Procesamiento Masivo de Web Spam

Washington Bastidas Santos
Jesús González Vera

hadoop

ESCUELA SUPERIOR POLITECNICA DEL LITORAL
ESPOL

amazon web services™

Agenda

- INTRODUCCIÓN
- PROBLEMA
- METODOLOGÍA
- IMPLEMENTACIÓN
- EVALUACIÓN Y RESULTADOS
- CONCLUSIÓN
- TRABAJO FUTURO
- BIBLIOGRAFÍA

1

INTRODUCCIÓN

INTRODUCCIÓN

→ Introducción

Problema

Metodología

Implementación

Evaluación y resultados

Conclusión

Trabajo Futuro

Bibliografía

- Acceso, recuperación y reutilización de la información.
- Máquinas de búsqueda.
- Falencias (PageRank) y servicios gratuitos (Blogspot).
- Incentivo Económico (Google $16,000M en el 2007).

WEBSPAM

“Es simplemente la asignación injustificable de relevancia a una o varias páginas produciendo resultados inesperados en las máquinas de búsqueda "”[1]

---

[1]: Z. Gyongyiand H. Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005

MOTIVACIÓN

Batalla de los buscadores.

Tecnología para procesamiento de grandes cantidades de datos y mejores algoritmos para manejo de información.

Cada vez que se encuentra una solución parcial a un problema, los spammers se encargan de buscar una forma de eludirlo.

Solo los expertos pueden detectar este tipo de problemas.

Suplir a las personas a través de una solución automatizada y que aprenda con el tiempo.

2

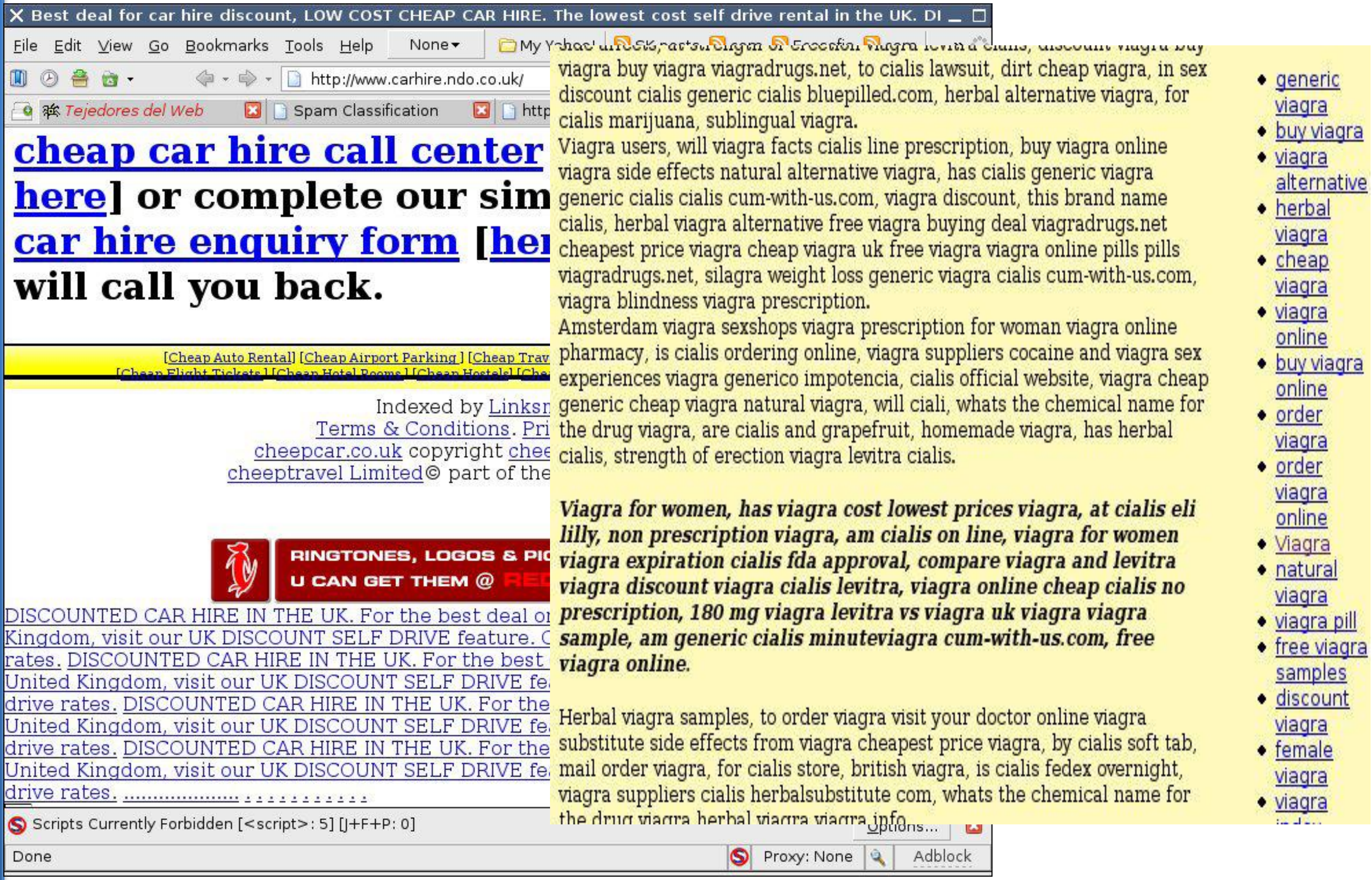
PROBLEMA

POR QUÉ EL WEBSPAM ES MALO?

- **Para el usuario**
  - Difícil satisfacer la información.
  - Experiencia de búsqueda frustrante.
- **Para la máquina de búsqueda**
  - Gasto de ancho de banda, procesamiento CPU, espacio de almacenamiento.
  - Distorsiona el ranking del resultado.

PROBLEMA

- Introducción
- → **Problema**
- Metodología
- Implementación
- Evaluación y resultados
- Conclusión
- Trabajo Futuro
- Bibliografía



cheap car hire call center
here] or complete our sim
car hire enquiry form [her
will call you back.

[Cheap Auto Rental] [Cheap Airport Parking] [Cheap Trav
[Cheap Flight Tickets] [Cheap Hotel Rooms] [Cheap Hostels] [Che

Indexed by Linksn
Terms & Conditions. Pri
cheepcar.co.uk copyright chee
cheeptravel Limited© part of the

RINGTONES, LOGOS & PIC
U CAN GET THEM @ RED

DISCOUNTED CAR HIRE IN THE UK. For the best deal on
Kingdom, visit our UK DISCOUNT SELF DRIVE feature. C
rates. DISCOUNTED CAR HIRE IN THE UK. For the best
United Kingdom, visit our UK DISCOUNT SELF DRIVE fea
drive rates. DISCOUNTED CAR HIRE IN THE UK. For the
United Kingdom, visit our UK DISCOUNT SELF DRIVE fea
drive rates. DISCOUNTED CAR HIRE IN THE UK. For the
United Kingdom, visit our UK DISCOUNT SELF DRIVE fea
drive rates. ....

Scripts Currently Forbidden [<script>: 5] [J+F+P: 0]

Done

viagra buy viagra viagradrugs.net, to cialis lawsuit, dirt cheap viagra, in sex discount cialis generic cialis bluepilled.com, herbal alternative viagra, for cialis marijuana, sublingual viagra.
Viagra users, will viagra facts cialis line prescription, buy viagra online viagra side effects natural alternative viagra, has cialis generic viagra generic cialis cialis cum-with-us.com, viagra discount, this brand name cialis, herbal viagra alternative free viagra buying deal viagradrugs.net cheapest price viagra cheap viagra uk free viagra viagra online pills pills viagradrugs.net, silagra weight loss generic viagra cialis cum-with-us.com, viagra blindness viagra prescription.
Amsterdam viagra sexshops viagra prescription for woman viagra online pharmacy, is cialis ordering online, viagra suppliers cocaine and viagra sex experiences viagra generico impotencia, cialis official website, viagra cheap generic cheap viagra natural viagra, will ciali, whats the chemical name for the drug viagra, are cialis and grapefruit, homemade viagra, has herbal cialis, strength of erection viagra levitra cialis.

Viagra for women, has viagra cost lowest prices viagra, at cialis eli lilly, non prescription viagra, am cialis on line, viagra for women viagra expiration cialis fda approval, compare viagra and levitra viagra discount viagra cialis levitra, viagra online cheap cialis no prescription, 180 mg viagra levitra vs viagra uk viagra viagra sample, am generic cialis minuteviagra cum-with-us.com, free viagra online.

