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 COO:000035

CUPOM FISCAL

ITEM CÓDIGO DESCRIÇÃO
QTD. UN. VL UNIT R$ ST VL ITEM R$

1 7890000001234 Padaria 1UN X 12,00 Ta 12,00
2 7890000004321 Verduras e Legumes
1 UN X 5,00 Tb 5,00
3 7890000001111 Frios em geral 1UN X 4,00 Tb 4,00

TOTAL R$ 21,00
Dinheiro 50,00
Obrigado - Volte Sempre
TROCO R$ 29,00
Ta=12,00% Tb=18,00%

CCD4CA 444A4D D1E13B BECAA5 3C1E24 4DB325
EPSON TM-T81 FBII
ECF-IF VERSÃO:01.00.04 ECF:001 LJ:001
!!!!!!!!!!!!!x#<!! 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 COO:004275
CNPJ/CPF Consumidor: 123.456.789-00
NOME: João da Silva

CUPOM FISCAL

ITEM CÓDIGO DESCRIÇÃO QTD UN VL UNIT R$ ST A/T VL ITEM R$
001 98765432109876 REFRI COCA 600ML
3 UN X 1,50 F1 T 4,50)
002 12345678901098 CARNE MOIDA
0,676 KG X 5,99 I1 A 4,05)
TOTAL R$ 8,55
Dinheiro 10,00
TROCO R$ 1,45

DARUMA DEVELOPERS COMMUNITY
Obrigada e Volte Sempreÿÿ

DarumaFramework - Mensagem Não Programada
DarumaFramework - Mensagem Não Programada
1A8 81974 38074 7C9026 8C D60198 262DA 1DFED DA8
DARUMA AUTOMAÇÃO MACH 2
ECF-IF VERSÃO:01.00.00 ECF:001 LJ:
8B8B8B8B6IFBCFGIHEI 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

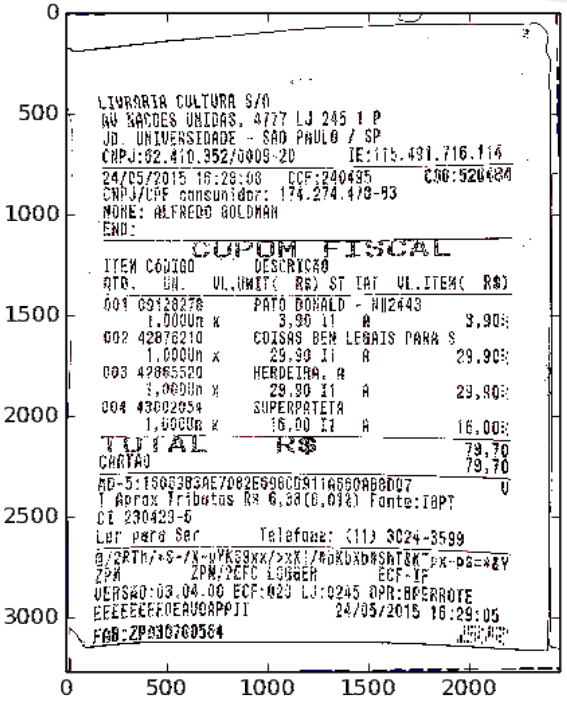
SPOLETO
CULINÁRIA ITALIANA
NOVA JORNADA RESTAURANTE LTDA.
CNPJ: 07.669.064/0010-10 B
GAVEA-RIO DE JANEIRO-RJ
IRF=64.02-NORTE
CNPJ: 07.669.064/0010-10 IE: 78.427.043
11/07/2010 19:09:38V CCF:117539 COO:121958
CUPOM FISCAL
ITEM CÓDIGO DESCRIÇÃO QTD.UN.VL UNIT(R$) ST VL ITEM(R$)
001 0000000000010 FETTUCCINE
1 un X 13,90 T13,00% 13,90)

Como encontrar os padrões?

