

Reconhecimento de Texto

Cupons Fiscais

Motivação

- Ferramentas de reconhecimento não conseguem se adaptar à textos com tabulações incomuns.
- Uma abordagem visando apenas o aprendizado do OCR não apresenta resultados satisfatórios.
- Cupons fiscais possuem o texto esparço, prejudicando muito a técnica que os OCRs utilizam para o reconhecimento.

O cupom fiscal

Epson do Brasil Ind. e Com.
Epson do Brasil
Av. Tucunaré 720
06460-020

CNPJ: 52.106.911/0001-00
IE: 206108738115
IM: 987654321098

07/07/2007 12:56:19 CCF:000010 C00:000035

CUPOM FISCAL

ITEM	CÓDIGO	QTD.	UN.	VL UNIT	RS	ST	VL ITEM	RS
1	78900000001234							
2	78900000004321							
3	78900000001111							

TOTAL R\$ 21,00
Dinheiro 50,00
TROCO R\$ 29,00
Ta=12,00% Tb=18,00%

Obrigado - Volte Sempre

CCD4CA 444A4D D1E13B BECAAS 3C1E24 4DB325
EPSON TM-T81 FBII
ECF-IF VERSÃO:01.00.04 ECF:001 LJ:001
!!!!!!!!!!!!!!*#<!! 07/07/2007 12:57:35
FAB:EP040710000000000963

urmet
DARUMA

Urmet Daruma
daruma.com.br sac@daruma.com.br
Av Independência, 3500 Taubaté
SAC ** (12) 3609-5050 **
CNPJ: 45.170.289/0001-25
IE: 688.023.460.111
IM: 3.633.72

23/11/2011 15:03:40V CCF:002612 C00:004275
CNPJ/CPF Consumidor: 123.456.789-00
NOME: João da Silva

CUPOM FISCAL

ITEM	CÓDIGO	QTD.	UN.	VL UNIT	RS	ST	A/T	VL ITEM	RS
001	98765432109876								
002	12345678901098								

TOTAL R\$ 8,55
Dinheiro 10,00
TROCO R\$ 1,45

Obrigada e Volte Sempre!!!

DARUMA DEVELOPERS COMMUNITY

DarumaFramework - Mensagem Não Programada
DarumaFramework - Mensagem Não Programada
IAS 81974 38074 7C9026 8C D6019B 262DA 1DFED DAB
DARUMA AUTOMACAO MACH 2
ECF-IF VERSÃO:01.00.00 ECF:001 LJ:
88888888IFBCFGIHEI 23/11/2011 15:03:42V
FAB:DR0910BR000088888888

Grandes problemas

- Não existe nenhum padrão de fonte e formato para os cupons fiscais
- A imagem tomada nem sempre possui uma qualidade adequada
- Processar a imagem leva muito tempo inviabilizando uma interação mais profunda com o usuário

Como melhorar o reconhecimento?

- Criar um servidor de armazenamento de imagens para, a priori, classificar manualmente muitos cupons fiscais
- A análise de muitas imagens deverá resultar em um bom aprendizado
- Segmentar a imagem, reduzindo o ruído

Tesseract

Existem outras formas que o OCR consegue trabalhar, diferente do que textos corridos:

-psm N

- 0 = Orientation and script detection (OSD) only.
- 1 = Automatic page segmentation with OSD.
- 2 = Automatic page segmentation, but no OSD, or OCR.
- 3 = Fully automatic page segmentation, but no OSD. (Default)
- 4 = Assume a single column of text of variable sizes.
- 5 = Assume a single uniform block of vertically aligned text.
- 6 = Assume a single uniform block of text.
- 7 = Treat the image as a single text line.
- 8 = Treat the image as a single word.
- 9 = Treat the image as a single word in a circle.
- 10 = Treat the image as a single character.

Fragmentação em palavras

SPOLETO
CULINÁRIA ITALIANA

NOVA JORNADA RESTAURANTE LTDA.

