Identificação de Escritor por Transformada SIFT e SVM-Linear na Língua Portuguesa

TRABALHO DE CONCLUSÃO DE CURSO DE ENGENHARIA ELÉTRICA

JOÃO PAULO LOPES SANCHEZ

ORIENTADOR: CELSO AP. DE FRANÇA

Problema e Objetivo

- **Problema**

  A autoria de textos manuscritos é um problema importante em diversas áreas:

  - Autoria de documentos legais;
  - Investigação de plágio;
  - Análise de documentos históricos.

- **Objetivo**

  Este trabalho propõe um método para identificação de autor em textos em português utilizando:

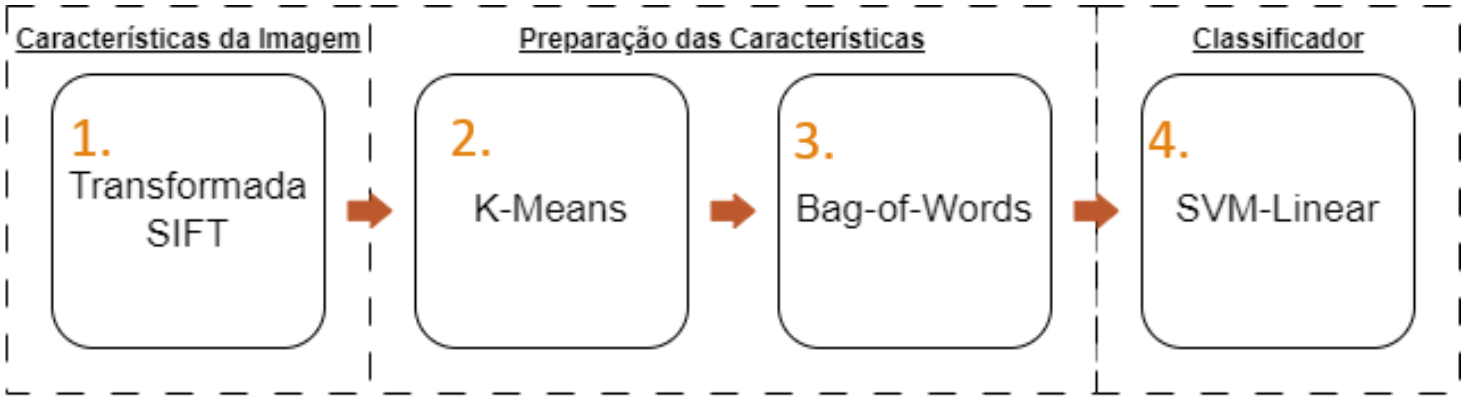
  - Transformada SIFT;
  - K-means;
  - Bag-of-Words;
  - SVM-Linear.

Definição de Escopo

- **Modelo de análise**
  - Identificação X Verificação de Escritores
- **Características da imagem**
  - Estrutura textual local X Estrutura textual global
- **Classificadores**
  - *Convolutional Neural Networks* X *Support Vector Machine*

Visão Geral – Metodologia

1. Extração de características.
2. Agrupamento das características
3. Criação de um modelo.
4. Treinamento de um classificador



Scale Invariant Feature Transform (SIFT)

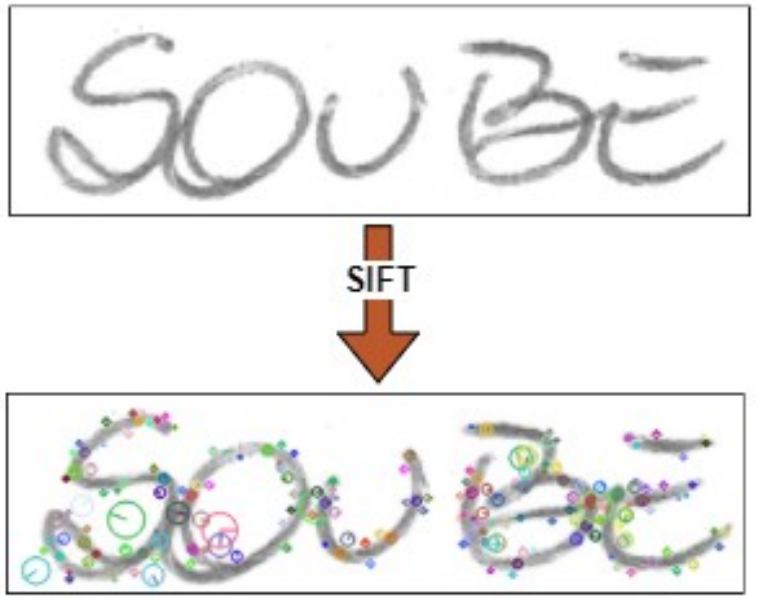
ALGORITMO PARA EXTRAÇÃO DE CARACTERÍSTICAS LOCAIS EM IMAGENS DIGITAIS.

INVARIANTE À ESCALA, ROTAÇÃO E PARCIALMENTE INVARIANTE À ILUMINAÇÃO E MUDANÇAS DE PONTO DE VISTA.

