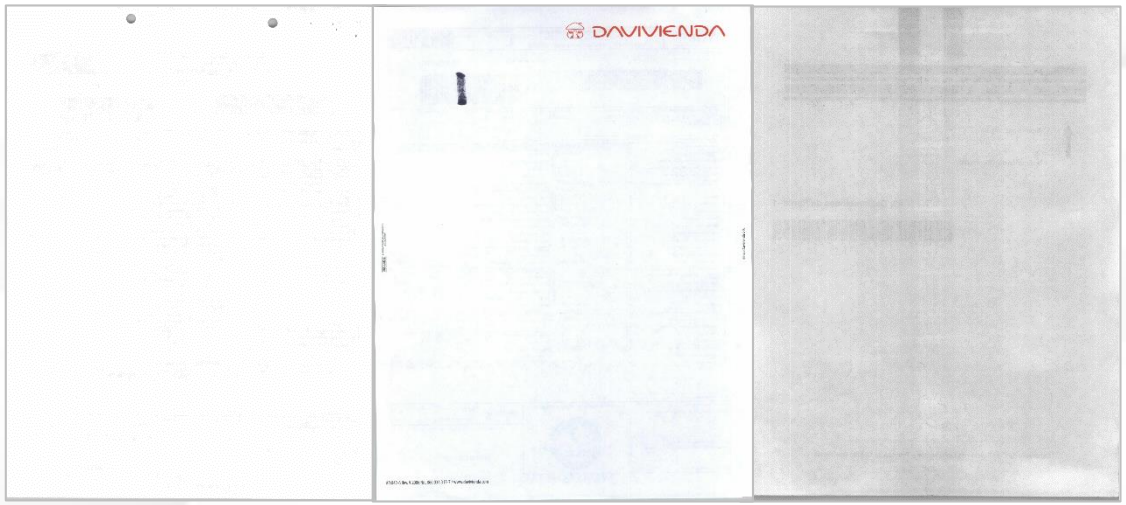
Prueba De Conocimientos
Datos No Estructurados I

