THE DEVELOPER'S CONFERENCE

OCR: o que é e como usar?

Trilha Machine Learning

Guilherme Malta
Graduando em Ciência da Computação (UFRGS)
Data Science Intern (Poatek)

Agenda

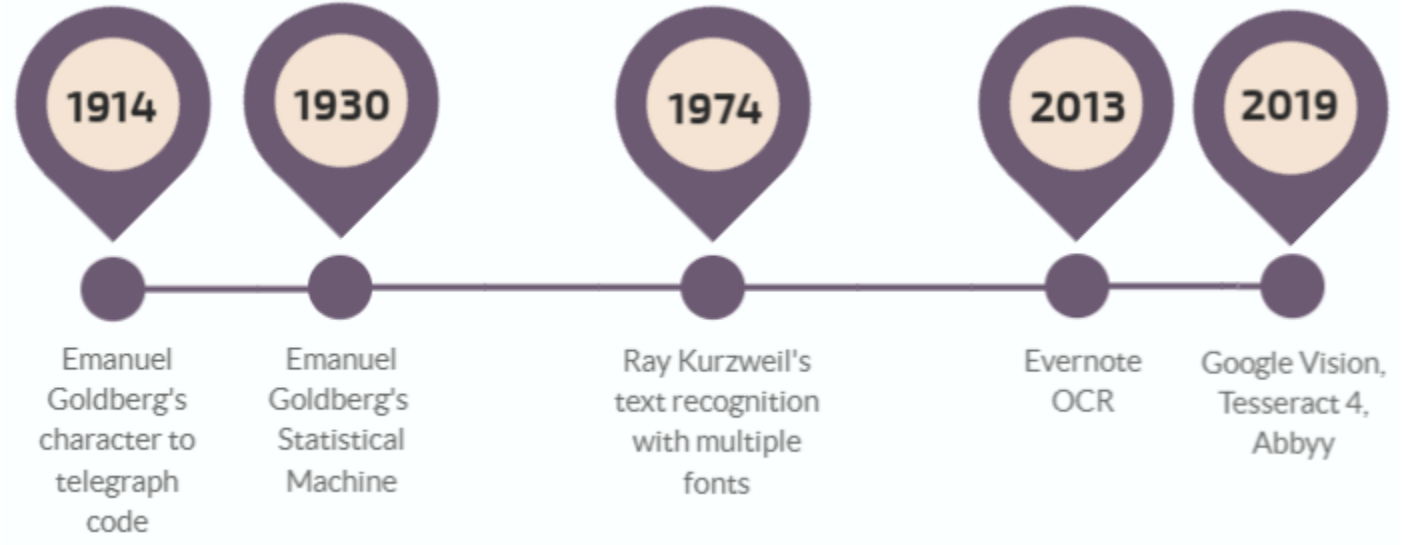
THE DEVELOPER'S CONFERENCE

- O que é OCR e para que serve?
- Tipos mais comuns de OCR
- Ferramentas de OCR usadas hoje
- Tesseract e Pytesseract
- Case: busca de informações em documentos escaneados usando Pytesseract

O que é OCR?

THE DEVELOPER'S CONFERENCE

- ***O**ptical **C**haracter **R**ecognition:* tecnologia de reconhecimento de texto em arquivos de imagem;



Tipos mais comuns

- Reconhecimento de dados em documentos pessoais (e.g. carteira de identidade, passaporte, carteira de motorista…);

towardsdatascience.com/how-to-use-optical-character-recognition-for-security-system-development-9dc57258afa3

