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

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 hiasfhalbjbla hodsaohfiasdoi hfdihasof hiasfhoias bla hodsaohfiasdoi hfdihasof hiasfhoias blabalbsaohfiasdoi hfdihasof hiasfhoias blabalbal balasdoi hfdihasof hiasfhoias blabalbal balbalbal hfdihasof hiasfhoias blabalbal balbalalbibal hfdihasof hiasfhoias blabalbal balbalbal albjbsof hiasfhoias blabalbal balbalbal albjbla hodsaohfiasfhoias blabalbal albjbla hodsaohfias blabalbal balbalbal albjbla hodsaohfias

(a) Original

balbal albjbla hodsaohfiasdoi hfdihasof hiasfhalbjbla hodsaohfiasdoi hfdihasof hiasfhoias bla hodsaohfiasdoi hfdihasof hiasfhoias blabalbsaohfiasdoi hfdihasof hiasfhoias blabalbal balasdoi hfdihasof hiasfhoias blabalbal balbalbal hfdihasof hiasfhoias blabalbal balbalalbal albjbsof hiasfhoias blabalbal albjbla hodasfhoias blabalbal albjbla hodasfhoias blabalbal albjbla hodsaohfias blabalbal balbalbal albjbla hodsaohfias

(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 Tumbleweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

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 Tumbleweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

(a) Original

(b) Binarizado

Segundo passo (opcional): dilatação horizontal



halhal albjbla bodsaobfiasdoi bfdibasof biasfb albjbla bodsaobfiasdoi bfdibasof biasfboias b la bodsaobfiasdoi bfdibasof biasfboias blabalb saobfiasdoi bfdibasof biasfboias blabalbal bal asdoi bfdibasof biasfboias blabalbal balbalbal bfdibasof biasfboias blabalbal balbalbal albjbla bod asfboias blabalbal balbalbal albjbla bodsaobfi asfboias blabalbal balbalbal albjbla bodsaobfi as blabalbal balbalbal albjbla bodsaobfias

(b) n = 5



(c)
$$n = 10$$

(d)
$$n = 20$$

Algoritmo de segmentação de linhas (sem dilatação)

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

$$se \frac{\sum pixeis_{branco}}{\sum pixeis_{preto}} >= limiar,$$

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

Resultados. Exemplo I

blablablabla blabla blablablabla blabla blablablabla blabla blablablabla blabla blabla bla

blablablabla blablablablablabla blabla blablablabla blabla blablabla bla blabla blablablablablablablablablablabla bla

blablablablablabla blablabla blablablabla blabla blablabla bla blabla blablablabla blabla blabla bla bla blabla blablabla bla blabla blablablabla bla blabla blabla blabla blablabla bla blabla blablablabla blabla

(a) Original

blablablabla blabla blablablabla blabla blablablabla blabla blablablabla blabla blabla bla

blablablabla blablablablabla blabla blablablabla blabla blablabla bla blabla blablablablablablablablablablabla

blablablablablabla blablabla blablablabla blabla blablabla bla blabla blablablabla blabla blabla bla bla blabla blablabla bla blabla blablablabla bla blabla blabla blabla blablabla bla blabla blablablabla blabla

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 Tumbleweed 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 Tumbleweed 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

Lorum 'sperim delor sit amet, ametlehir sadifraing lille, sed diam novimy virinad tempor invision ut labore et delore magna aliquijam erad, sed diam refujchia. At revo ses et acausam et fusdo shi delores et u rehim. Net dida bassel gubergum, no seu tabimata sankhus est horem ipsum delor sit amet, horem ipsum delor sit amet, consective sadificang state, sed diam nevung termed tempor invisioned set labore th delore magna diquipam erad, sed diam reluptuu. At nevo ses et acausam et fusdo due delores et u nehim. Sed dita hasel qu'auquem en no seu takumatu samblus set Lorum ipsum delor sit amet.

(a) Original

Lorem trsum dolor sit amed, consolerur sadistang eletr, sad diam nonumy sirmad temper invidunt ut labore et dolori magna diiguyam erat, sid siam ortuntua ett euro esi et acusam et pusto diu dolori et sa rehim. Het dida hasid quitorgen, no seu sadimata sambbus est ficrem irsum ador sit amel. Lorem irsum dolor sit amel. Lorem irsum ador sit amel. Lorem irsum dolor sit amel. Lorem irsum dolor sit amel. Lorem irsum dolor sit amel.

(b) Linhas segmentadas. Limiar = 1%

Resultados. Exemplo IV

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait réprésenté 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 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

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait réprésenté 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 divinités (féminine et masculine) qui président aux unions. Aphrodite et fros. Dans cette série de huit couples divins, il en est un qui fait pro-

(b) Linhas segmentadas. Limiar =10%

Segmentação de palavras (com dilatação)

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

Resultados. Exemplo I

form ifrsem dolor sid amol, considerur sadipsaing elite, sed diam nowing armed temper invitant at labore the dolore magna aliquyam erad, sed diam metaphua Al mos les et accusam et finale due solores et eu rehim. Het elite based queequen, no seu takimata sonkhus ett harem ipsum dolor sit omet hirem ipsum dolor sit omet harem ipsum adipayam erad, sed diam notuptua. At vere ses et accusam et pisto aus dolors et sa rehim. Itel that hard queuque ren no seu takumata samblus est toram ipsum dolor sit amet.

(a) Original

(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 Tumbleweed 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, 142, president and CEO, is the master Net lempire 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 Weblor mobile Net-based ventures in Japan and the U.S., including Tumbleweed 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 réprésenté les Douze Dieux. Entre le Solcil (Hélios) et la Lune (Sélénè) les diouze divinités, groupées deux à deux, s'ordonnaient en six couples : di dieu une décsse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions: Aphrodite et folgs. 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 lla grande Istatue de Zeus, là l'Olympie, Phidias avait réprésentélles Douzel Dieux. Entre le Soleil (Hélios) et lla Lune (Séléné) les gouzel divinités, groupées deux la deux is fordonnaientens ixi couples l'un dieu une décase. l'Aucentre de la frise, en sumombre, les deux divinités (étaininne let masculine) qui président aux unions l'Approdire et gross. Dans (cette lécric de l'huit couples divins, il ten est un qui fait pro-

(c) Palavras segmentadas (com erros)

Resultados. Exemplo IV

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 balbalal albjb sof hiasfhoias blabalbal balbal albjbla hod asfhoias blabalbal balbal albjbla hodsaohfi as blabalbal balbal albjbla hodsaohfiasdoi

(a) Original

balbal albibla hodsachfilasdol hfdihasof hilasfh albibla hodsachfilasdol hfdihasof hilasfholas b la hodsachfilasdol hfdihasof hilasfholas blabalba sachfilasdol hfdihasof hilasfholas blabalbal bal asdol hfdihasof hilasfholas blabalbal balbalbal hfdihasof hilasfholas blabalbal balbalbal albib sof hilasfholas blabalbal balbalbal albibla asfholas blabalbal balbalbal albibla hodsachfilasdol

(b) Characteres segmentados, N = 1 (dil)

Material

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

Fim