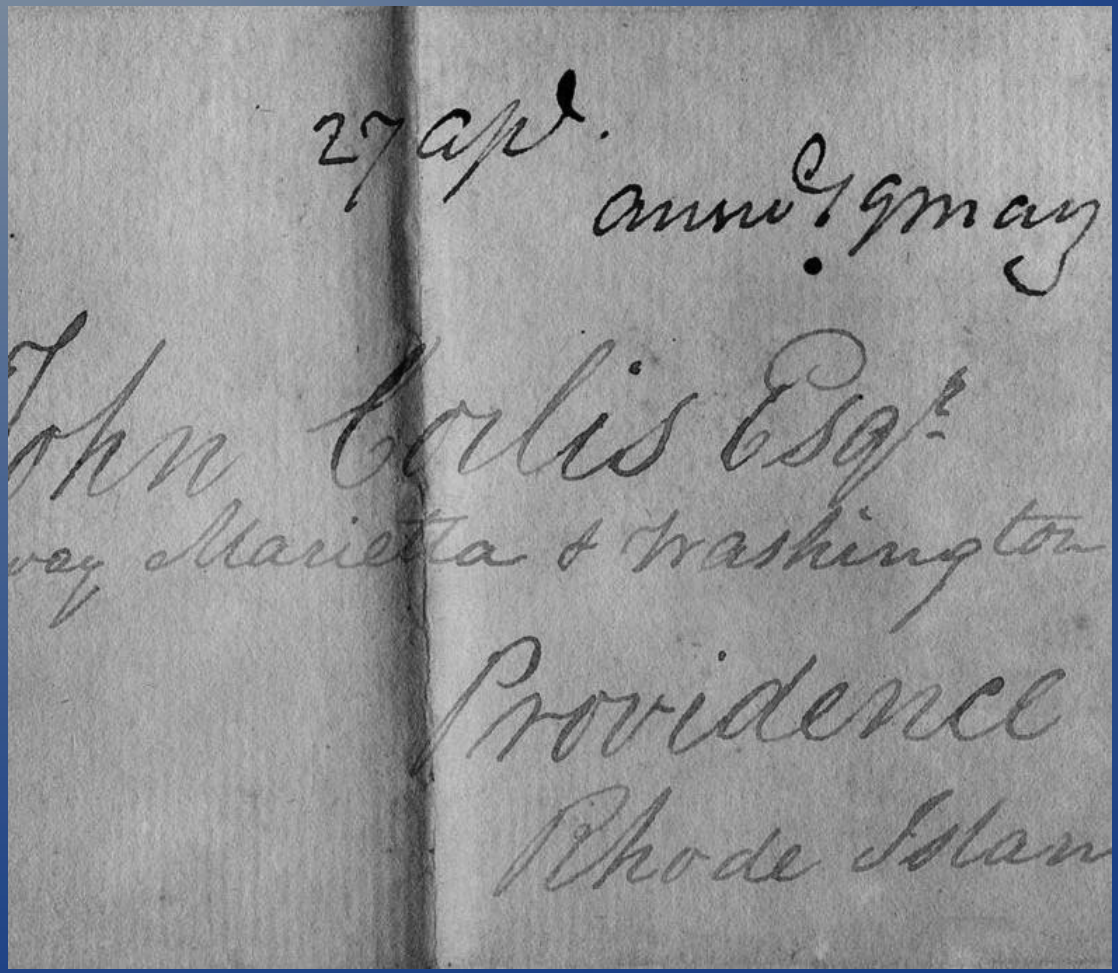
Documentos históricos - Exemplos

- Pouco contraste entre texto e fundo



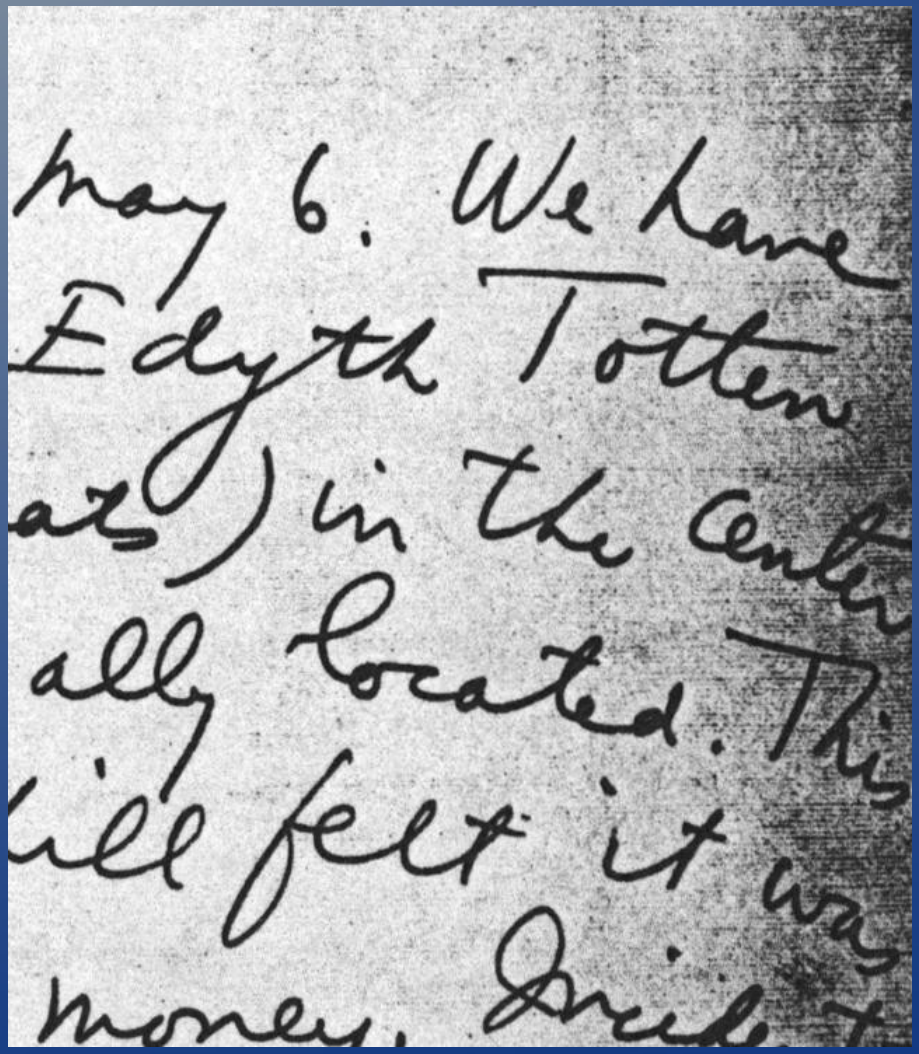
Documentos históricos - Exemplos

- Dobras



Documentos históricos - Exemplos

- Iluminação desigual



Competições DIBCO - Objetivos

- Averiguar interesse da pesquisa nesta área
- Avaliar o estado da arte de métodos de binarização de documentos históricos
- Estabelecer um sistema de avaliação padronizado, composto de métricas e conjunto de dados

Competições DIBCO - Resultados

- Há grande interesse e métodos muito bons
- Métodos são inconsistentes

Seleção Proposta - Objetivos

- Adicionar consistência e confiabilidade à binarização
- Melhorar a média
- Tornar-se melhor que cada indivíduo

Seleção Proposta – Ideia Principal

- Para selecionar, é necessário um critério de decisão
- Foram utilizados Precision e Recall
- Assumem que existe uma resposta ideal (ground truth)
- Para relaxar esta suposição, estes valores são estimados

Seleção Proposta – Estimativa

- Calcular a probabilidade de cada pixel estar correto
- Considere cada resultado de método como uma contribuição para estimativa do texto real
- Além destes, adicione dois resultados artificiais: uma imagem branca e uma imagem preta

Seleção Proposta – Escolha

- Considerando os métodos igualmente confiáveis, então pode-se usar uma distribuição uniforme
- Precision e Recall podem ser calculados utilizando probabilidade condicional
- Resultados ordenados por F-Measure
- Maior F-Measure considerado o resultado mais correto