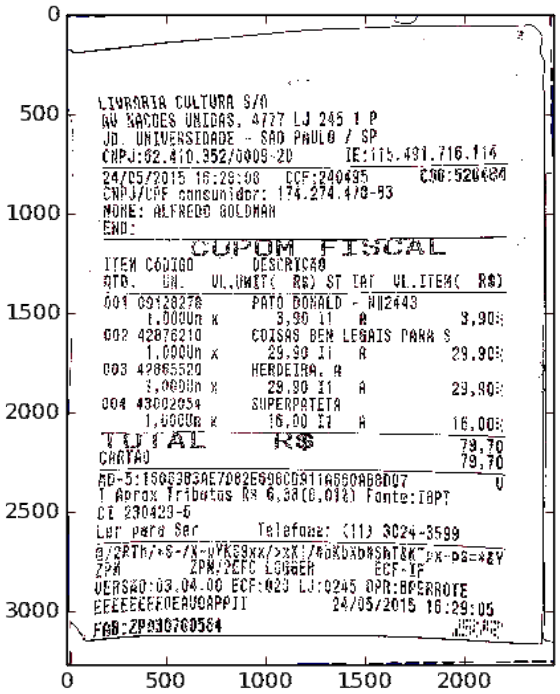
1. Melhorar o alinhamento
2. Identificar a altura das linhas
3. Identificar as palavras
4. Recortar
5. Processar utilizando o OCR
6. Reconhecer os padrões de interesse

