ALBERTe

Pré-processamento

Aquisição da Imagem

Segmentação

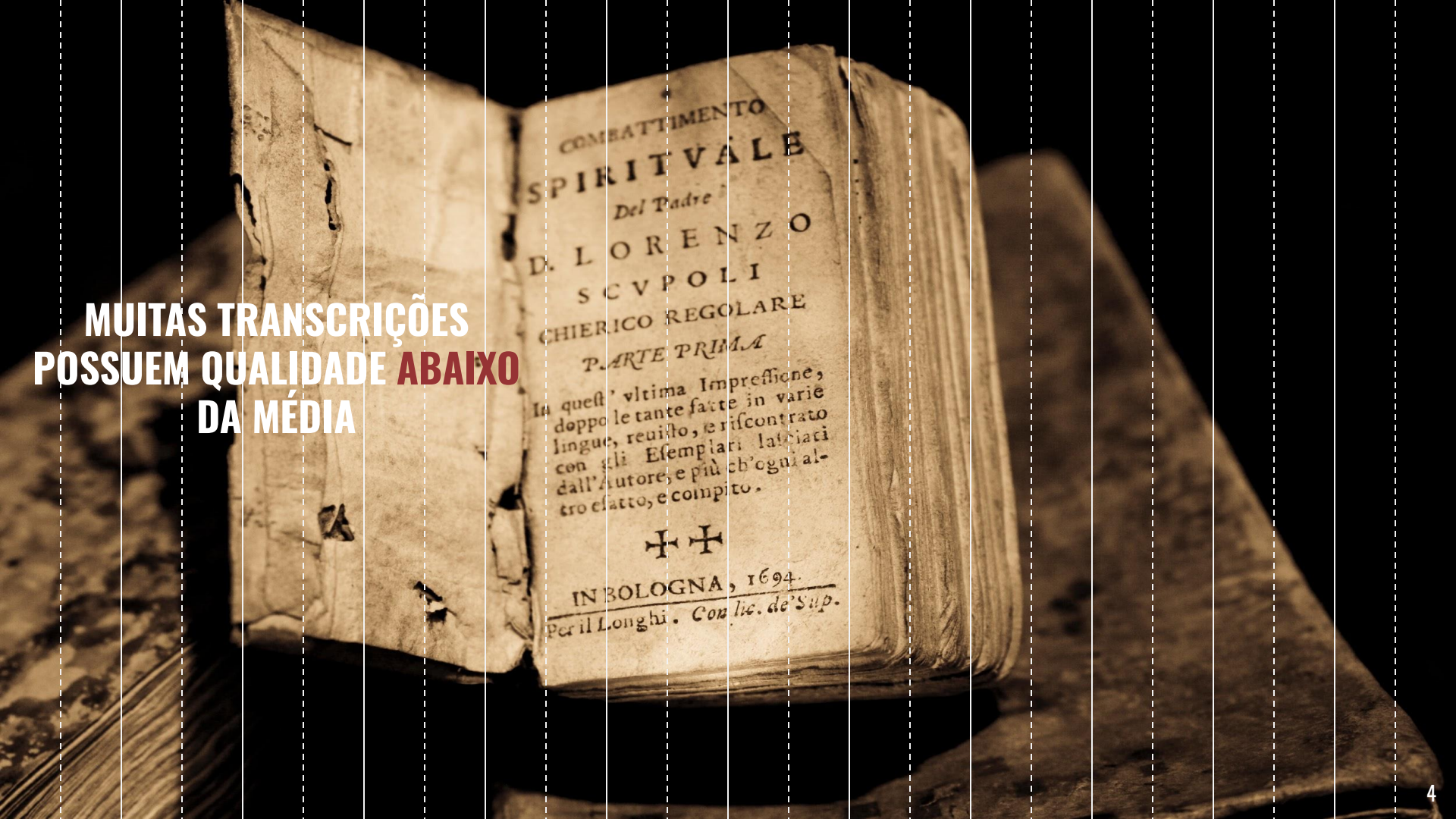
Pós-Processamento

Extração de Características

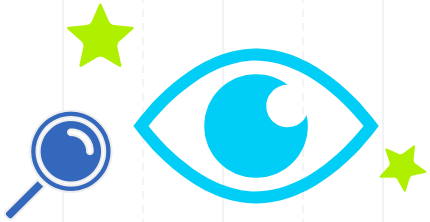
Treinamento da Rede Neural

A precisão das tecnologias de Optical Character Recognition (OCR) afeta consideravelmente o modo como os documentos digitais são indexados, acessados e explorados.