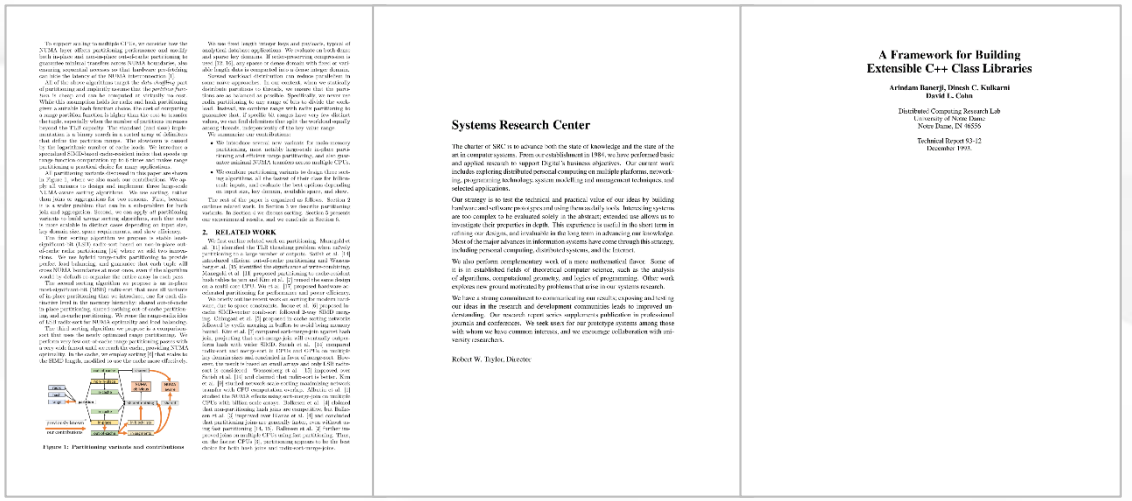
Laura Ruales

Exploración de Datos

Blanco: 147 imágenes

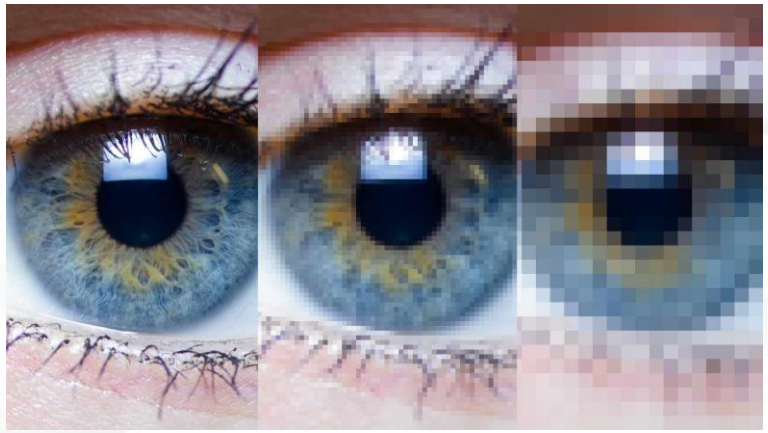


Documento: 100 imágenes



Técnicas Usadas

Entendimiento



Los computadores “ven” las imágenes como un arreglo o una matriz de números que representa un color

Una misma imagen modificada para ser mas grande o más pequeña, será entendida como una imagen distinta para el computador

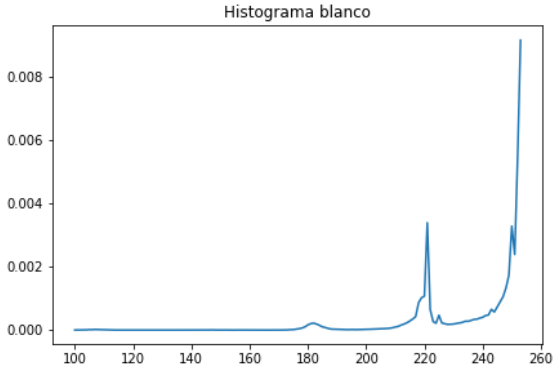
Técnicas Usadas

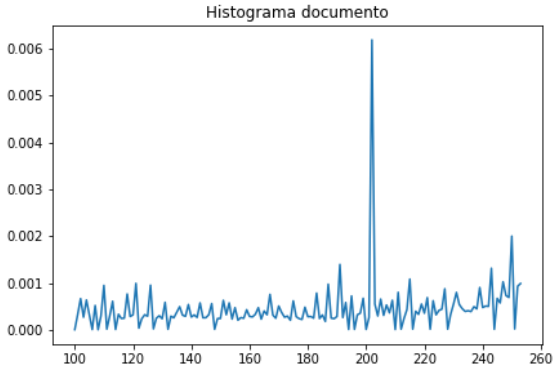
Representación

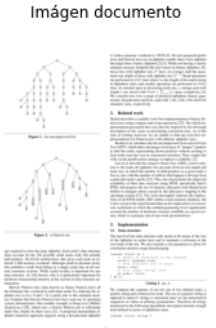
Histograma normalizado: Contar la cantidad de pixeles de la imagen para cada intensidad (valor del pixel) y normalizar (dividir) por el tamaño de la imagen