Herbal viagra samples, to order viagra visit your doctor online viagra substitute side effects from viagra cheapest price viagra, by cialis soft tab, mail order viagra, for cialis store, british viagra, is cialis fedex overnight, viagra suppliers cialis herbalsubstitute com, whats the chemical name for the drug viagra herbal viagra viagra info

- generic viagra
- buy viagra
- viagra alternative
- herbal viagra
- cheap viagra
- viagra online
- buy viagra online
- order viagra
- order viagra online
- Viagra
- natural viagra
- viagra pill
- free viagra samples
- discount viagra
- female viagra
- viagra

PALABRAS CLAVES + LINKS

PROBLEMA

- Creación de páginas para que otras tenga un mejor ranking.
- Produce "optimización de buscadores":
  - Gran mayoría de tráfico generado por buscadores.
  - Usuarios solo observan las 3 primeras paginas de búsqueda.
- Dos tipos de Web spam:
  - Spam basado en contenido.
  - Spam basado en links.

SPAM BASADO EN CONTENIDO

Introducción

→ Problema

Metodología

Implementación

Evaluación y resultados

Conclusión

Trabajo Futuro

Bibliografía

- Keywords Repetidos.
- Palabras como: “googel”, “accomodation” , “trabel”.
- Análisis estadístico.
- Éxito por no filtros de spam en las consultas más populares y mejor pagadas.

Keyword Stuffing

Results 1 - 10 of about 2,320,000 for sample lay out of a medical certificate. (0.33 seconds)

Did you mean: sample **layout** of a medical certificate

Free Wordpress Theme: WP **Medical** Doctor | Wordpress Themed
Live Demo: Features: 3-column wordpress theme **medical** wordpress. ... **medical certificate** template, doctors **certificate layout**, free premium wp themes ...
www.wordpressthemed.com/free-wordpress-theme-wp-**medical**-doctor/ - 14k -
Cached - Similar pages - Note this

Sponsored Links

Certificate Samples
Huge selection of **Certificate** items
Yahoo.com

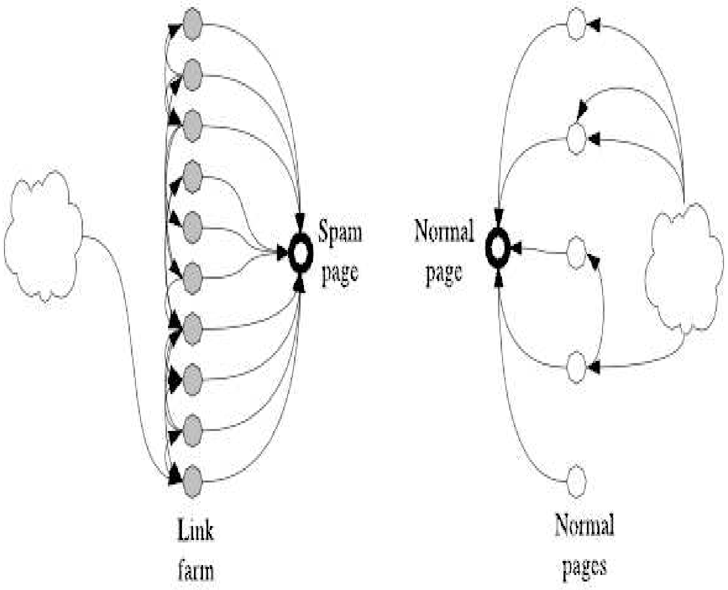
- medicaldoctor
- free blogger medicine templates
- wood word template of wordpress
- medicine dosing schedule template
- wp themes doctor
- free girls ppc themes
- medical css based tamplate free
- templet thema
- free templates on doctors and health
- doctor templates free downloads
- sample lay out of a medical certificate
- free downloads medical templates
- free medical template for wordpress
- free medical Themes for Wordpress
- WordPress docter
- torrent wordpress template
- medecine- ebooks blogspot
- free medical photos
- how to chane template to rtl in wordpress
- wordpress templates pharmacy free

SPAM BASADO EN LINK

Introducción
→ Problema
Metodología
Implementación
Evaluación y resultados
Conclusión
Trabajo Futuro
Bibliografía

- Google y su algoritmo PageRank basado en links.
- Otros buscadores siguieron el modelo.
- El modelo de cómo trabaja es conocido por los spammers.
- Ejemplo granja de enlaces.

Granja de Enlaces

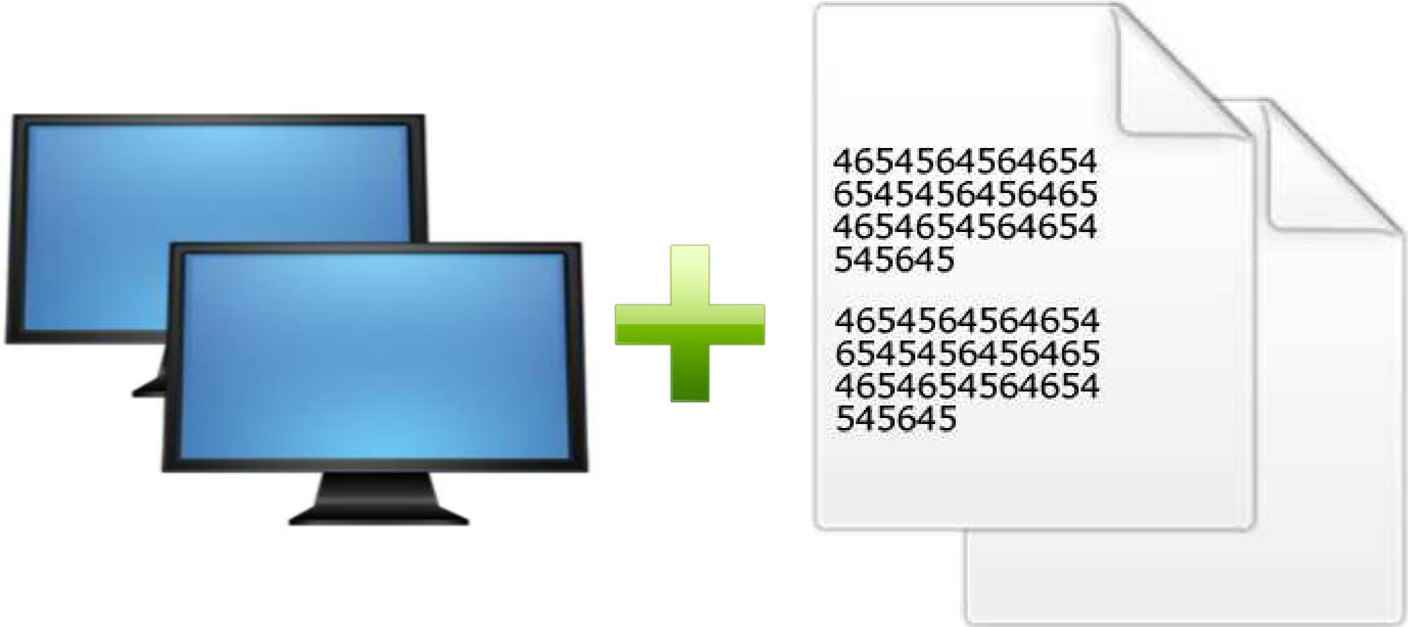


3

METODOLOGÍA

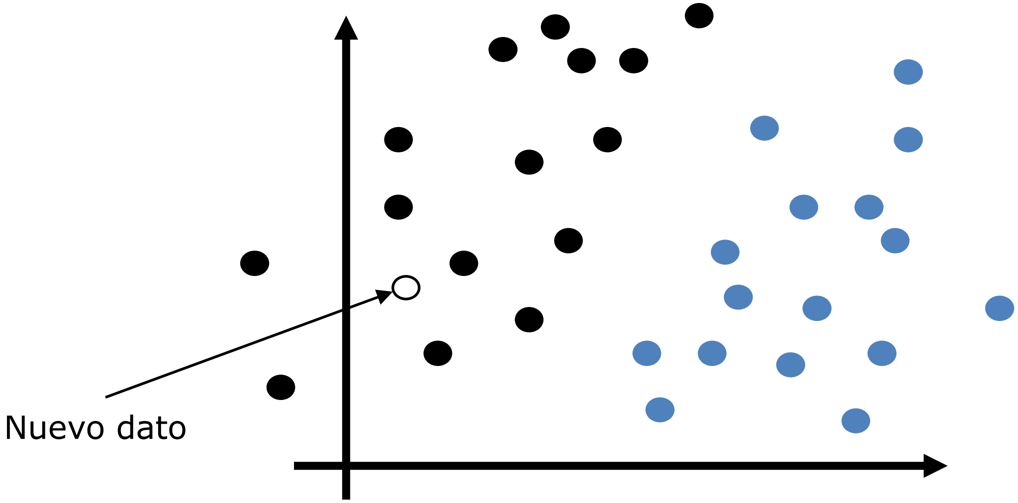
APRENDIZAJE AUTOMÁTICO

- **El *aprendizaje o entrenamiento* es el mejoramiento en base a la experiencia de alguna tarea.**
  - Algoritmos Supervisado: Función correspondencia



