

Algoritmo para segmentação de linhas de textos

Rafael G. Nagel

Instituto Federal de Santa Catarina

rafael.gustavo.nagel@gmail.com

05 de novembro, 2020

Tópicos

1 Aplicação

2 Filtros

3 Algoritmo

Aplicação

Pré-processamento para a aplicação de Reconhecimento Ótico de Caracteres (do inglês OCR)



Primeiro passo: binarização. Exemplo I

balbal albjbla hodsaohfiasdoi hfdihasof hiasfhoias
albjbla hodsaohfiasdoi hfdihasof hiasfhoias b
la hodsaohfiasdoi hfdihasof hiasfhoias blabalb
saohfiasdoi hfdihasof hiasfhoias blabalbal bal
asdoi hfdihasof hiasfhoias blabalbal balbalbal
hfdihasof hiasfhoias blabalbal balbalbal albjb
sof hiasfhoias blabalbal balbalbal albjbla hod
asfhoias blabalbal balbalbal albjbla hodsaohfi
as blabalbal balbalbal albjbla hodsaohfiasdoi

(a) Original

balbal albjbla hodsaohfiasdoi hfdihasof hiasfhoias
albjbla hodsaohfiasdoi hfdihasof hiasfhoias b
la hodsaohfiasdoi hfdihasof hiasfhoias blabalb
saohfiasdoi hfdihasof hiasfhoias blabalbal bal
asdoi hfdihasof hiasfhoias blabalbal balbalbal
hfdihasof hiasfhoias blabalbal balbalbal albjb
sof hiasfhoias blabalbal balbalbal albjbla hod
asfhoias blabalbal balbalbal albjbla hodsaohfi
as blabalbal balbalbal albjbla hodsaohfiasdoi

(b) Binarizado

Primeiro passo: binarização. Exemplo II

MASAYOSHI SON, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

YASUMITSU SHIGETA, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumblweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

(a) Original

MASAYOSHI SON, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

YASUMITSU SHIGETA, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumblweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

(b) Binarizado

Segundo passo (opcional): dilatação horizontal

balbal albjbla hodsaohfiasdoi hfdihasof hiasfh
albjbla hodsaohfiasdoi hfdihasof hiasfhoias b
la hodsaohfiasdoi hfdihasof hiasfhoias blabalb
saohfiasdoi hfdihasof hiasfhoias blabalbal bal
asdoi hfdihasof hiasfhoias blabalbal balbalbal
hfdihasof hiasfhoias blabalbal balbalbal albjb
sof hiasfhoias blabalbal balbalbal albjbla hod
asfhoias blabalbal balbalbal albjbla hodsaohf
as blabalbal balbalbal albjbla hodsaohfiasdoi

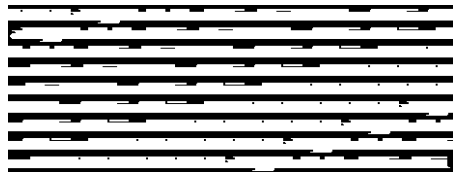
(a) Original binarizado



(c) $n = 10$

balbal albjbla hodsaohfiasdoi hfdihasof hiasfh
albjbla hodsaohfiasdoi hfdihasof hiasfhoias b
la hodsaohfiasdoi hfdihasof hiasfhoias blabalb
saohfiasdoi hfdihasof hiasfhoias blabalbal bal
asdoi hfdihasof hiasfhoias blabalbal balbalbal
hfdihasof hiasfhoias blabalbal balbalbal albjb
sof hiasfhoias blabalbal balbalbal albjbla hod
asfhoias blabalbal balbalbal albjbla hodsaohf
as blabalbal balbalbal albjbla hodsaohfiasdoi

(b) $n = 5$



(d) $n = 20$

Algoritmo de segmentação de linhas (sem dilatação)

- 1 Demonstração no GIMP
- 2 Percorre todas linhas horizontais de pixels para verificar se tem texto ou não
- 3 Limiar de contagem de pixels para cada linha (de pixels) definido como:

$$se \frac{\sum pixels_{branco}}{\sum pixels_{preto}} \geq limiar,$$

Então linha de pixels faz parte do texto e não do fundo.