CNPJ: 07.669.064/0010-10 B

GAVEA-RIO DE JANEIRO-RJ
IRF=64.02-NORTE

CNPJ: 07.669.064/0010-10 IE: 78.427.043
11/07/2010 19:09:38V CCF:117539 C00:121958

CUPOM FISCAL

ITEM	CÓDIGO	DESCRIÇÃO	QTD.	UN.	VL.	UNIT(R\$)	ST	VL.	ITEM(R\$)
001	00000000000010	FETTUCCINE							
1	un	X	13,90	T13,00%					13,90)

Como encontrar os padrões?

1. Melhorar o alinhamento
2. Identificar a altura das linhas
3. Identificar as palavras
4. Recortar
5. Processar utilizando o OCR
6. Reconhecer os padrões de interesse

Alinhando a imagem

0 500 1000 1500 2000 2500 3000

LIVRARIA CULTURA S/A
AV NACDES UNIDAS, 4777 LJ 245 1 P
JD. UNIVERSIDADE - SAO PAULO / SP
CNPJ: 62.410.352/0009-20 IE: 115.491.716.114
24/05/2015 16:28:08 CCF: 240495 CCO: 520484
CNPJ/CPF consumidor: 174.274.478-83
NOME: ALFREDO GOLDMAN
END:

CUPOM FISCAL

ITEM	CODIGO	QTD.	UN.	VL.UNIT(R\$)	ST	IAT	VL.ITEM(R\$)
001	09128278	1,000Un x					
				3,90	II	A	3,90:
002	42876210	1,000Un x					
				29,90	II	A	29,90:
003	42865520	1,000Un x					
				29,90	II	A	29,90:
004	43002054	1,000Un x					
				16,00	II	A	16,00:
TOTAL							R\$
							79,70
CARTAO							79,70
HD-5:16069B3AE7D82E69CD911A660AB8D07							U
1 Aprox Tributos R\$ 6,38(6,01%) Fonte:IBPT							
CI 230429-6							
Ler para Ser							
Telefone: (11) 3024-3599							
3/2RTh/*S-/X-vvK9xx/>xx!/#oKbXb#ShT&K*px-ps=*8Y							
ZPM ZPM/2EFC LOGGER ECF-IF							
VERSAO:03.04.00 ECF:020 LJ:0245 OPR:BPERROTE							
EEEEEEEOEAUOPPII							
AB:ZP030700564 24/05/2015 16:29:05							

0 500 1000 1500 2000

0 500 1000 1500 2000 2500 3000

LIVRARIA CULTURA S/A
AV NACDES UNIDAS, 4777 LJ 245 1 P
JD. UNIVERSIDADE - SAO PAULO / SP
CNPJ: 62.410.352/0009-20 IE: 115.491.716.114
24/05/2015 16:28:08 CCF: 240495 CCO: 520484
CNPJ/CPF consumidor: 174.274.478-83
NOME: ALFREDO GOLDMAN
END:

CUPOM FISCAL

ITEM	CODIGO	QTD.	UN.	VL.UNIT(R\$)	ST	IAT	VL.ITEM(R\$)
001	09128278	1,000Un x					
				3,90	II	A	3,90:
002	42876210	1,000Un x					
				29,90	II	A	29,90:
003	42865520	1,000Un x					
				29,90	II	A	29,90:
004	43002054	1,000Un x					
				16,00	II	A	16,00:
TOTAL							R\$
							79,70
CARTAO							79,70
HD-5:16069B3AE7D82E69CD911A660AB8D07							U
1 Aprox Tributos R\$ 6,38(6,01%) Fonte:IBPT							
CI 230429-6							
Ler para Ser							
Telefone: (11) 3024-3599							
3/2RTh/*S-/X-vvK9xx/>xx!/#oKbXb#ShT&K*px-ps=*8Y							
ZPM ZPM/2EFC LOGGER ECF-IF							
VERSAO:03.04.00 ECF:020 LJ:0245 OPR:BPERROTE							
EEEEEEEOEAUOPPII							
FAB:ZP030700564 24/05/2015 16:29:05							