Seleção Proposta – Problemas

- Escolha de imagens muito apagadas ou com muito ruído
- Acabou ficando em último nos testes

Seleção Proposta – Soluções

- Filtrar os resultados discrepantes de Recall
- Não utilizar uma imagem preta
- Imagens homogêneas ou mapas de contraste

Resultados - Métodos

- Para comparar, foram implementados métodos clássicos
- Otsu como método global
- Niblack, Sauvola e White como métodos locais
- Su, o vencedor do H-DIBCO 2010

Resultados - Métricas

- De todos os anos
- F-Measure (com GT), PSNR
- NRM, MPM e DRD

Resultados - 2009

	posição	placar
Proposta	1	5
Su	2	8
White	3	14
Otsu	4	15
Sauvola	5	19
Niblack	6	23

	posição	placar
Proposta	1	60
Su	2	69
Whte	3	93
Otsu	4	111
Sauvola	5	137
Niblack	6	190

Resultados - 2010

	posição	placar
Su	1	9
Proposta	2	10
Otsu	3	16
White	4	20
Sauvola	5	22
Niblack	6	28

	posição	placar
Su	1	85
Proposta	2	98
Otsu	3	105
White	4	144
Sauvola	5	187
Niblack	6	229

Resultados - 2011

	posição	placar
Proposta	1	5
White	2	11
Su	3	12
Sauvola	4	14
Otsu	5	18
Niblack	6	24

	posição	placar
Su	1	115
Proposta	2	125
Otsu	3	173
White	4	174
Sauvola	5	179
Niblack	6	319

Resultados – Simulação DIBCO 2011

	posição	placar
Proposta	1	8
11	2	10
4	3	15
3	4	23
6	5	24
2	6	28

	posição	placar
10	1	316
8	2	351
11	3	428
Proposta	4	466
6	5	467
4	6	486

Conclusões

- Embora possa escolher resultados um pouco abaixo do ótimo, aumenta a consistência e confiabilidade por evitar os resultados inaceitáveis
- Melhor desempenho (pesos, métodos heterogêneos, outras estratégias)