SUPERVISADO Y CLASIFICACIÓN

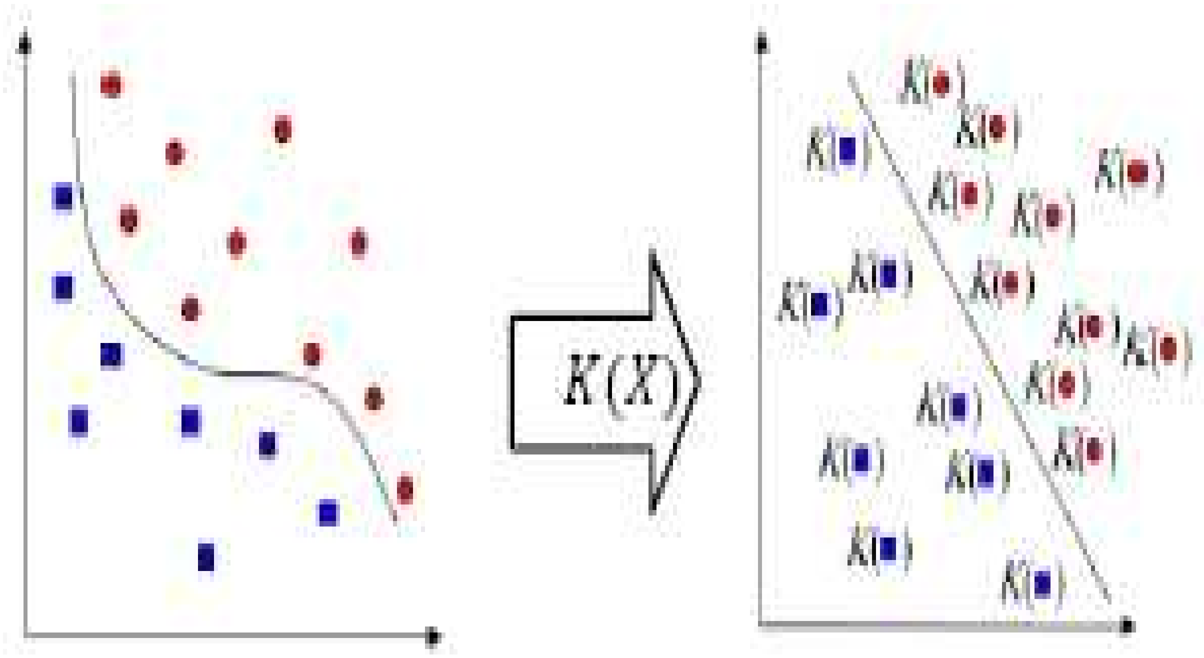
- Algoritmos que razonan a partir de ejemplos y producen hipótesis.
- Un tipo de aprendizaje supervisado es ***Clasificación:***
  - *Construir **modelo** para predecir la clase de un nuevo dato*



MÁQUINA DE VECTORES DE APOYO

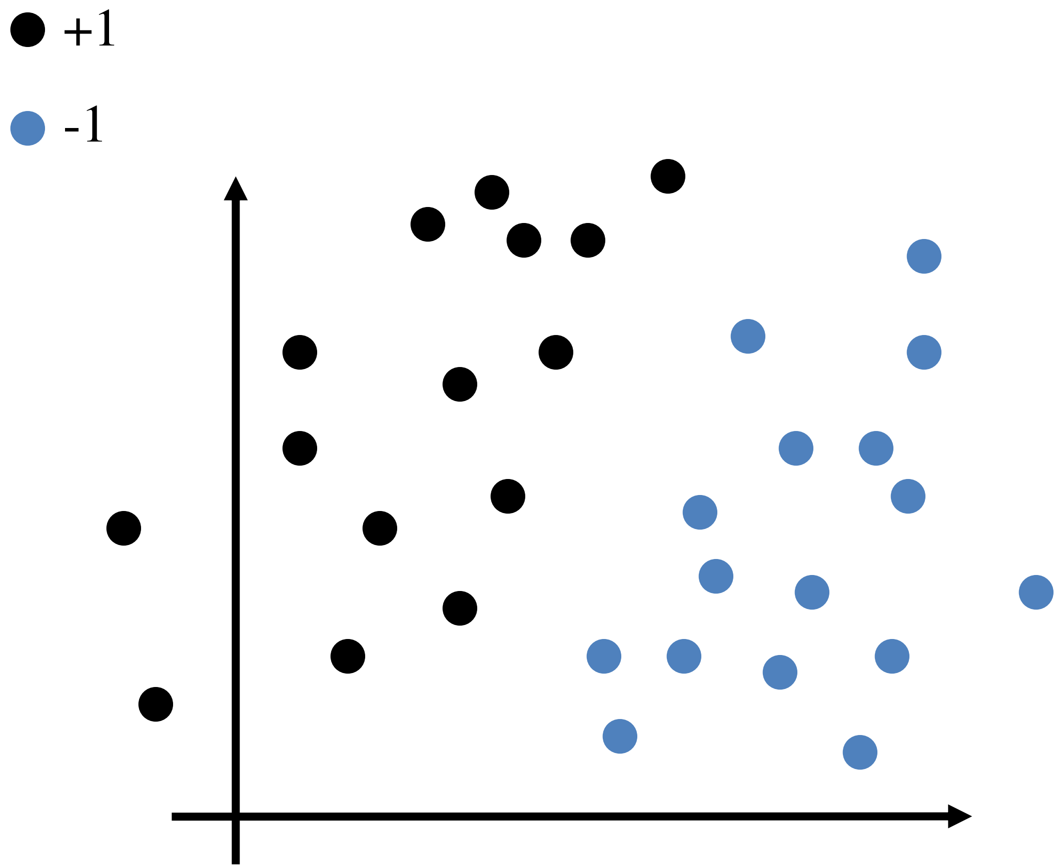
Introducción
Problema
→ Metodología
Implementación
Evaluación y resultados
Conclusión
Trabajo Futuro
Bibliografía

- Máquinas de vectores de apoyo (SVM, siglas inglés)
- Desarrolladas por Vapnik están basadas en la teoría de aprendizaje estadístico.
- Utilizan funciones Kernel para datos dispersos:



MÁQUINA DE VECTORES DE APOYO

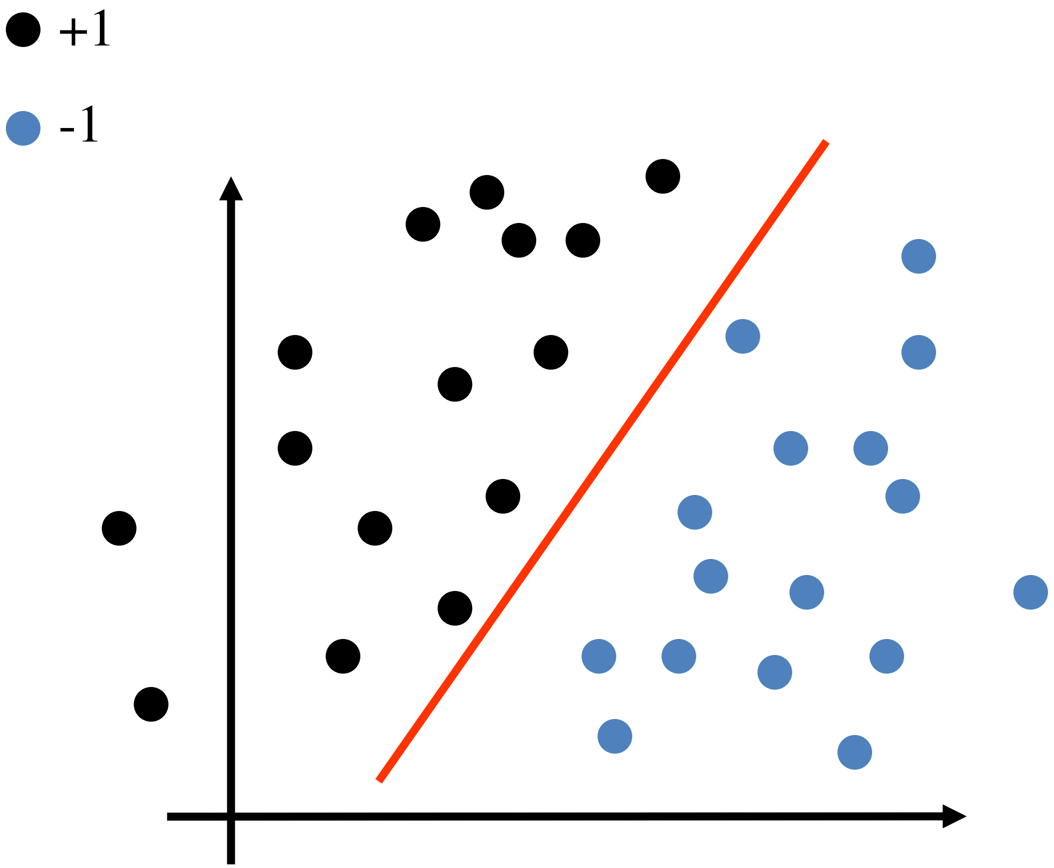
¿cómo clasificar estos datos?