0 500 1000 1500 2000

Limiarização

0 500 1000 1500 2000 2500 3000

LIVRARIA CULTURA S/A
AV NACÕES UNIDAS, 4777 LJ 245 1 P
JD. UNIVERSIDADE - SÃO PAULO / SP
CNPJ:62.410.352/0009-20 IE:115.491.716.114
24/05/2015 16:28:08 CCF:240495 C00:520484
CNPJ/CPF consumidor: 174.274.478-93
NONE: ALFREDO GOLDMAN
END:

CUPOM FISCAL						
ITEM	CODIGO	DESCRICAO	QTD.	UN.	VL.UNIT(R\$)	ST IAT VL.ITEM(R\$)
001	09128278	PATO DONALD - WH2443	1,000	Un x	3,90	11 A 3,90:
002	42876210	COISAS BEM LEGAIS PARA S	1,000	Un x	29,90	11 A 29,90:
003	42865520	HERDEIRA, A	1,000	Un x	29,90	11 A 29,90:
004	43002054	SUPERPATETA	1,000	Un x	16,00	11 A 16,00:
TOTAL					R\$	79,70
CARTAO						79,70
HD-5:1506393AE7D82E696CD911A660AB8D07 U						
1 Aprox Tributos R\$ 6,58(8,01%) Fonte:IBPT						
CI 230429-6						
Ler para Ser Telefone: (11) 3024-3599						
@/2RTh/*S-X-vYK89xx/>xxX!/#bKbXb#ShT&K*pk-ps=*8Y						
ZPM ZPM/2EFC 1066ER ECF-IF						
VERSAO:03.04.00 ECF:020 LJ:0245 OPR:BPERROTE						
EEEEEEEEE0EAU0APP11 24/05/2015 16:29:05						
FAB:ZP030700564						

0 500 1000 1500 2000

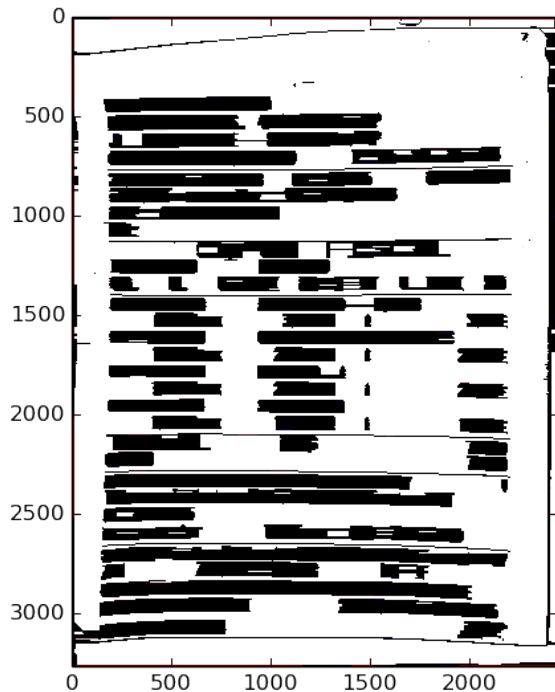
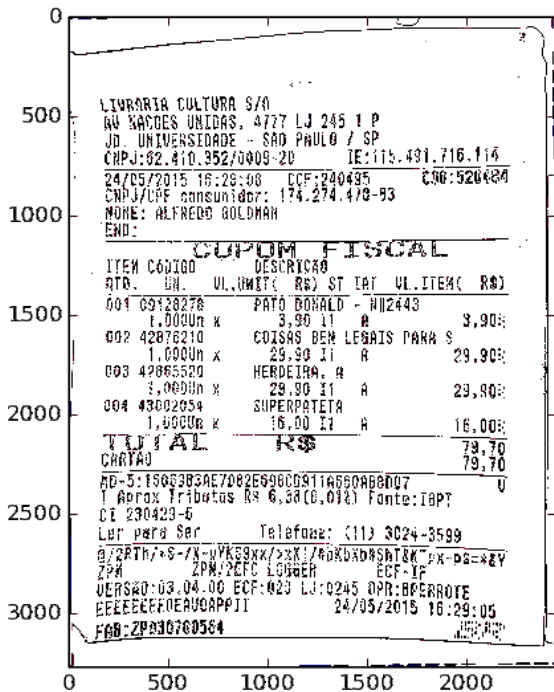
0 500 1000 1500 2000 2500 3000