SIFT
Transformada SIFT

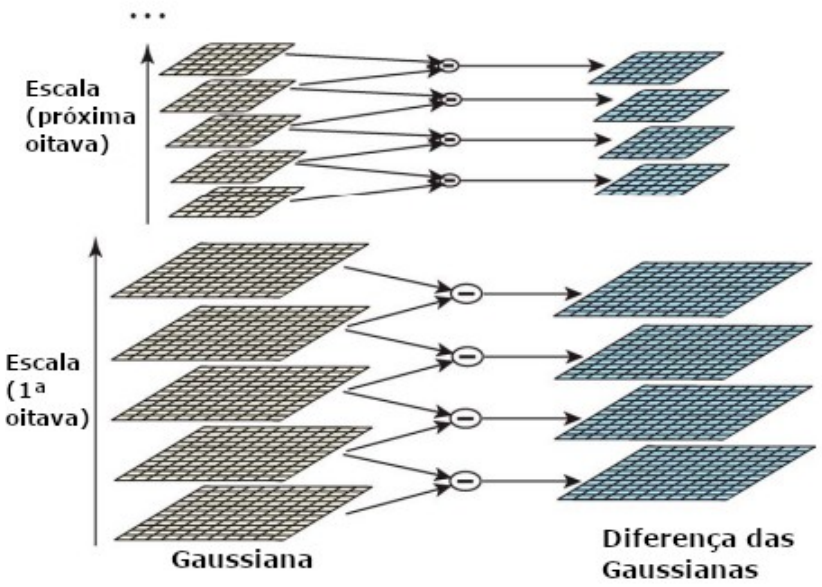
1. Detecção de pontos de interesse:
   - Localização de pontos com alta variação de escala e espacial.
   - A sub-amostragem em diferentes escalas aumenta a robustez.

2. Descritor de características:
   - Histograma de gradientes em torno do ponto de interesse.
   - Alta dimensionalidade (128 elementos) para melhor discriminação.

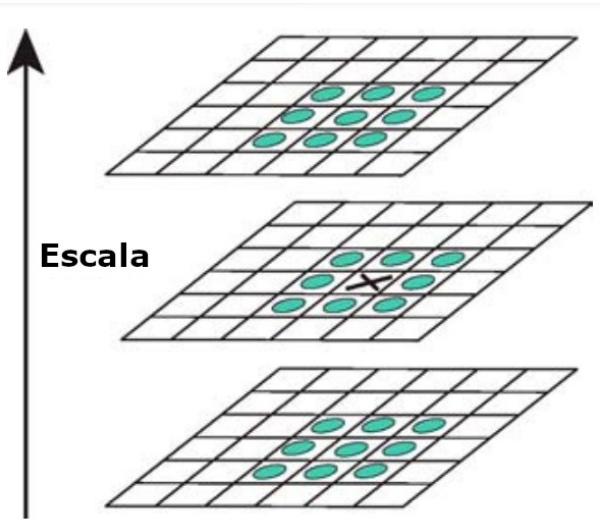


SIFT
Definição dos Pontos-Chave

- Suavização da imagem com filtros gaussianos em diferentes imagens
- Cálculo do mapa de diferença de escala

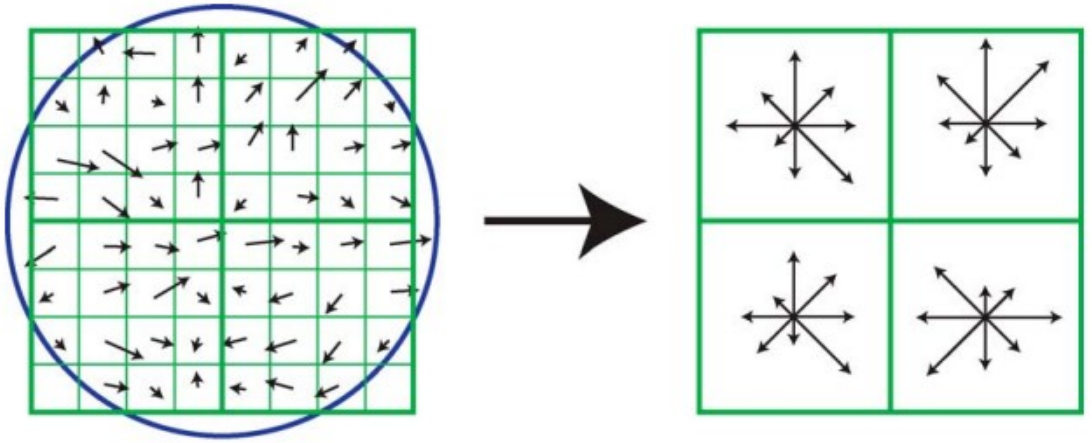


- Comparação com todos pontos adjacentes em escala e proximidade.
- Filtragem dos pontos-chave com critérios de estabilidade e contraste.



SIFT
Cálculo do Descritor de Ponto-Chave

- Janela Espacial:
  - Definição de uma janela espacial em torno do ponto de interesse.
  - Divisão da janela em sub-regiões.
- Histograma de Gradientes em Sub-Regiões:
  - Cálculo do histograma de gradientes em cada sub-região.
  - Concatenação dos histogramas para formar um vetor de características.
- Descritor de Ponto-Chave:
  - Vetor com 128 elementos que descreve a textura local da imagem.