MÁQUINA DE VECTORES DE APOYO

- Introducción
- Problema
- → **Metodología**
- Implementación
- Evaluación y resultados
- Conclusión
- Trabajo Futuro
- Bibliografía

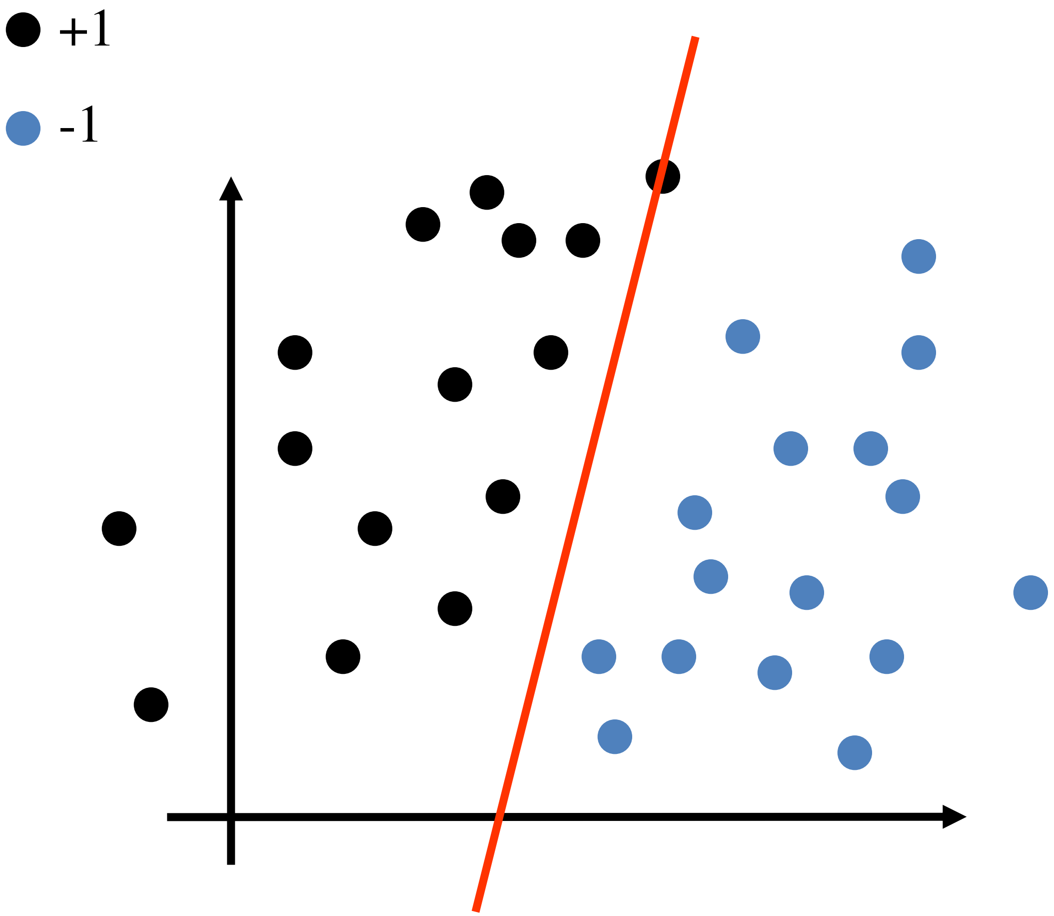
¿cómo clasificar estos datos?



MÁQUINA DE VECTORES DE APOYO

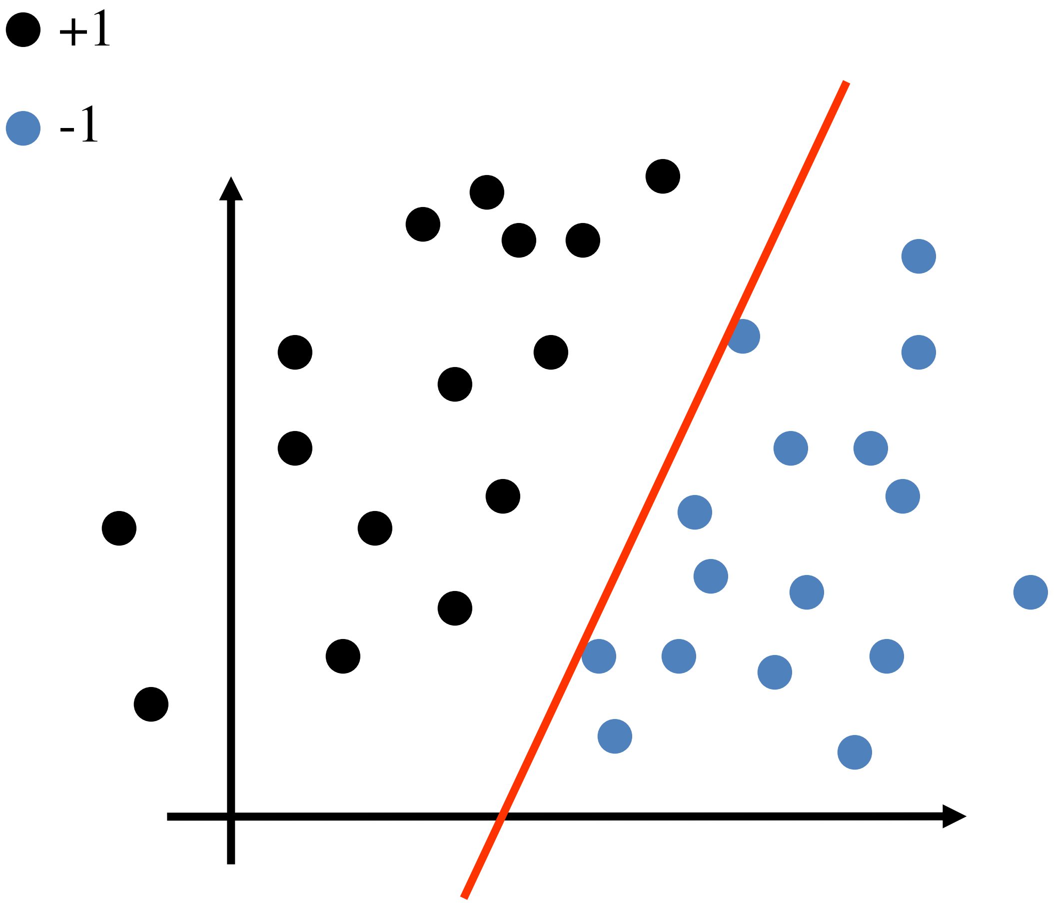
- Introducción
- Problema
- → **Metodología**
- Implementación
- Evaluación y resultados
- Conclusión
- Trabajo Futuro
- Bibliografía

¿cómo clasificar estos datos?



MÁQUINA DE VECTORES DE APOYO

¿cómo clasificar estos datos?