LIVRARIA CULTURA S/A
AV NACÕES UNIDAS, 4777 LJ 245 1 P
JD. UNIVERSIDADE - SÃO PAULO / SP
CNPJ:62.410.352/0009-20 IE:115.491.716.114
24/05/2015 16:28:08 CCF:240495 C00:520484
CNPJ/CPF consumidor: 174.274.478-93
NONE: ALFREDO GOLDMAN
END:

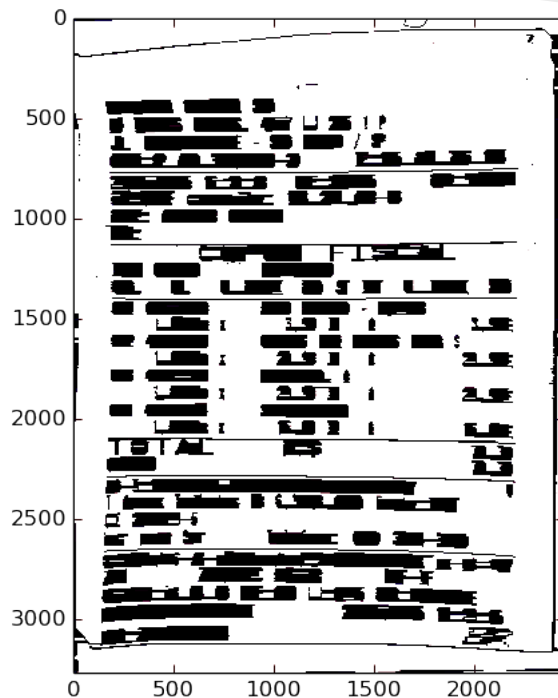
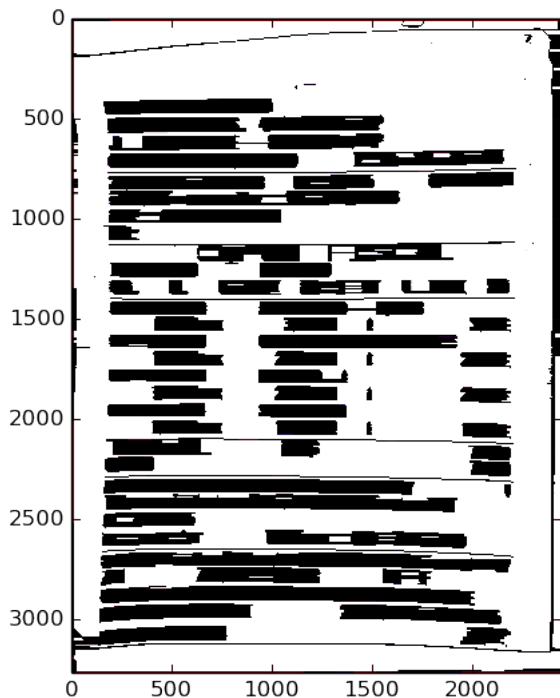
CUPOM FISCAL						
ITEM	CODIGO	DESCRICAO	QTD.	UN.	VL.UNIT(R\$)	ST IAT VL.ITEM(R\$)
001	09128278	PATO DONALD - WH2443	1,000	Un x	3,90	11 A 3,90:
002	42876210	COISAS BEM LEGAIS PARA S	1,000	Un x	29,90	11 A 29,90:
003	42865520	HERDEIRA, A	1,000	Un x	29,90	11 A 29,90:
004	43002054	SUPERPATETA	1,000	Un x	16,00	11 A 16,00:
TOTAL					R\$	79,70
CARTAO						79,70
HD-5:1506393AE7D82E696CD911A660AB8D07 U						
1 Aprox Tributos R\$ 6,58(8,01%) Fonte:IBPT						
CI 230429-6						
Ler para Ser Telefone: (11) 3024-3599						
@/2RTh/*S-X-vYK89xx/>xxX!/#bKbXb#ShT&K*pk-ps=*8Y						
ZPM ZPM/2EFC 1066ER ECF-IF						
VERSAO:03.04.00 ECF:020 LJ:0245 OPR:BPERROTE						
EEEEEEEEE0EAU0APP11 24/05/2015 16:29:05						
FAB:ZP030700564						