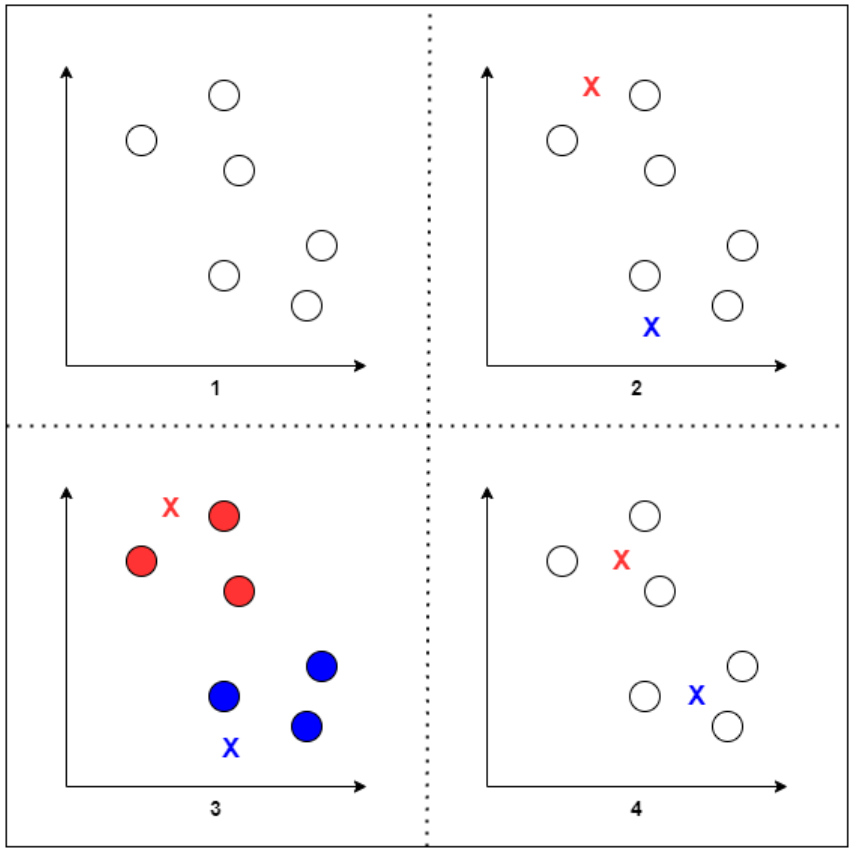
K-Means

ALGORITMO DE AGRUPAMENTO NÃO-SUPERVISIONADO QUE DIVIDE OS DADOS EM K GRUPOS.

OBJETIVO: MINIMIZAR A SOMA DAS DISTÂNCIAS ENTRE CADA PONTO E O CENTROIDE DE SEU CLUSTER.

K-Means Algoritmo

1. Escolha de k: Definir o número de clusters desejados.
2. Inicialização: Selecionar aleatoriamente k centroides no espaço de dados.
3. Atribuição: Atribuir cada ponto ao cluster com o centroide mais próximo.
4. Atualização: Recalcular os centroides como a média dos pontos em cada cluster.
5. Repetição: Repetir os passos 3 e 4 até que os centroides não se movam significativamente ou um número máximo de iterações seja atingido.



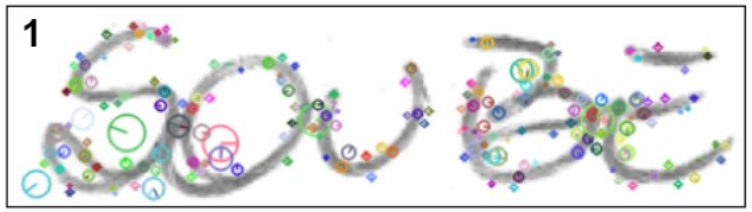
Bag-of-(Visual)-Words

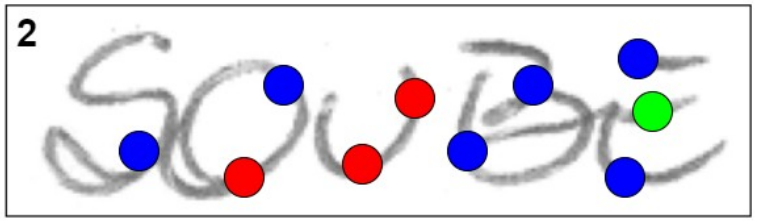
MODELO PARA REPRESENTAR DOCUMENTOS COMO COLEÇÕES DE “PALAVRAS VISUAIS”.

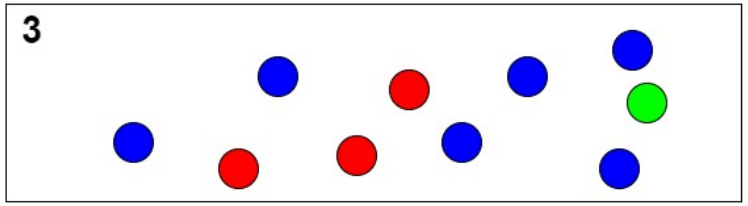
OBJETIVO: AGRUPAR “PALAVRAS” DE FORMA ABSTRATA E POR SEMELHANÇA NAS CARACTERÍSTICAS DA IMAGEM

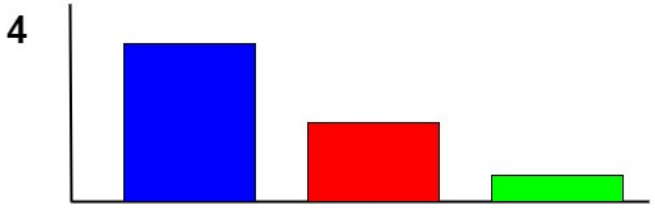
Bag-of-Words Funcionamento

- Um vocabulário é criado a partir das características mais frequentes.
- As "palavras visuais" são características visuais extraídas da imagem.
- O *codebook* é um dicionário que mapeia características visuais para "palavras visuais".
- O histograma de "palavras visuais" é usado para representar a imagem.