MÁQUINA DE VECTORES DE APOYO

Introducción

Problema

→ **Metodología**

Implementación

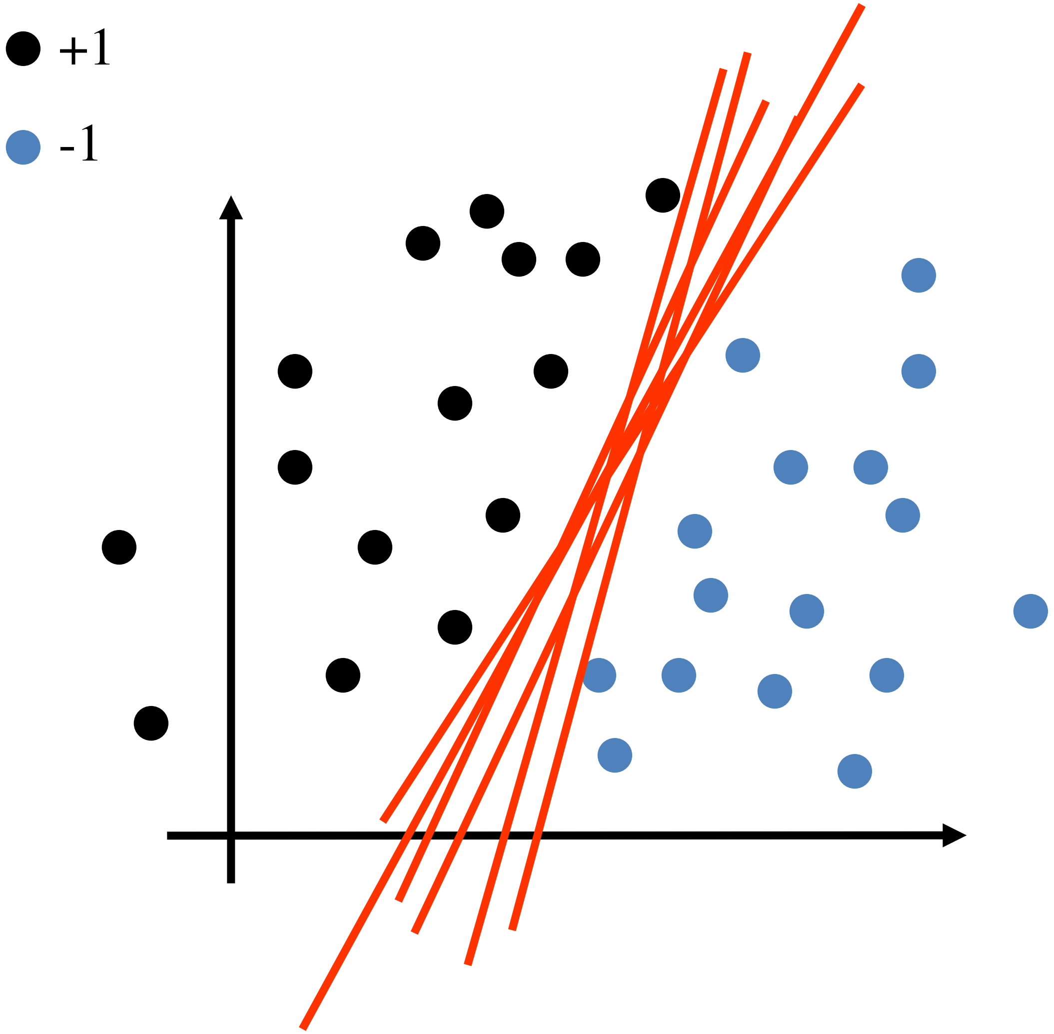
Evaluación y resultados

Conclusión

Trabajo Futuro

Bibliografía

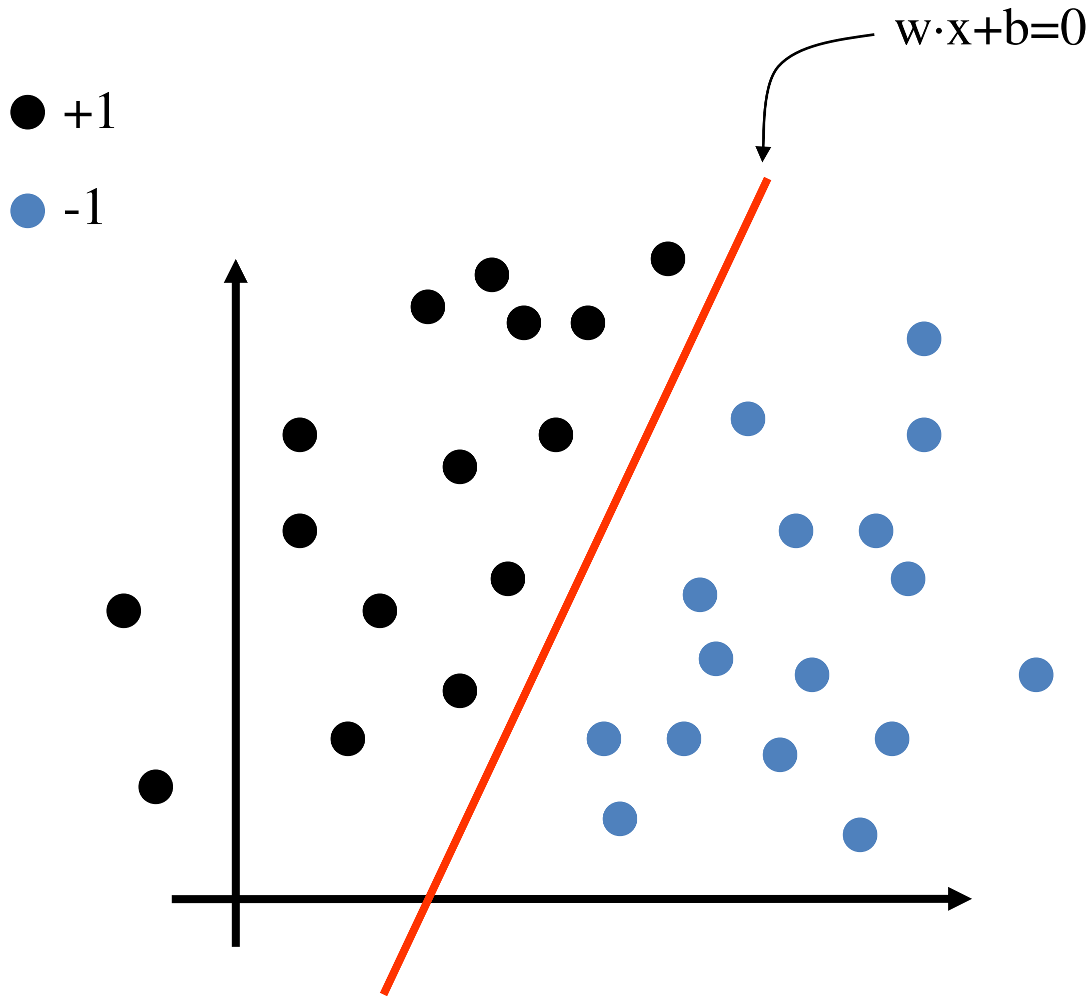
¿cómo clasificar estos datos?



MÁQUINA DE VECTORES DE APOYO

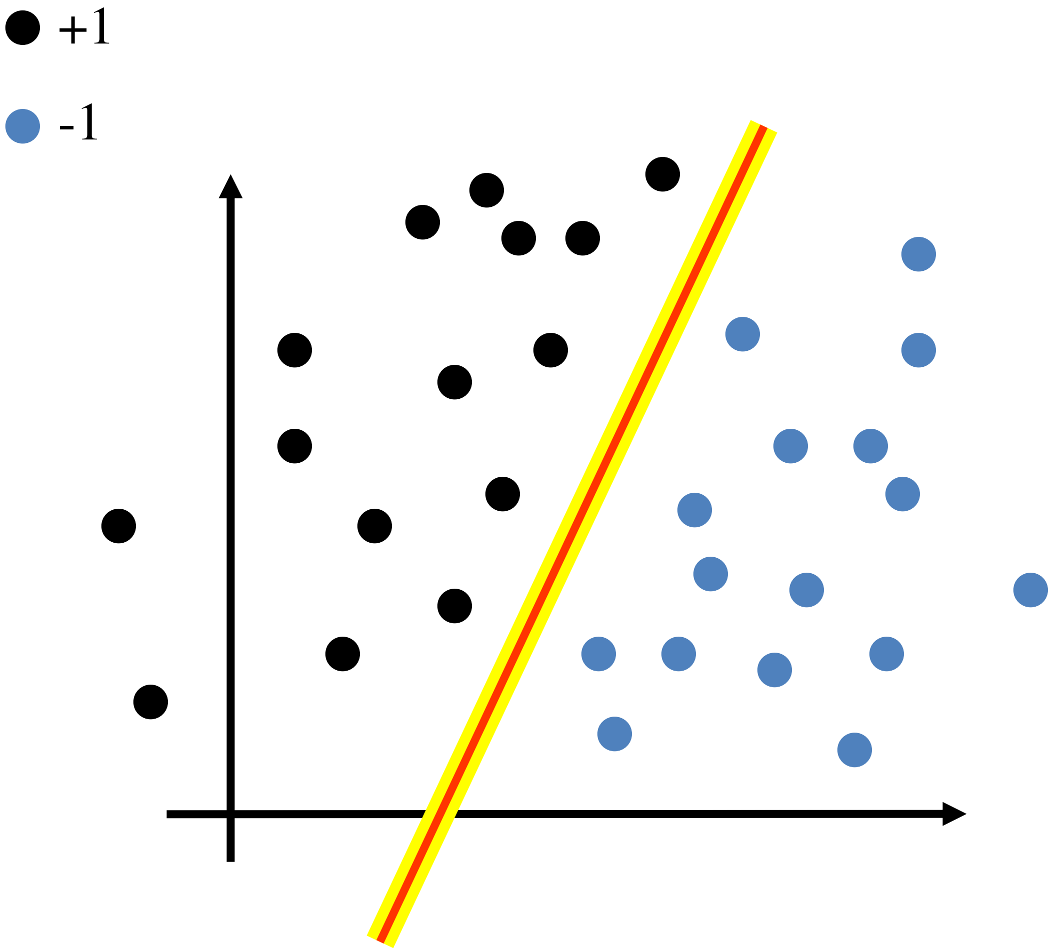
Introducción
Problema
→ Metodología
Implementación
Evaluación y resultados
Conclusión
Trabajo Futuro
Bibliografía