Todos los histogramas se almacenan dentro de un dataset

Ejemplo histograma e imagen para cada clase

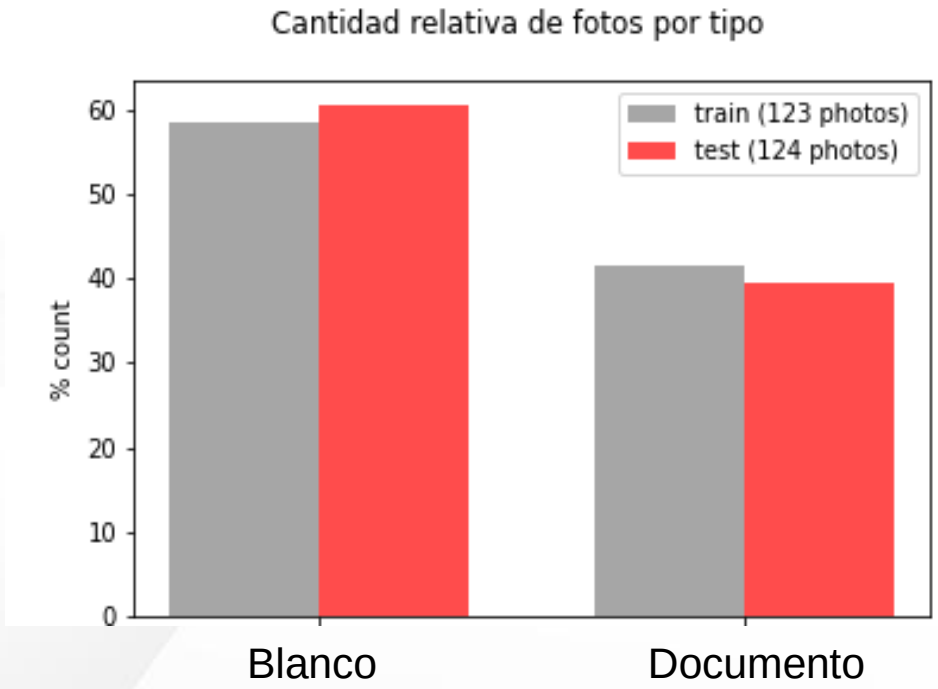






Técnicas Usadas

Clasificación



División del dataset en entrenamiento y prueba.

Creación del modelo

Entrenamiento

Prueba con datos no vistos

Obtención de métricas

Dificultades

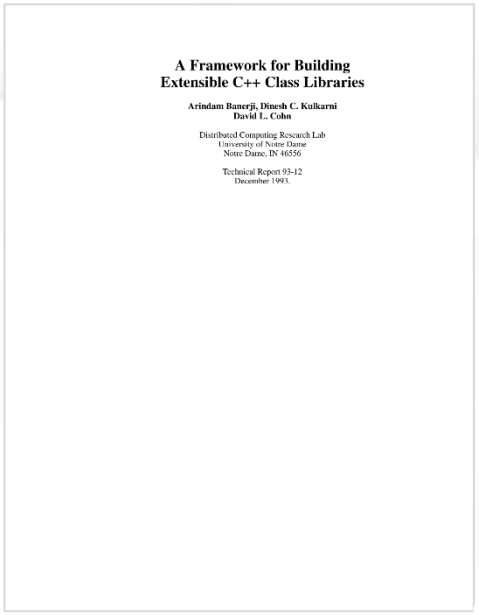
Dificultades Resueltas

- Las imágenes tienen distintos tamaños
- Las imágenes están en distintos formatos (rgb, rgba y grises)

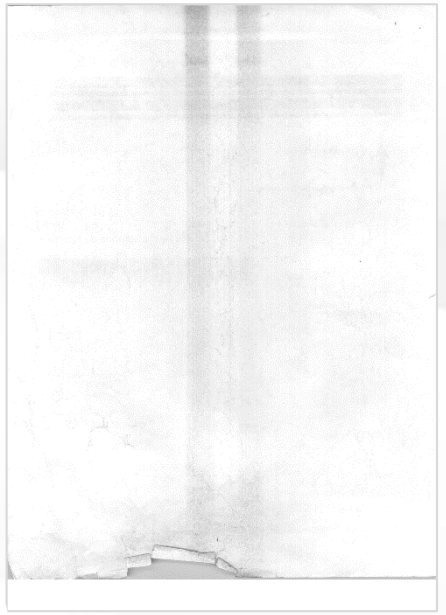
Dificultades Latentes

En algunos casos, las imágenes clasificadas como en blanco tienen bastante ruido. Mientras algunas imágenes clasificadas como documento tienen poco texto