Tipos mais comuns

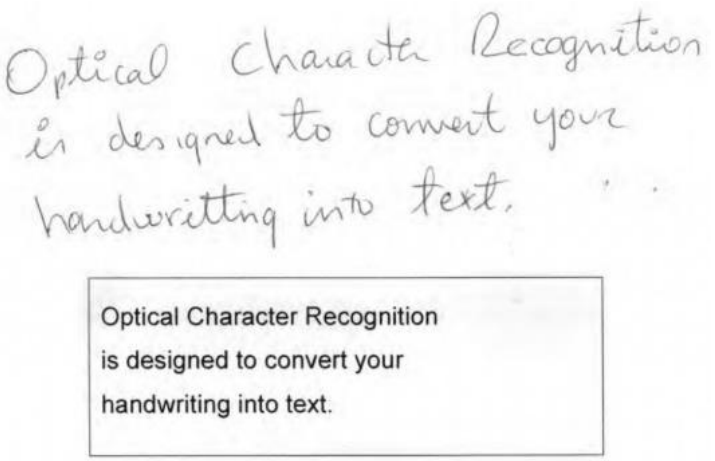
- Reconhecimento de placas de veículos;

https://towardsdatascience.com/automatic-license-plate-detection-recognition-using-deep-learning-624def07eaaf

Tipos mais comuns

THE DEVELOPER'S CONFERENCE

- Reconhecimento de textos escritos à mão;



www.rsipvision.com/deep-learning-for-ocr/

Tipos mais comuns

THE DEVELOPER'S CONFERENCE

- Reconhecimento de texto "*in-the-wild*";

towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201aa

Tipos mais comuns

THE DEVELOPER'S CONFERENCE

- Transformação de documentos escaneados em documentos pesquisáveis;

medium.com/@InDataLabs/ocr-algorithm-improve-and-automate-business-processes-c212c7e355a5

Opções

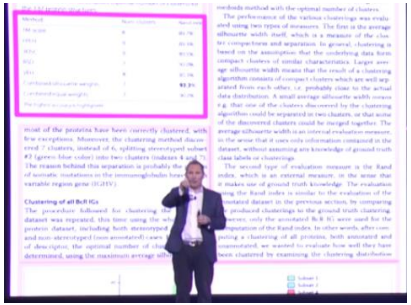
THE DEVELOPER'S CONFERENCE

- Treinar modelo from scratch
- Especializar modelo pré-treinado
- Utilizar modelo/ferramenta treinada

Ferramentas atuais

THE DEVELOPER'S CONFERENCE

- Google Vision
- Amazon Textract
- Tesseract 4
  - Open source



Method	Num. clusters	Rand index
TM-score	8	89.7%
FPFH	9	89.3%
3DSC	9	89.5%
RSD	7	92.0%
VFH	8	85.3%
Combined silhouette weights	7	92.2%
Combined equal weights	7	90.2%

aws.amazon.com/pt/textract
https://cloud.google.com/vision
https://github.com/tesseract-ocr/tesseract

Ferramentas atuais

THE DEVELOPER'S CONFERENCE

Research of Optimal Machine Learning OCR Approach

Name/ Dataset	Internet_ driver_ids (333 lines)	Recognition accuracy, %	Internet_ pasport (1016 lines)	Recognition accuracy, %	Driver_ids of our project (474 lines)	Recognition accuracy, %
Microsoft computer-vision	147	44,1%	164	16,1%	188	39,7%
OCR Space	141	42,3%	372	36,6%	146	30,8%
Abbyy	174	52,3%	497	48,9%	168	35,44%
Google Vision	269	80,8%	657	64,7%	381	80,4%
Cloud Mersive	79	23,7%	163	16%	56	11,81%
Tesseract 3 (open source)	65	19,5%	243	23,9%	59	12,44%
Tesseract 4 (open source)	123	36,9%	465	45,8%	119	25,1%

mobidev

towardsdatascience.com/how-to-use-optical-character-recognition-for-security-system-development-9dc57258afa3

Tesseract (1995)

THE DEVELOPER'S CONFERENCE

- Desenvolvido pela HP à partir de 1984;
- Lançamento open source em 2005;
- Desenvolvido pela Google desde então

https://static.googleusercontent.com/media/research.google.com/pt-BR//pubs/archive/33418.pdf

Tesseract V 2.0 (2007)

THE DEVELOPER'S CONFERENCE

- Ray Smith (Google Inc.)