Definimos el hiperplano



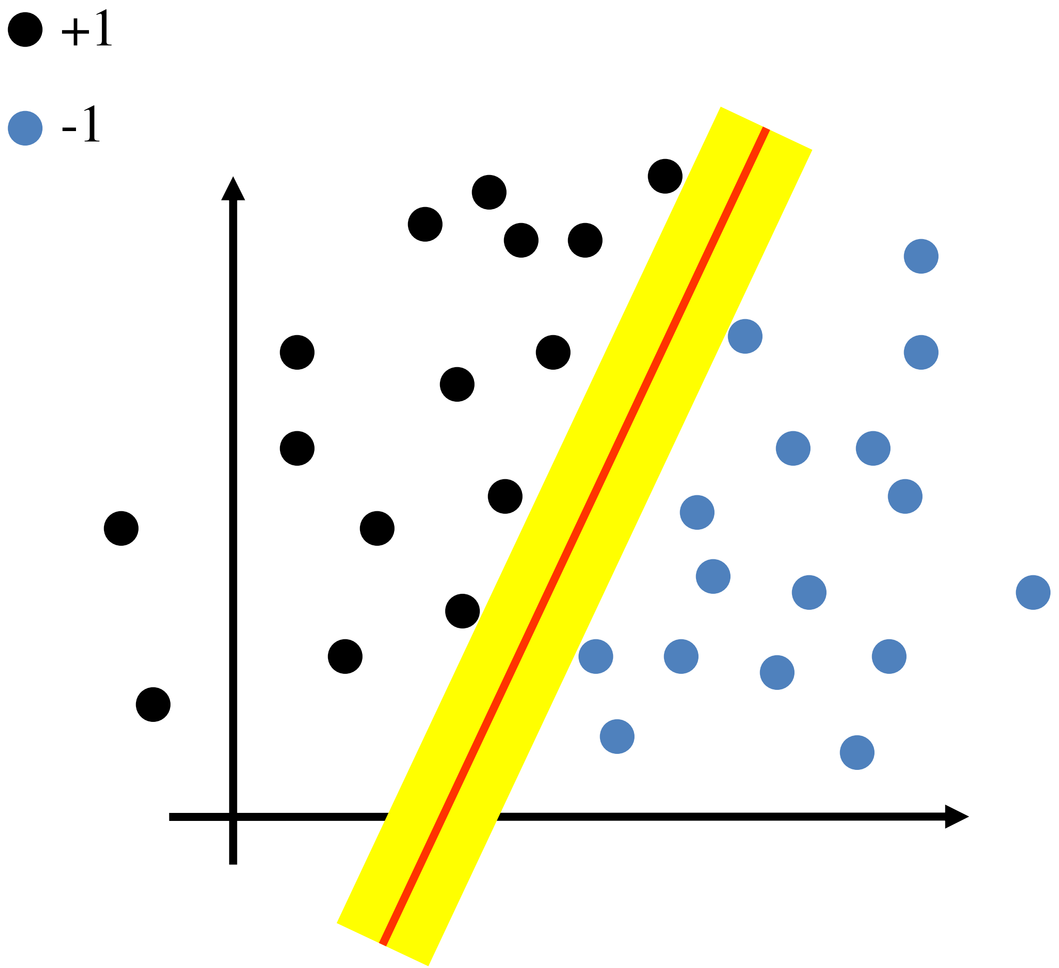
MÁQUINA DE VECTORES DE APOYO

Definimos el margen



MÁQUINA DE VECTORES DE APOYO

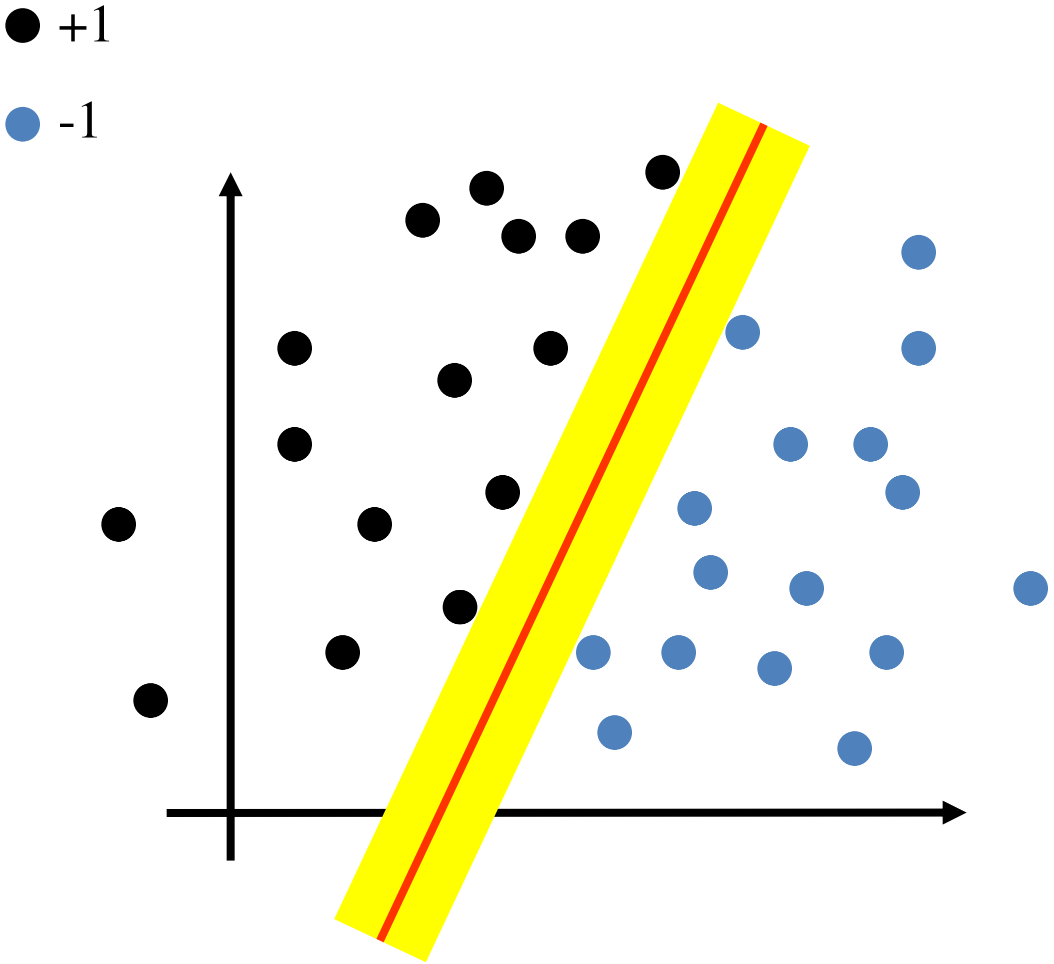
La idea es maximizar el margen.