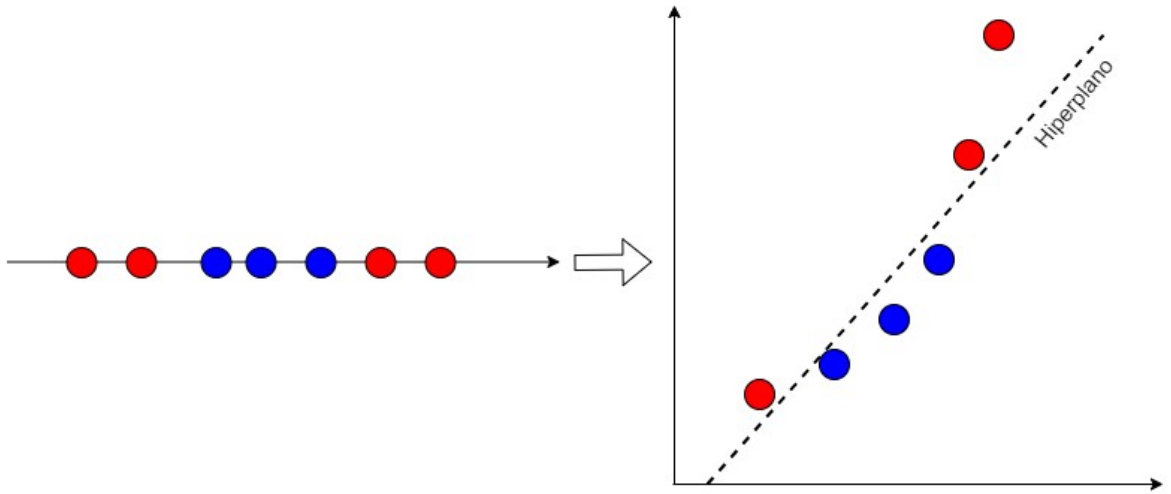
Support Vector Machine

ALGORITMO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO PARA CLASSIFICAÇÃO E REGRESSÃO.

OBJETIVO: ENCONTRAR UM HIPERPLANO QUE MAXIMIZA A MARGEM ENTRE AS CLASSES DE TREINAMENTO

Support Vector Machine Funcionamento

- Um kernel é definido
- O hiperplano é encontrado maximizando a margem entre as classes.
- Novos pontos de dados são classificados de acordo com o lado do hiperplano em que se encontram.



Support Vector Machine Linear

- Kernel linear não é eficaz para conjuntos não-linearmente separáveis
- Para altas dimensões apresenta bons resultados

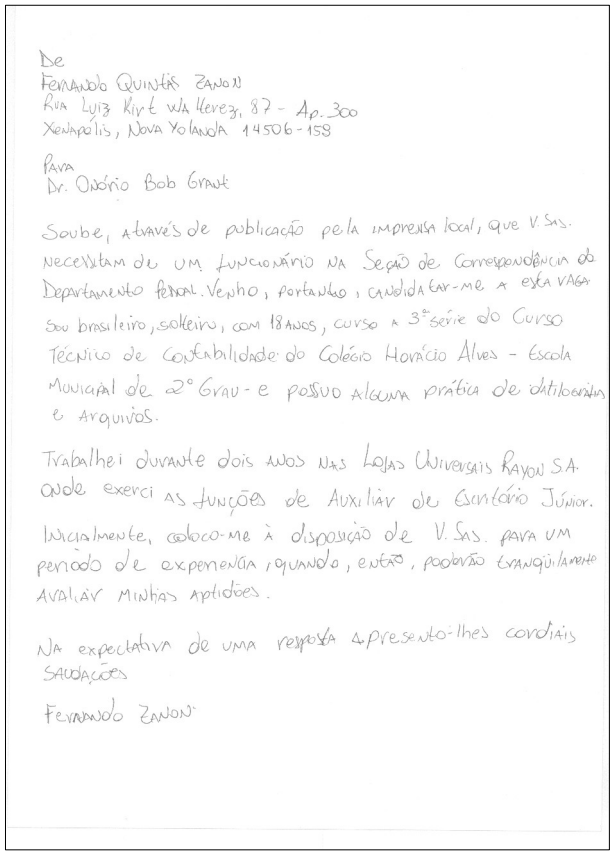
$$K(x_1, x_2) = \varphi(x_1)^T \varphi(x_2) \rightarrow \varphi(x) = x$$

Brazilian Forensic Letter Database (BFL)

CONTÉM CARTAS FORENSES DE DIFERENTES AUTORES COM AMPLA VARIEDADE DE ESTILOS DE ESCRITA, CALIGRAFIAS E PADRÕES DE ESCRITA MANUAL.

ÚTIL PARA INVESTIGAÇÕES CRIMINAIS, CASOS DE FRAUDE E ANÁLISE DE DOCUMENTOS.

BFL
Exemplos de Documentos



De
Fernando Quintas Zanon
Rua Luiz Kirt Wallerez, 87 - Ap. 300
Xenápolis, Nova Yolanda 14506-159

Para
Dr. Onório Bob Grant

Soube, através de publicação pela imprensa local, que V. Sas. necessitam de um funcionário na Seção de Correspondência do Departamento Pessoal. Venho, portanto, candidatar-me a esta vaga.

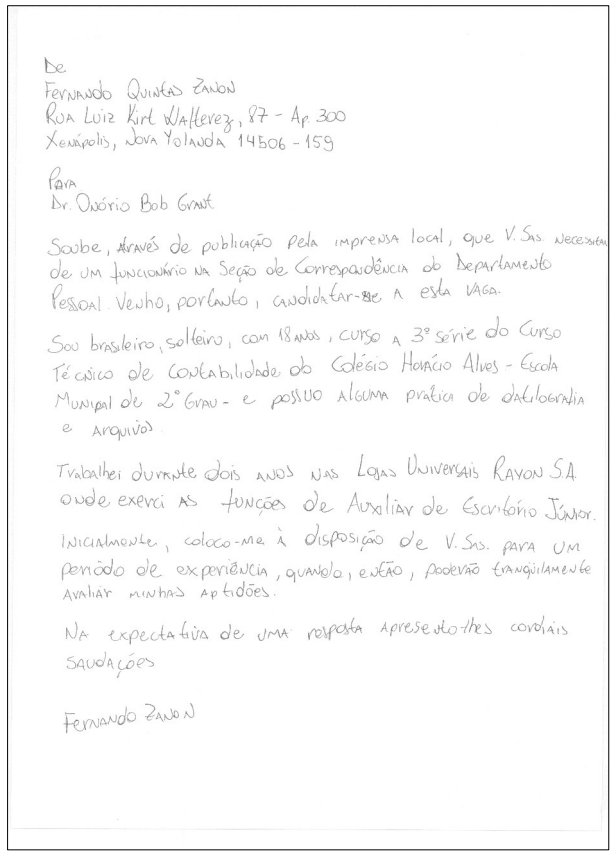
Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Horácio Alves - Escola Municipal de 2º Grau - e possuo alguma prática de datilografia e arquivos.