Documento



Blanco

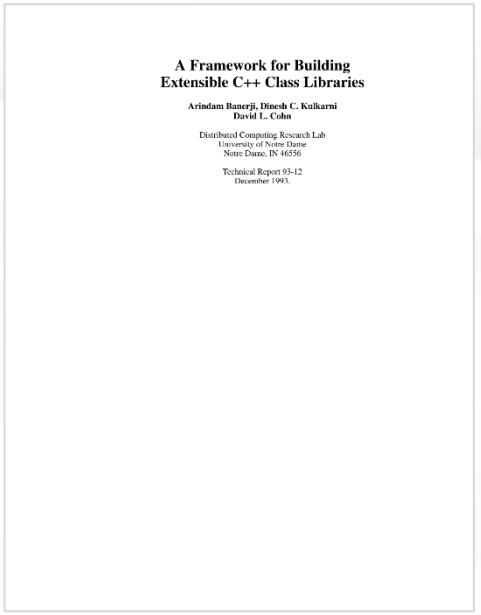


Aspectos importantes

No podemos asumir que si el modelo lo ha hecho bien hasta ahora, lo seguirá haciendo bien en el futuro

A medida que la cantidad de datos aumenten y con ellos **incrementen las variaciones** para cada clase (Documento o Blanco). El modelo actual puede ser insuficiente para **separar patrones que no ha visto antes.**

Documento

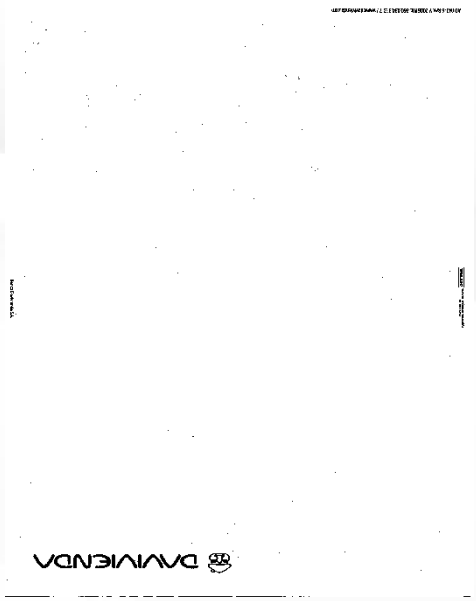

A Framework for Building
Extensible C++ Class Libraries

Arindam Banerji, Dinesh C. Kulkarni
David L. Cohn

Distributed Computing Research Lab
University of Notre Dame
Notre Dame, IN 46556

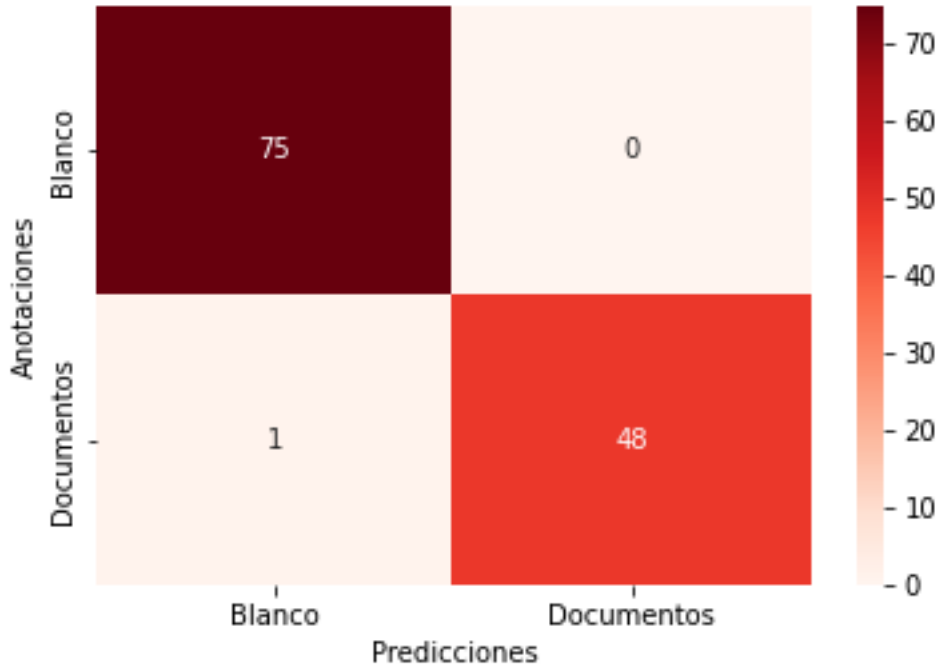
Technical Report 93-12
December 1993.