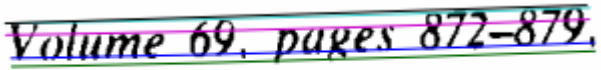


Fig. 1. An example of a curved fitted baseline.

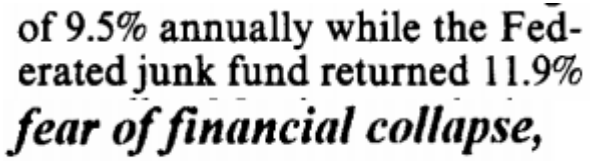


Fig. 3. Some difficult word spacing.

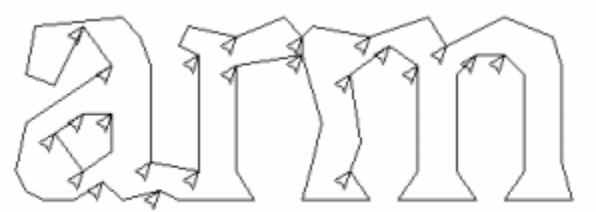


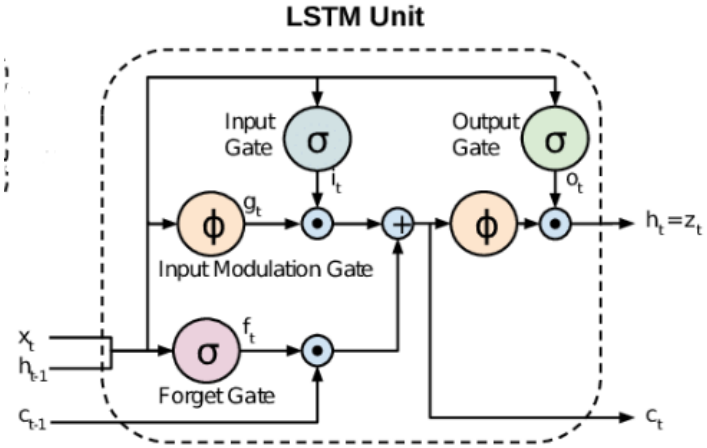
Fig. 4. Candidate chop points and chop.

https://static.googleusercontent.com/media/research.google.com/pt-BR//pubs/archive/33418.pdf

Tesseract release notes Oct 29 2018 - V4.0.0

THE DEVELOPER'S CONFERENCE

- **New OCR engine**
  - Added a new OCR engine that uses neural network system based on LSTMs, with major accuracy gains.
  - This includes new training tools for the LSTM OCR engine. A new model can be trained from scratch or by fine tuning an existing model.
  - Added trained data that includes LSTM models to 123 languages.



https://www.researchgate.net/figure/RNN-simple-cell-versus-LSTM-cell-4_fig2_317954962

Pytesseract

THE DEVELOPER'S CONFERENCE

- Wrapper para Tesseract
- Aceita formatos .jpeg, .png, .gif, .bmp, .tiff e outros
- Fornece estruturas de dados que permitem solução para problemas mais complexos
  - Bounding boxes
  - Lista de caracteres reconhecidos
  - Metadados das palavras reconhecidas

+ Image processing

THE DEVELOPER'S CONFERENCE

- Técnicas de processamento de imagem frequentemente usadas
  - Resize
  - Filtros (gaussiano, mediana, etc)
  - Cropping
  - Outros

THE
DEVELOPER'S
CONFERENCE

Case

THE DEVELOPER'S CONFERENCE

Guilherme Malta

Data Science Intern
Poatek

linkedin.com/in/gmalta

github.com/GhMalta

guilherme.malta@poatek.com

Obrigado!

THE DEVELOPER'S CONFERENCE