0 500 1000 1500 2000

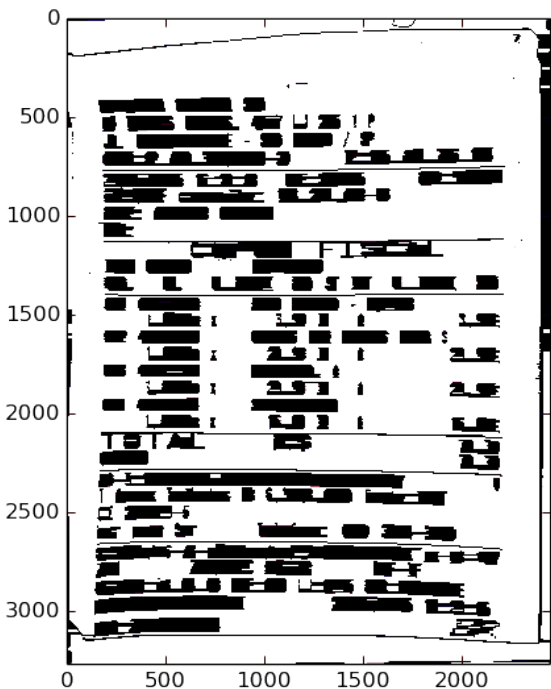
Altura das linhas



Fragmentando palavras



Extraindo palavras



0

500

1000

1500

2000

2500

3000

0 500 1000 1500 2000

ATIVIDADE CULTURA S7A

AD. RACIOLAS UNIDAS 4777 245 1 8

AD. UNIVERSIDADE - SAO PAULO 0 6A

CNPJ: 62.410.352/0009-20 E: 115.491.716.014

24/05/2015 16:28:08 ECF: 240495 EOD: 520484

CNPJ/CPF Consumidor: 174.274.478-83

NOME: ALFREDO GOLDMAN

END:

CUPOM FISCAL

ITEM	QUANT	DESCRIÇÃO	VL. UNIT	RS	QUANT	VL. TOTAL	RS
001	09128278	PATO DONATO - NIT2443	0.90	1	0	0.90	
002	82876210	MUSAS 6ER CERRAS PARA S	29.90	1	0	29.90	
003	82865520	HERDEIRA	29.90	1	0	29.90	
004	83002054	SUPERPATELA	16.00	1	0	16.00	
TOTAL						R\$	79.70
PAGAR							79.70