Trabalhei durante dois anos nas Lojas Universais Rayon S.A. onde exerci as funções de Auxiliar de Escritório Júnior.

Inicialmente, coloco-me à disposição de V. Sas. para um período de experiência, quando, então, poderão tranqüilamente avaliar minhas aptidões.

Na expectativa de uma resposta apresento-lhes cordiais saudações

Fernando Zanon



De
Fernando Quintas Zanon
Rua Luiz Kirt Wallerez, 87 - Ap. 300
Xenápolis, Nova Yolanda 14506-159

Para
Dr. Onório Bob Grant

Soube, através de publicação pela imprensa local, que V. Sas. necessitam de um funcionário na Seção de Correspondência do Departamento Pessoal. Venho, portanto, candidatar-me a esta vaga.

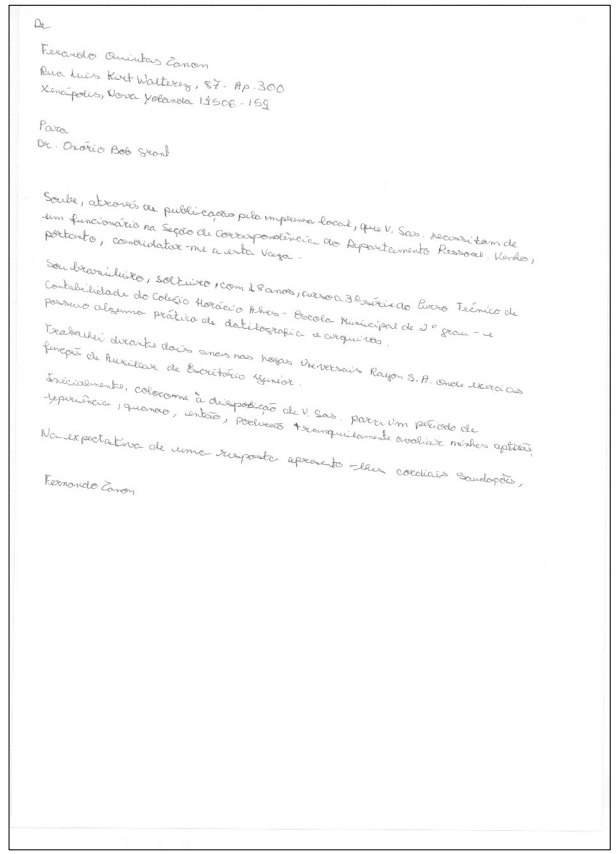
Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Horácio Alves - Escola Municipal de 2º Grau - e possuo alguma prática de datilografia e arquivo.

Trabalhei durante dois anos nas Lojas Universais Rayon S.A. onde exerci as funções de Auxiliar de Escritório Júnior.

Inicialmente, coloco-me à disposição de V. Sas. para um período de experiência, quando, então, poderão tranqüilamente avaliar minhas aptidões.

Na expectativa de uma resposta apresento-lhes cordiais saudações

Fernando Zanon



De
Fernando Quintas Zanon
Rua Luiz Kirt Wallerez, 87 - Ap. 300
Xenápolis, Nova Yolanda 14506-159

Para
Dr. Onório Bob Grant

Soube, através de publicação pela imprensa local, que V. Sas. necessitam de um funcionário na Seção de Correspondência do Departamento Pessoal. Venho, portanto, candidatar-me a esta vaga.

Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Horácio Alves - Escola Municipal de 2º grau - e possuo alguma prática de datilografia e arquivos.

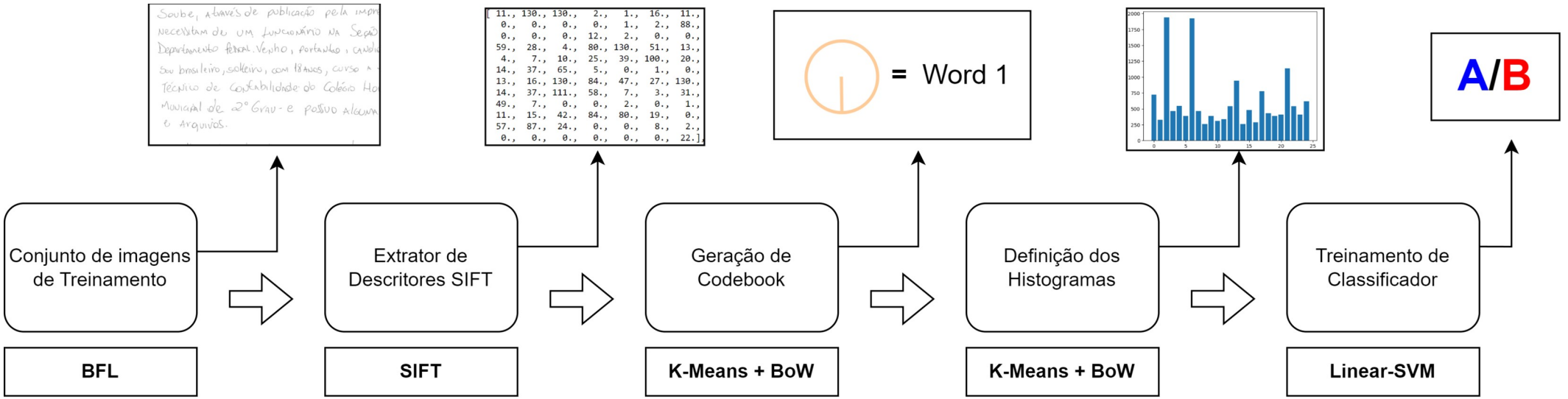
Trabalhei durante dois anos nas Lojas Universais Rayon S.A. onde exerci as funções de Auxiliar de Escritório Júnior.