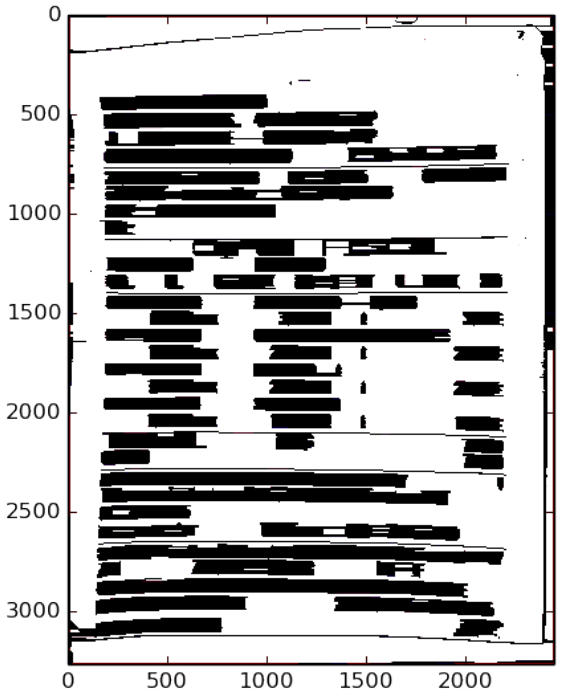
Alinhando a imagem

Limiarização

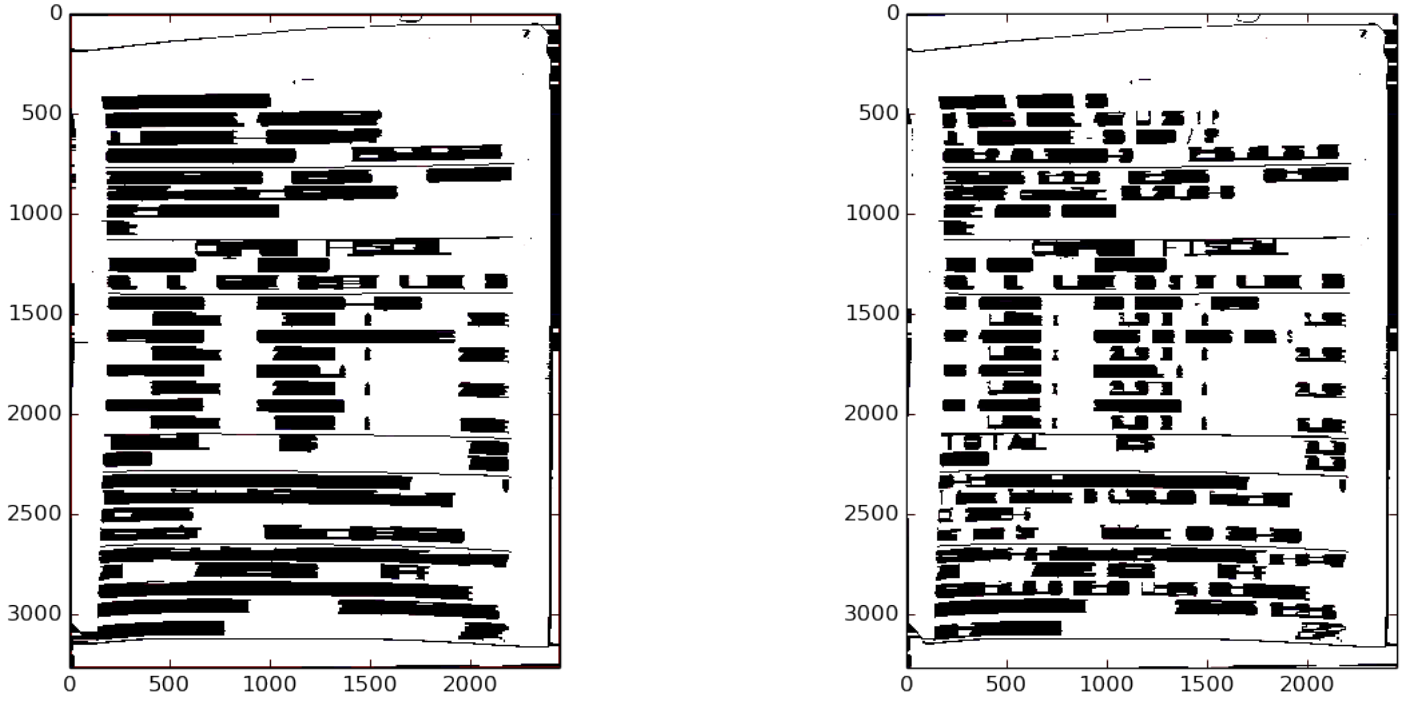


Altura das linhas

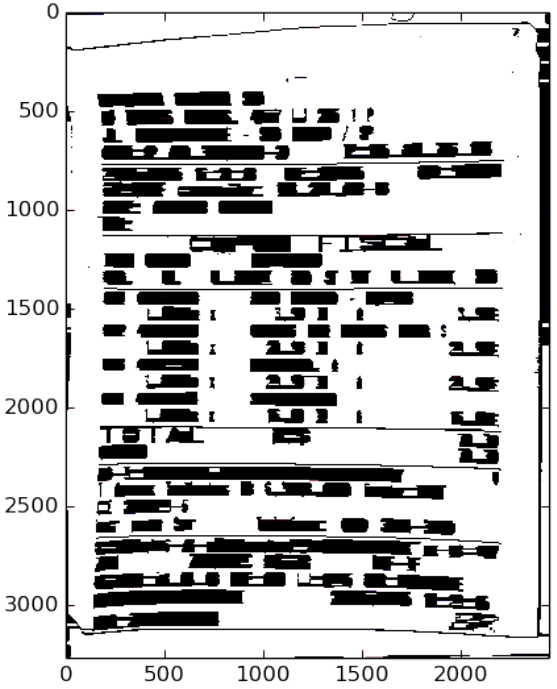


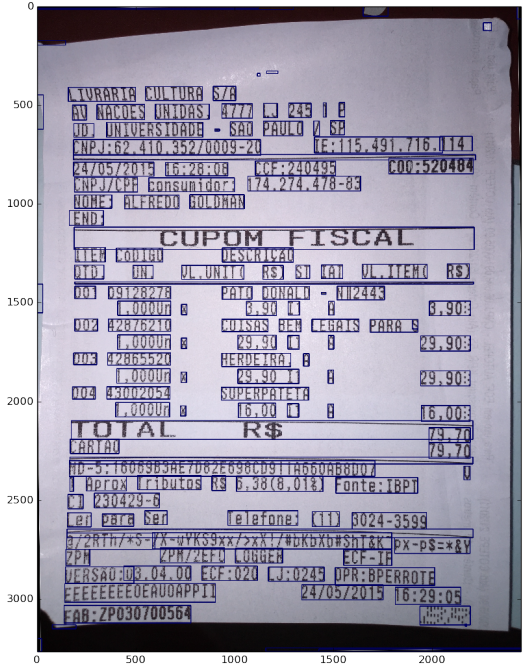


Fragmentando palavras



Extraindo palavras

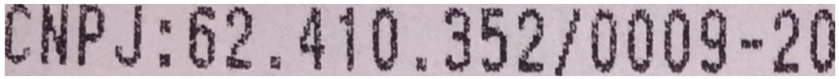




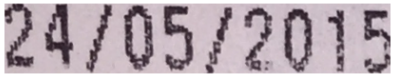
Resultado da extração

Aplicando o OCR

Encontrando padrões

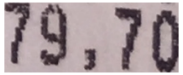


CNPJ:52.410~352¡0009~20



24/0512015

CDU: 520484



?9,70


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

Melhorias

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