Impact analysis of OCR quality on research tasks in digital archives, 2015



MUITAS TRANSCRIÇÕES POSSUEM QUALIDADE ABAIXO DA MÉDIA



Pós-processamento de OCR beneficia tanto documentos digitalizados como não-digitalizados.

ORIGINAL VS CORREÇÃO

There were **oncc** a man and a **wman** who had long in vain wished for a child.

At length the woman hoped that God was about to grant her desire.

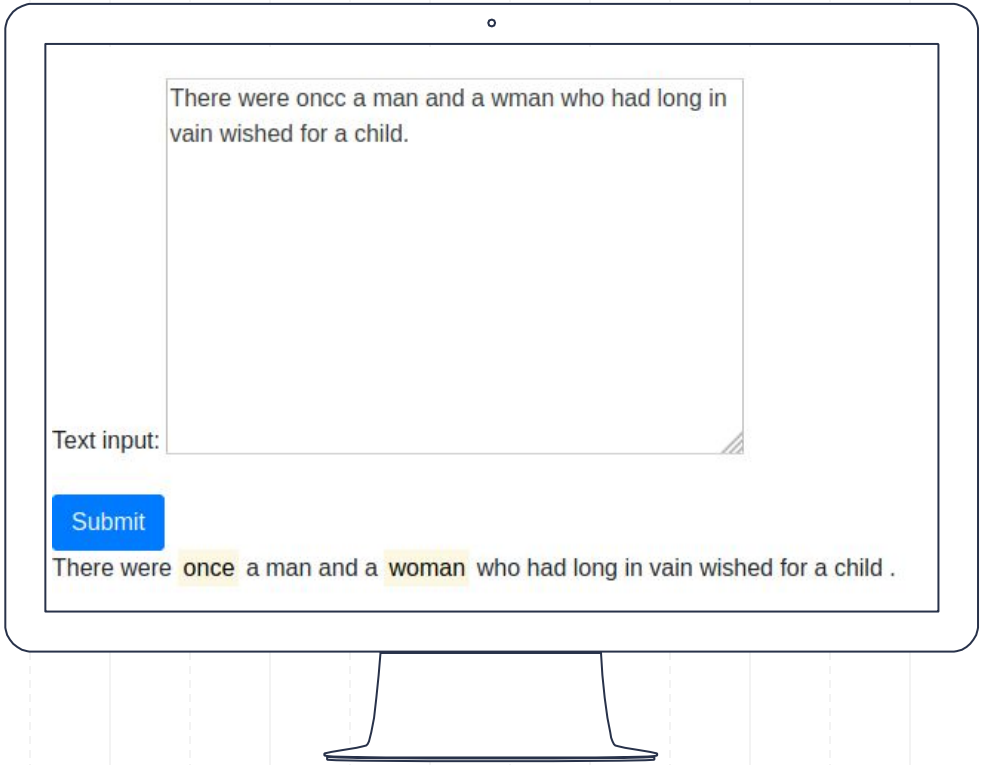
There were **once** a man and a **woman** who had long in vain wished for a child .

At length the woman hoped that God was about to grant her desire .

TEXTO ORIGINAL

ALBERT + PyEnchant

TEXTO CORRIGIDO

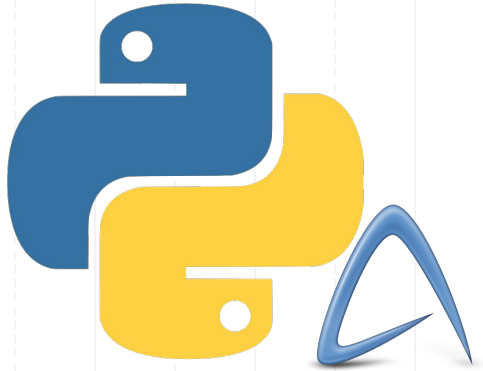


ALBERT

Usa camadas repetidas, o que resulta em uma pequena área ocupada por memória. Entretanto, o custo computacional permanece semelhante a uma arquitetura semelhante ao BERT.

PyEnchant

PyEnchant combina toda a funcionalidade da biblioteca Enchant subjacente com a flexibilidade do Python e uma ótima interface orientada a objetos “Pythonic”.



F1-SCORE - DOCUMENTO ORIGINAL

	Valor
Média	0.97
Mediana	1.0
Desvio Padrão	0.162

F1-SCORE - DOCUMENTO ORIGINAL COM ALBERT+PyEnchant

	Valor
Média	0.91
Mediana	1.0
Desvio Padrão	0.190

OBRIGAD@!

Dúvidas?

Ariany Ferreira

Giuseppe Fiorentino

João Victor Galdino

-

AiBox Summer School