Id=5.16088638E7D62E696C051A660A88007

1 Aprox Tributos RS 6.38(8.01%) Fonte:IBPT

230429-6

PA 06R 6ER (telefone) (11) 8024-3599

3/28th/4S-VX-WYK69xx/xx1/8KbXb#ShT&K px-ps=x&y

ZPM ZPM/ZERO LOGEN ECF-1A

VERSÃO: 03.04.00 ECF: 020 L: 0243 DPR: 8PERROTE

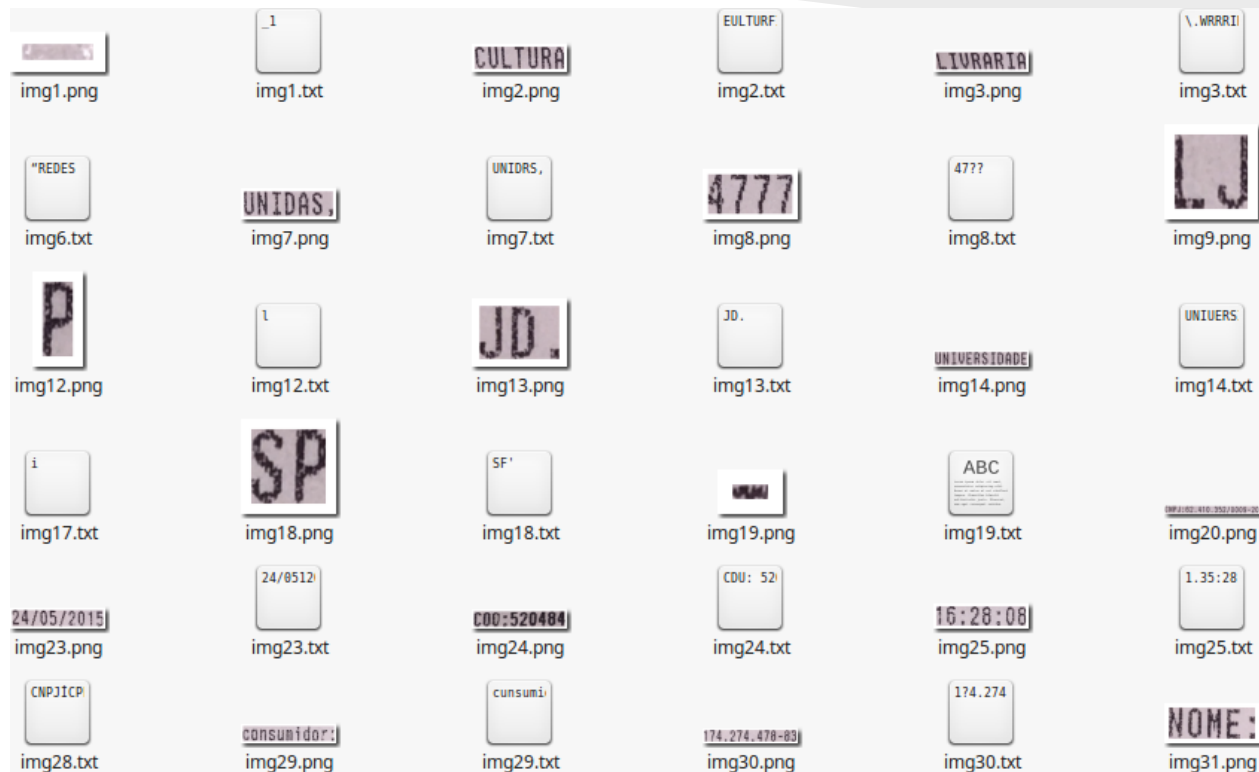
EEEEEEEEEEADAPP11 24/05/2015 16:29:08

EAB: ZP030700564

Resultado da extração



Aplicando o OCR



Encontrando padrões

CNPJ:62.410.352/0009-20

CNPJ:52.410~352;0009~20

24/05/2015

24/0512015

C00:520484

CDU: 520484

79,70

?9,70

```
{  
  "data":{  
    "cnpj" : "62.410.352/0009-20",  
    "data" : "24/05/2015",  
    "coo" : "C00: 520484",  
    "total" : "79,70"  
  }  
}
```


Melhorias

- Minimizar o uso de calculos replicados
- Paralelizar algoritmos
- Usar cálculo em GPU
- Inserir novos testes