AiBox Lab

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

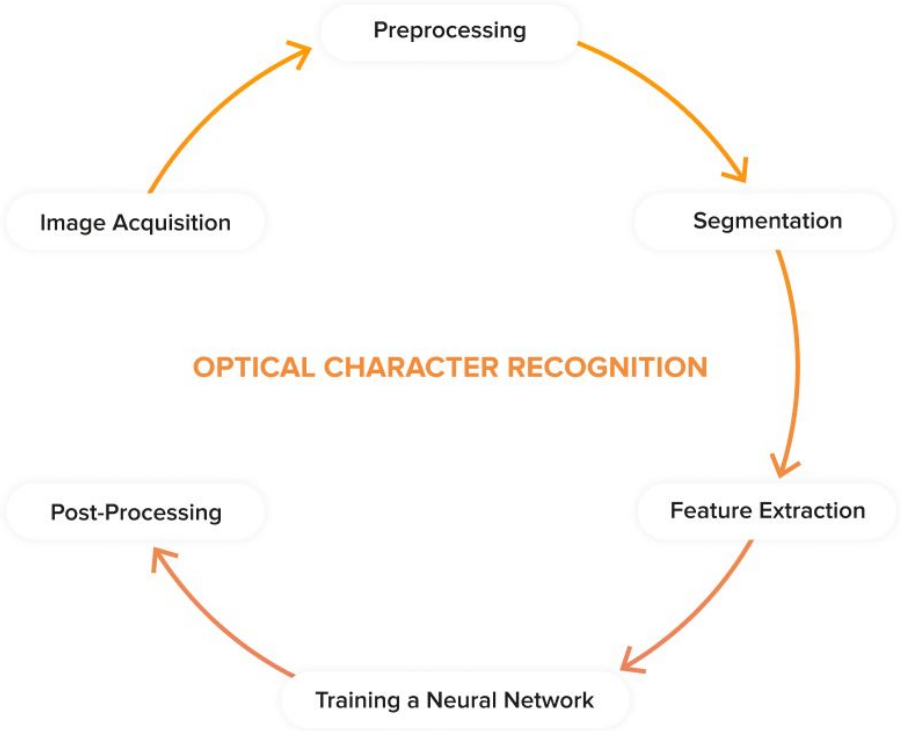
Pós-processamento de OCR

AiBox Summer School

Equipe 3

- Ariany Ferreira
- Gabriel Moura
- Giuseppe Fiorentino
- João Victor Galdino

Etapas para criar um mecanismo de OCR

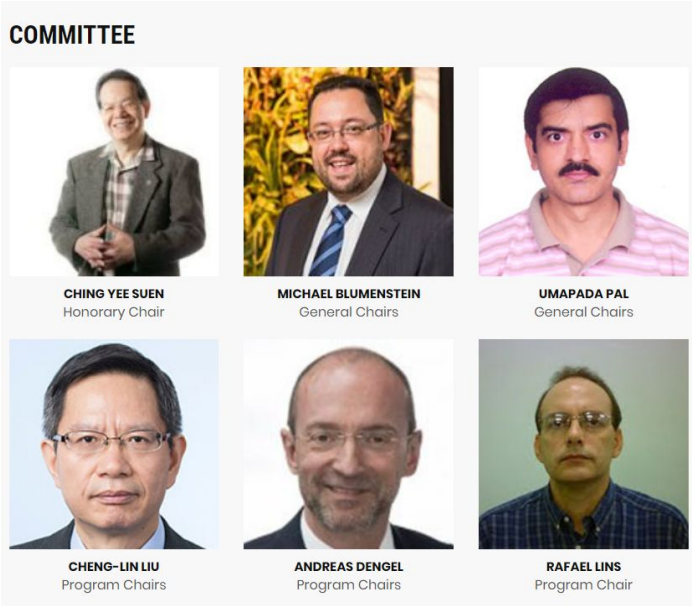


Natas vs CCC

- Natas nos dá uma lista de sugestões para que possa ser utilizado na correção. A indicação é utilizar o primeiro elemento da lista como o elemento a substituir.
  - Método de processamento de documentos históricos com foco no estudo no neologismo.
- CCC (Context-based Character Correction) é um método de correção de carácter baseado no contexto.
  - Modelo desenvolvido pelo time: Clova AI, NAVER/LINE Corp., South Korea

International Conference on Document Analysis and Recognition - ICDAR

- ICDAR 2019 Competition on Post-OCR Text Correction



ICDAR 2019

15th International Conference on Document Analysis and Recognition
20-25 September 2019 | International Convention Centre Sydney, Australia

CCC

BERT - *Bi-directional Encoder Representations from Transformers* - é um algoritmo de pesquisa. Ele ajuda a entender pesquisas de uma forma mais parecida com os humanos.

Base de dados

- Os dados utilizados foram **manualmente modificados** gerando erros aleatórios simulando assim os erros gerados por uma OCR.
- Artigos científicos na área da **educação**.

Análise exploratória/descritiva da base

- Colab

Métrica de avaliação

- F1 Score de cada documento
  - Média harmônica do precision e recall
    - Precision
      - Que proporção de identificações positivas estava realmente correta?
    - Recall
      - Que proporção de positivos reais foi identificada corretamente?

Resultados da comparação

- Natas
  - Média 0.259
  - Mediana 0.167
  - Desvio Padrão 0.247
- CCC
  - Não houve dados suficientes para realizar as análises
    - Muito consumo de hardware e um modelo não otimizado.

Obrigado!