Resultados. Exemplo 1

```
blablabla blabla blablabla blabla blablabla blabla  
blablabla blabla  
blabla bla  
  
blablabla blablablablabla blabla  
blablabla blabla blabla bla blabla  
blablablablablablablablablabla  
bla  
  
blablablablabla blablabla  
blablabla blabla blabla bla blabla blablabla blabla  
blabla bla bla blabla blabla bla blabla blablabla bla  
blabla blabla blabla blabla bla blabla blablabla blabla
```

(a) Original

```
blablabla blabla blablabla blabla blablabla blabla  
blablabla blabla  
blabla bla  
  
blablabla blablablablabla blabla  
blablabla blabla blabla bla blabla  
blablablablablablablablablabla  
bla  
  
blablablablabla blablabla  
blablabla blabla blabla bla blabla blablabla blabla  
blabla bla bla blabla blabla bla blabla blablabla bla  
blabla blabla blabla blabla bla blabla blablabla blabla
```

(b) Linhas segmentadas. Limiar = 0,1%

Resultados. Exemplo II

MASAYOSHI SON, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

YASUMITSU SHIGETA, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumblweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

(a) Original

MASAYOSHI SON, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

YASUMITSU SHIGETA, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumblweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

(b) Linhas segmentadas. Limiar = 0,1%

Resultados. Exemplo III

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

(a) Original

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

(b) Linhas segmentadas. Limiar = 1%

Resultados. Exemplo IV

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités¹ (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait pro-

(a) Original

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités¹ (féminine et masculine) qui président aux unions : Aphrodite et Eros². Dans cette série de huit couples divins, il en est un qui fait pro-

(b) Linhas segmentadas. Limiar = 10%

Segmentação de palavras (com dilatação)

- 1 Processo executado após segmentação das linhas (visto anteriormente)
- 2 Exige dilatação
- 3 Dentro de cada segmento de linha, verifica-se as colunas de pixels para verificar se tem texto ou não
- 4 Demonstração no GIMP

Resultados. Exemplo I

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

(a) Original

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

(b) Palavras segmentadas. $N = 10$ (dil.)

Resultados. Exemplo II

MASAYOSHI SON, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

YASUMITSU SHIGETA, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumblweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

(a) Original

MASAYOSHI SON, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

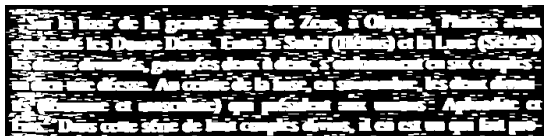
YASUMITSU SHIGETA, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumblweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

(b) Palavras segmentadas. $N = 7$ (dil.)

Resultados. Exemplo III

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples: un dieu-une déesse. Au centre de la frise, en surmembre, les deux divinités (féminine et masculine) qui président aux unions: Aphrodite et Eros. Dans cette série de huit couples divins, il en est un qui fait pro-

(a) Original



(b) Palavras dilatadas, $N = 6$

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Séléné) les douze divinités, groupées deux à deux, s'ordonnaient en six couples: un dieu-une déesse. Au centre de la frise, en surmembre, les deux divinités (féminine et masculine) qui président aux unions: Aphrodite et Eros. Dans cette série de huit couples divins, il en est un qui fait pro-

(c) Palavras segmentadas (com erros)

Resultados. Exemplo IV

balbal albjbla hodsaohfiasdoi hfdihasof hiasfhoias
albjbla hodsaohfiasdoi hfdihasof hiasfhoias b
la hodsaohfiasdoi hfdihasof hiasfhoias blabalb
saohfiasdoi hfdihasof hiasfhoias blabalbal bal
asdoi hfdihasof hiasfhoias blabalbal balbalbal
hfdihasof hiasfhoias blabalbal balbalbal albjb
sof hiasfhoias blabalbal balbalbal albjbla hod
asfhoias blabalbal balbalbal albjbla hodsaohfi
as blabalbal balbalbal albjbla hodsaohfiasdoi

(a) Original

balbal albjbla hodsaohfiasdoi hfdihasof hiasfhoias
albjbla hodsaohfiasdoi hfdihasof hiasfhoias b
la hodsaohfiasdoi hfdihasof hiasfhoias blabalb
saohfiasdoi hfdihasof hiasfhoias blabalbal bal
asdoi hfdihasof hiasfhoias blabalbal balbalbal
hfdihasof hiasfhoias blabalbal balbalbal albjb
sof hiasfhoias blabalbal balbalbal albjbla hod
asfhoias blabalbal balbalbal albjbla hodsaohfi
as blabalbal balbalbal albjbla hodsaohfiasdoi

(b) Caracteres segmentados, $N = 1$ (dil)

Material

- <https://github.com/RGNagel/line-segmentation-in-text-images>
- <https://github.com/RGNagel/apresentacao-SNCT-2020>

Fim