MÁQUINA DE VECTORES DE APOYO

- Introducción
- Problema
- → **Metodología**
- Implementación
- Evaluación y resultados
- Conclusión
- Trabajo Futuro
- Bibliografía

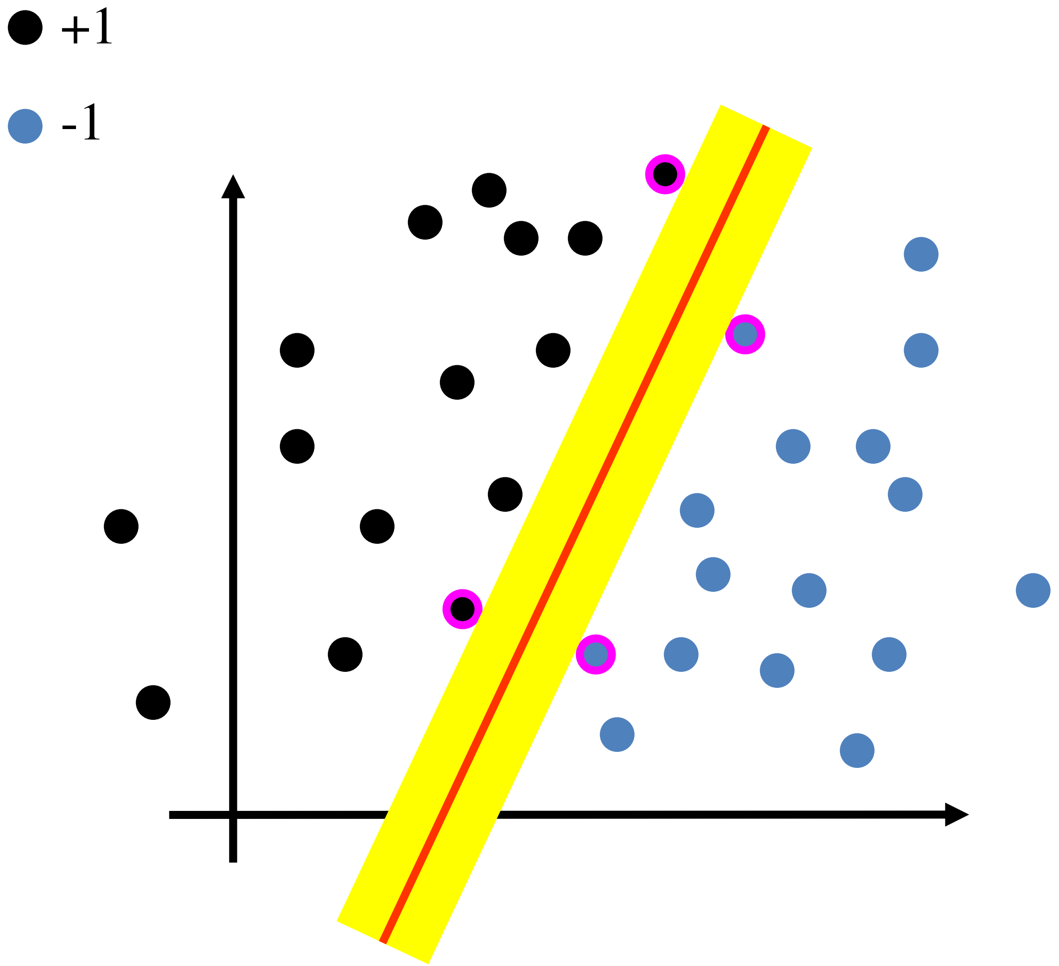
El hiperplano que tenga el mayor margen es el mejor clasificador de los datos.



MÁQUINA DE VECTORES DE APOYO

- Introducción
- Problema
- → **Metodología**
- Implementación
- Evaluación y resultados
- Conclusión
- Trabajo Futuro
- Bibliografía

Los vectores de apoyo son los puntos que tocan el límite del margen.



MÁQUINA DE VECTORES DE APOYO

Introducción

Problema

→ Metodología

Implementación

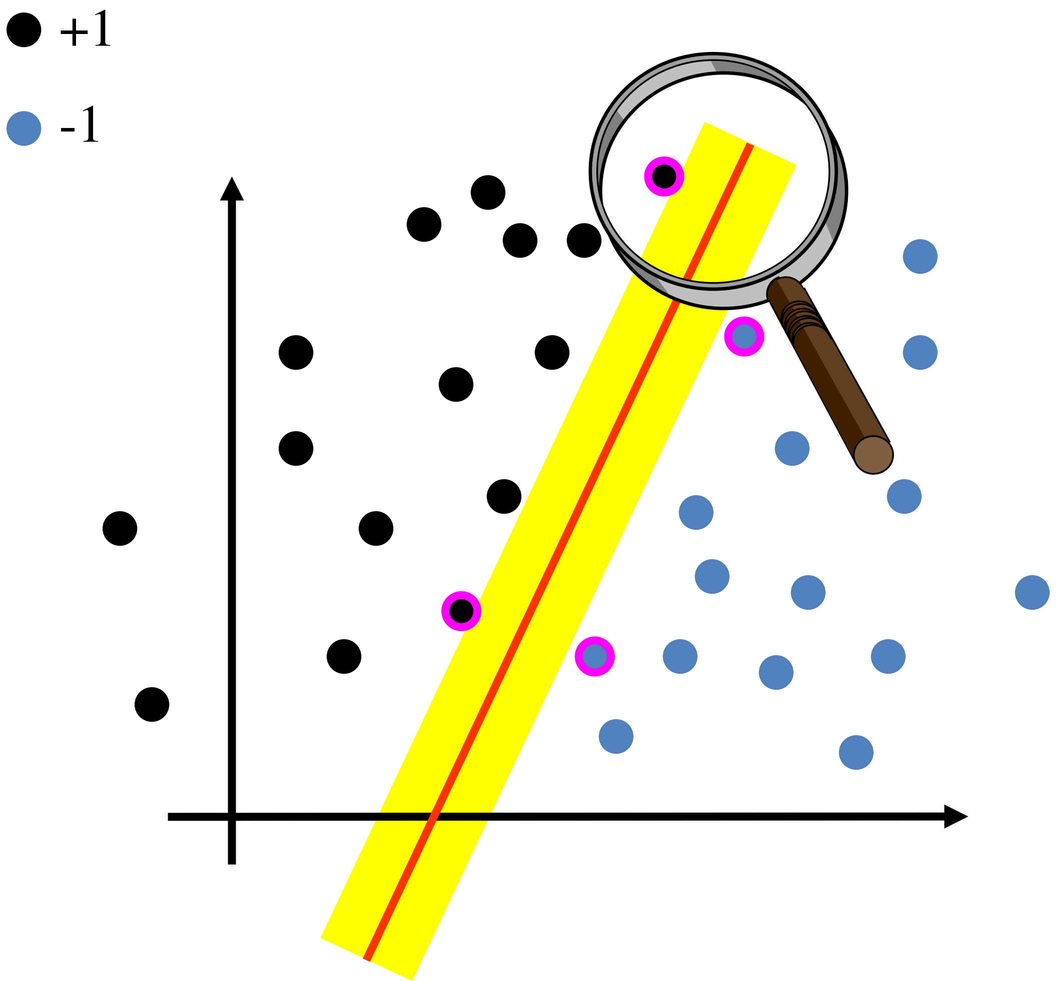
Evaluación y resultados

Conclusión

Trabajo Futuro

Bibliografía

Veamos los hiperplanos “positivo” y “negativo”