Inicialmente, coloco-me à disposição de V. Sas. para um período de experiência, quando, então, poderão tranquilamente avaliar minhas aptidões.

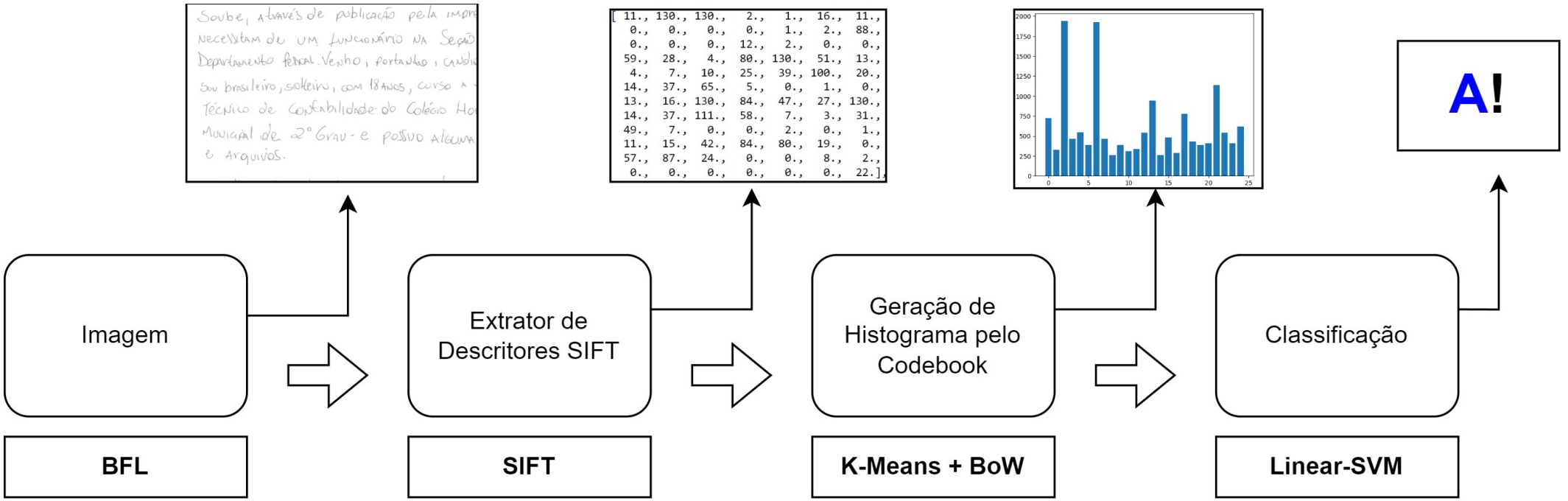
Na expectativa de uma resposta apresento-lhes cordiais saudações,