Blanco



Resultados

Matriz de Confusión.



Las predicciones del modelo son consistentes con las anotaciones de los datos, lo que sugiere que el modelo funciona bien

Resultados Adicionales

Modelo	Exactitud entrenamiento	Exactitud prueba	Precision	Cobertura	F-medida	Tiempo entrenamiento
Naïve Bayes	1,000	0,992	1,000	0,980	0,990	0,003
Bosques Aleatorios	1,000	0,992	1,000	0,980	0,990	0,013
Potenciación del gradiente	1,000	0,984	0,980	0,980	0,980	1,028
Red neuronal	1,000	0,984	0,980	0,980	0,980	2,038
Arboles de decisión	0,967	0,960	0,940	0,959	0,949	0,001
K –vecinos mas cercanos	0,780	0,782	0,645	1,000	0,784	0,004
Linear SVM	0,789	0,774	0,640	0,980	0,774	1,178

Los tiempos de entrenamiento fueron eficientes

* Modelos entrenados con el 50% de los datos y probados con el 50% de los datos.

Modelo	Exactitud entrenamiento	Exactitud prueba	Precision	Cobertura	F-medida	Tiempo entrenamiento
Naive Bayes	1,00	1,00	1,00	1,00	1,00	0,004
Arboles de decisión	1,00	0,96	0,96	0,96	0,96	0,013
Potenciación del gradiente	1,00	0,96	0,96	0,96	0,96	0,999
Bosques aleatorios	1,00	0,96	0,96	0,96	0,96	1,547
K-vecinos mas cercanos	0,99	0,96	0,96	0,96	0,96	0,001
SVM	0,78	0,80	0,70	1,00	0,82	0,006
Red neuronal	0,78	0,80	0,70	1,00	0,82	1,080

Los tiempos de entrenamiento fueron eficientes

* Modelos entrenados con el 80% de los datos y probados con el 20% de los datos.

Resultados Adicionales

Modelo	Exactitud entrenamiento	Exactitud prueba	Precision	Cobertura	F-medida	Tiempo entrenamiento
Naïve Bayes	1,000	0,992	1,000	0,980	0,990	0,003
Bosques Aleatorios	1,000	0,992	1,000	0,980	0,990	0,013
Potenciación del gradiente	1,000	0,984	0,980	0,980	0,980	1,028
Red neuronal	1,000	0,984	0,980	0,980	0,980	2,038
Arboles de decisión	0,967	0,960	0,940	0,959	0,949	0,001
K –vecinos mas cercanos	0,780	0,782	0,645	1,000	0,784	0,004
Linear SVM	0,789	0,774	0,640	0,980	0,774	1,178

Los modelos tienen buenas métricas de desempeño

* Modelos entrenados con el 50% de los datos y probados con el 50% de los datos.

Modelo	Exactitud entrenamiento	Exactitud prueba	Precision	Cobertura	F-medida	Tiempo entrenamiento
Naive Bayes	1,00	1,00	1,00	1,00	1,00	0,004
Arboles de decisión	1,00	0,96	0,96	0,96	0,96	0,013
Potenciación del gradiente	1,00	0,96	0,96	0,96	0,96	0,999
Bosques Aleatorios	1,00	0,96	0,96	0,96	0,96	1,547
K-vecinos mas cercanos	0,99	0,96	0,96	0,96	0,96	0,001
SVM	0,78	0,80	0,70	1,00	0,82	0,006
Red neuronal	0,78	0,80	0,70	1,00	0,82	1,080

Los modelos tienen buenas métricas de desempeño

* Modelos entrenados con el 80% de los datos y probados con el 20% de los datos.


Gracias