Fernando Zanon

Metodologia Treinamento



Metodologia
Classificação de Nova Imagem



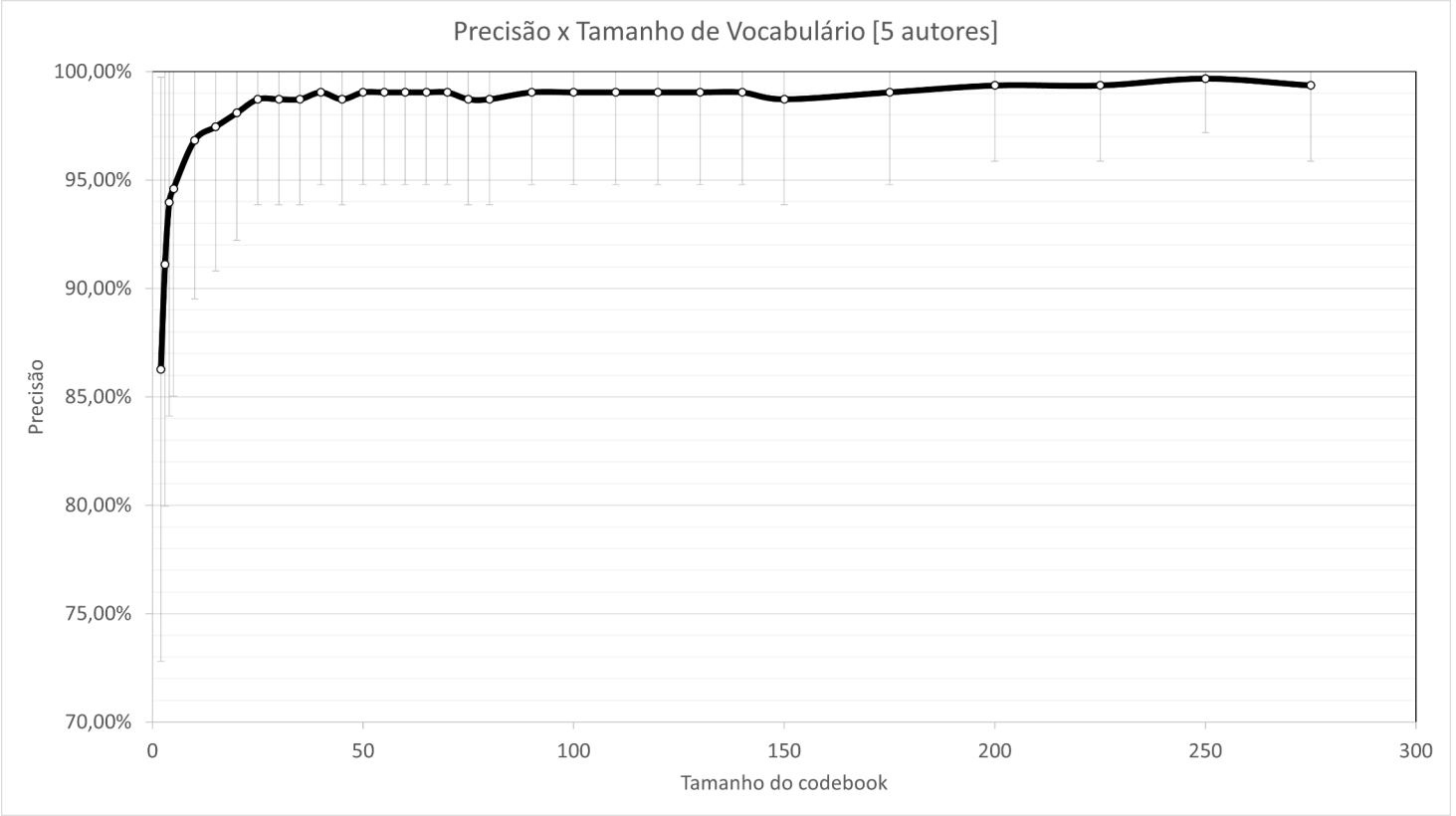
Resultados

MÉTODO FOI AVALIADO NO BANCO DE DADOS BRAZILIAN FORENSIC LETTERS.

VALIDAÇÃO DA METODOLOGIA PROPOSTA EM AMOSTRAS DE 5, 10 E 315 ESCRITORES

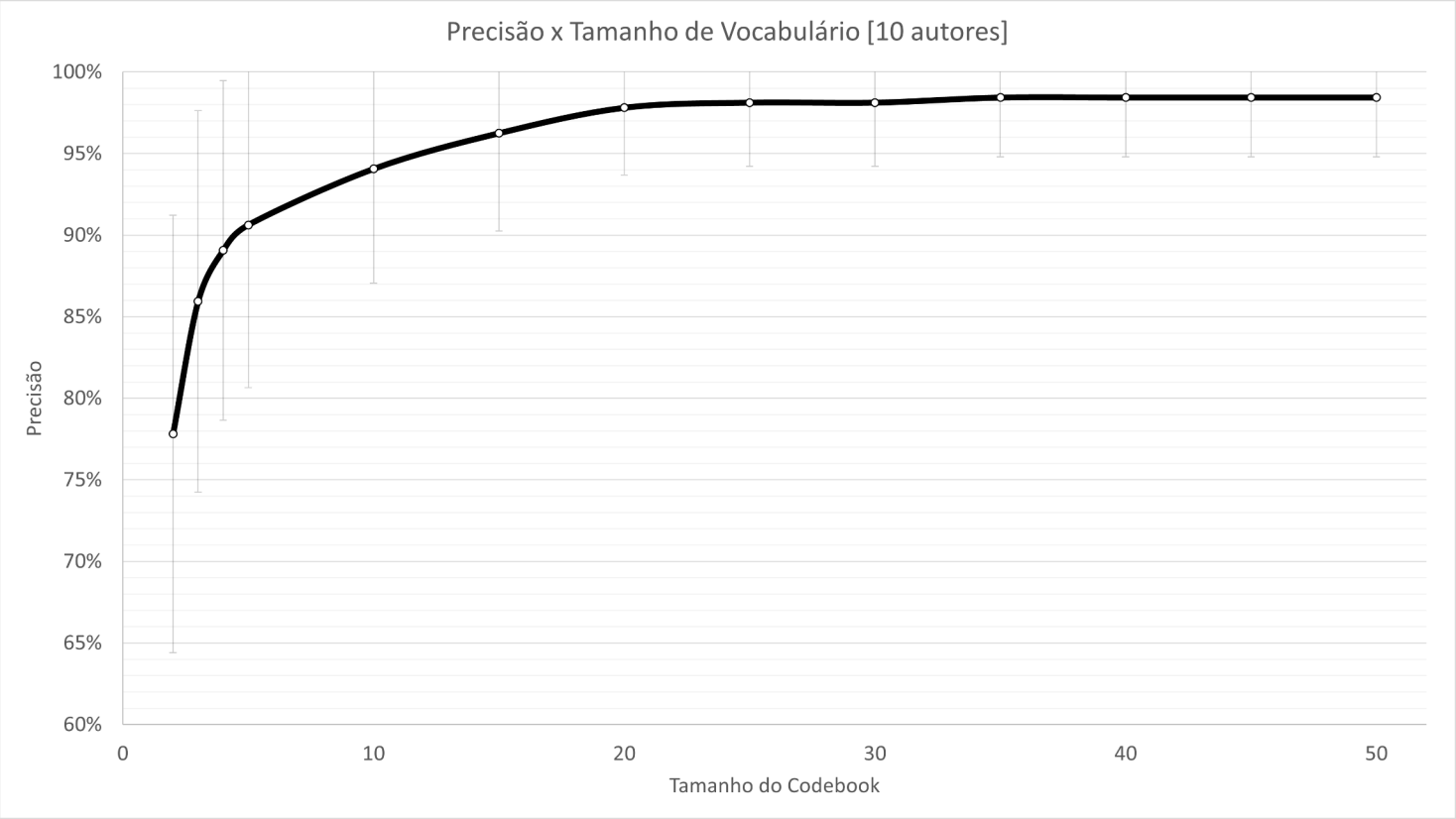
Resultados
População de 5 Autores

- Estabilização da Curva: 25 palavras
- Precisão: 98,73%
- Replicado: 63 vezes
- Pico: 99,68% (Overfitting)
- Validado até 275 palavras



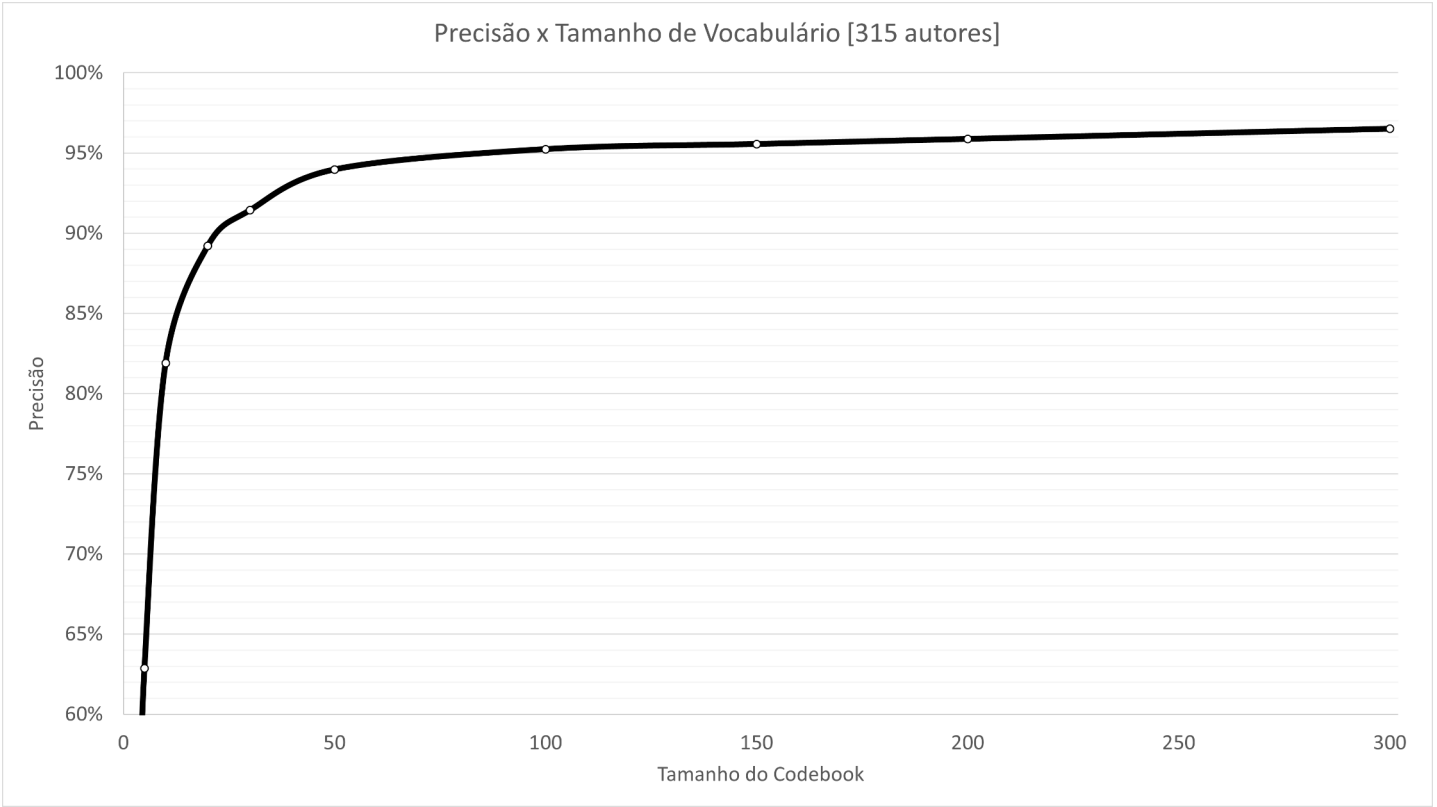
Resultados
População de 10 Autores

- Estabilização da Curva: 25 palavras
- Precisão: 98,12%
- Replicado: 31 vezes
- Validação até 50 palavras



Resultados
População de 315 Autores

- Precisão: 96,51 %
- Ponto de validação: 300 palavras



Resultados
Comparação com literatura no BFL

Autor	Método	Precisão
Kessentini et al.	Polygon	97,78%
Kessentini et al.	Chain Code	93,02%
Bertolini et al.	LPQ + Texture	99,20%
Kessentini et al.	Edge-Hinge + RL	98,41%
Amaral et al.	Graphometry	76%
Modelo proposto	**SIFT + Linear-SVM**	**96,51%**

Conclusões

Resultados:

- Precisão de 96,51% para 315 autores (BFL completo);
- Precisão de 98,73% para amostras menores (5 e 10 autores).

Comparação com a Literatura:

- Menos eficaz que o LPQ + Textura e outros métodos, mas ainda eficiente.

Possibilidades de Aprimoramento:

- Alterar modelo de extração de características;
- Alterar modelo de classificação final;
- Analisar influência do tamanho da amostra e do codebook.

Conclusão:

- Modelo tem potencial para ser uma solução eficaz para identificação de autores em português.

Obrigado!