MÁQUINA DE VECTORES DE APOYO

Introducción

Problema

→ Metodología

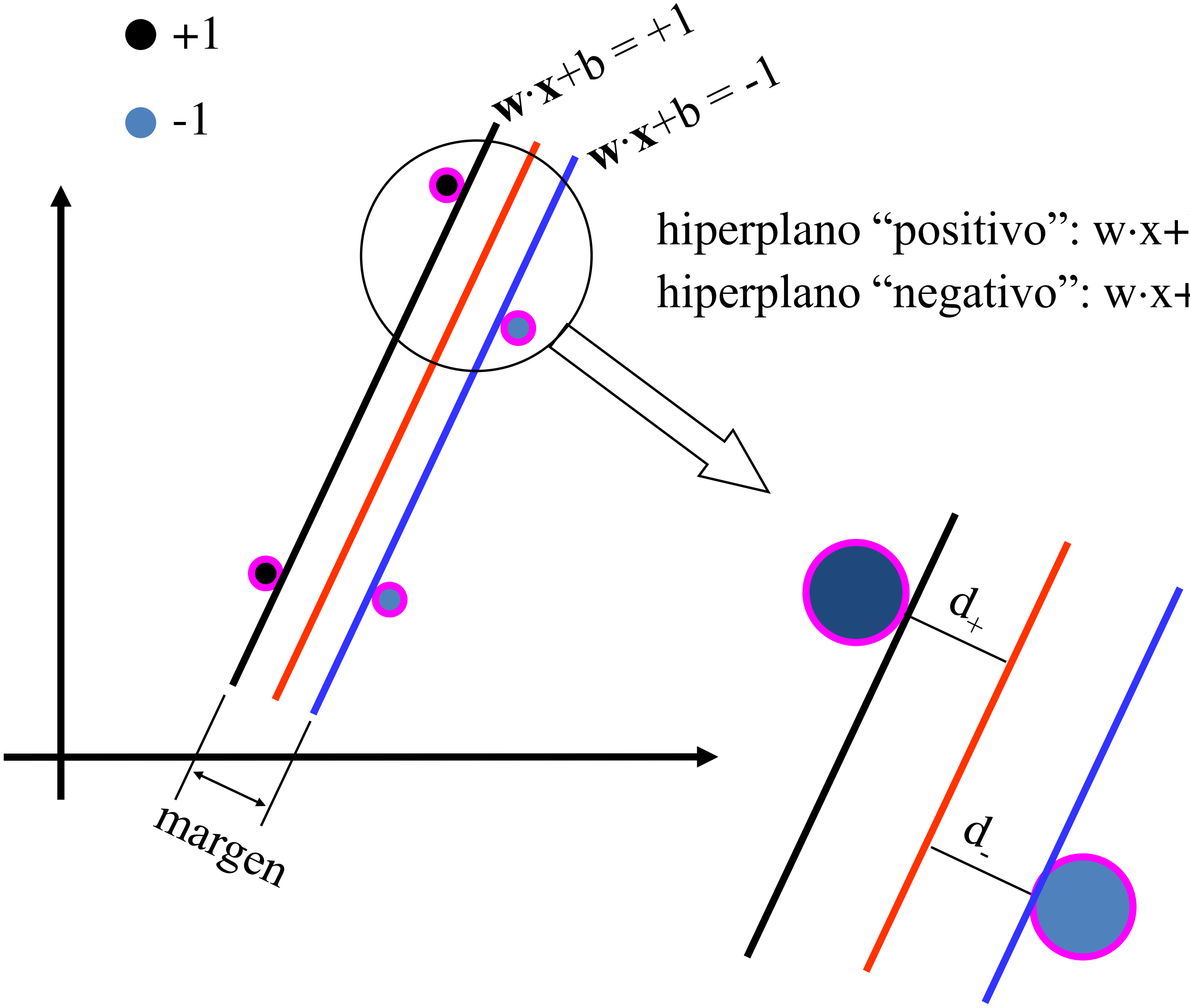
Implementación

Evaluación y resultados

Conclusión

Trabajo Futuro

Bibliografía



hiperplano “positivo”: $w \cdot x+b = +1$

hiperplano “negativo”: $w \cdot x+b = -1$

MÁQUINA DE VECTORES DE APOYO

Al final lo que resulta es una función de correspondencia para la clasificación

$$F(x) = sgn(\sum_{i=1}^{l} y_i \alpha_i (k(x_i,x) + b)).$$

α_i	Son los valores de son los multiplicadores de LaGrange de la ecuación
$k(x_i,x)$	Es la función Kernel utilizada y ***b*** la variable independiente.
	Los vectores de apoyo están implícitos en la función Kernel

SVM EN CASCADA

Introducción

Problema

→ **Metodología**

Implementación

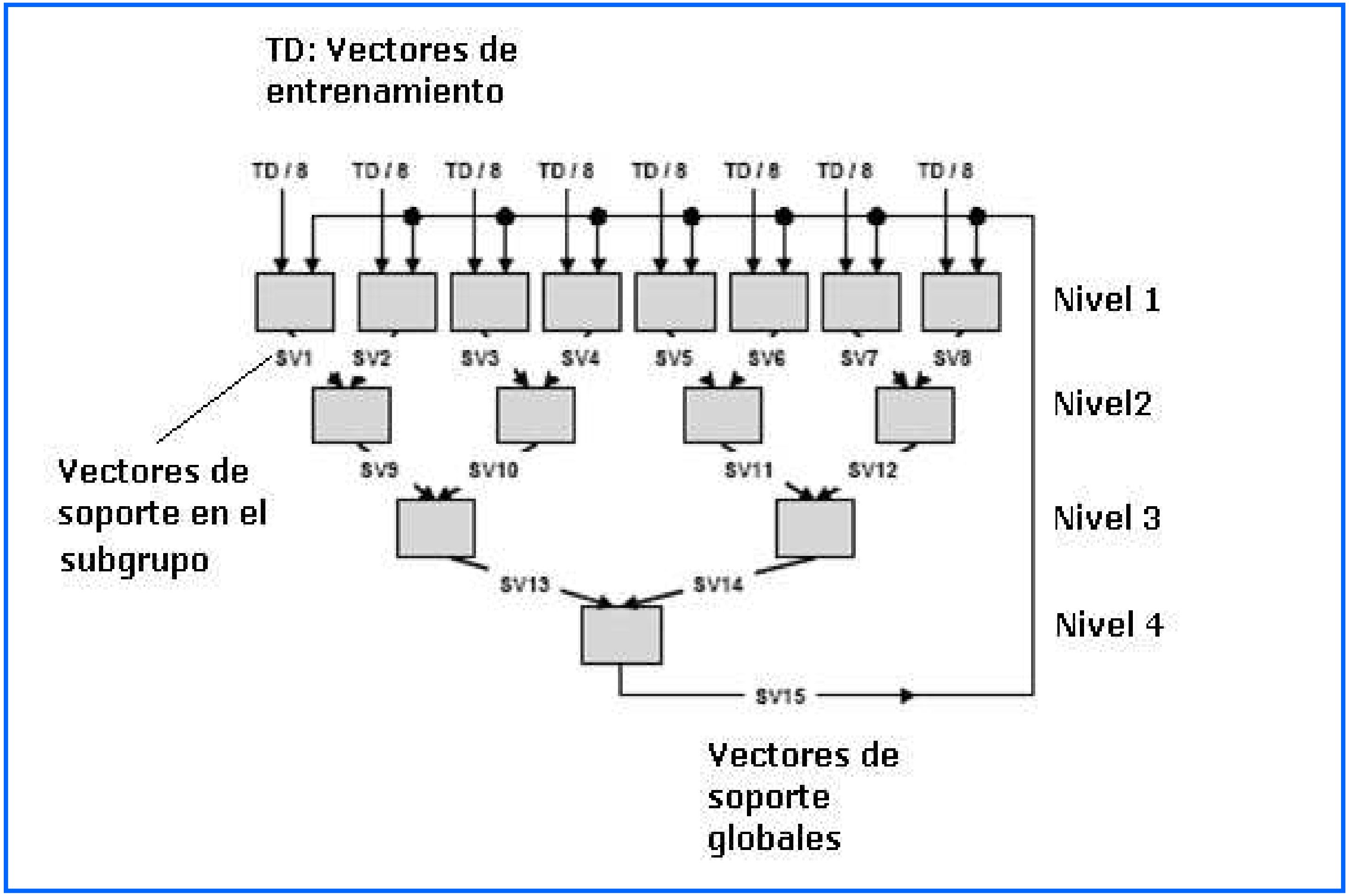
Evaluación y resultados

Conclusión

Trabajo Futuro

Bibliografía

- Consumo recursos demasiado elevado de SVM
- Alternativa de paralelización: ***SVM en cascada***



VECTORES DE CARACTERÍSTICAS

- Introducción
- Problema
- → **Metodología**
- Implementación
- Evaluación y resultados
- Conclusión
- Trabajo Futuro
- Bibliografía

- Vectores constituidos por datos numéricos.
- Expresiones regulares.

```
Regex _TagParser = new Regex("<([a-zA-Z]\w*?)>")
```

- Datos características + etiqueta (Spam o No Spam).
- Características Seleccionadas:
  - *Número de palabras en la página.*
  - *Número de palabras en el título.*
  - *Promedio de palabras*
  - *Fracción del texto anclado.*
  - *Porcentaje de texto oculto.*

EL UNIVERSO | Noticias del Ecuador y del mundo - Mozilla Firefox

Ingresar | Registrarse

Ediciones Anteriores | Clasificados | Servicios

EL UNIVERSO

Sábado 19 de septiembre del 2009
Guayaquil, Ecuador
Hora Local 07:25

Edición Impresa

Noticias | Opinión | Guayaquil | Deportes | Entretenimiento | Vida | Fotogalerías

Política | Economía | País | Migración | Internacional | Agropecuarios | The New York Times

Aguajes ponen en riesgo a 42 casas en Jambelí

La erosión deja prácticamente al archipiélago sin playa. Hay zonas declaradas en riesgo.

Conatel dio 'luz verde' al proceso de juzgamiento contra Teleamazonas

Resolvió ayer disponer a la Superintendencia de Telecomunicaciones que inicie un proceso de juzgamiento para determinar si cometió o no infracciones.

Sin invitados de EE.UU., FAE asumió control total de base

Militares de EE.UU. se reunieron ayer en un hotel, mientras que la FAE hizo su ceremonia.

Con resguardo policial asumen dos rectores

Las autoridades a las que se encargaron los rectorados de los colegios Vicente Rocafuerte y Aguirre Abad fueron posesionadas bajo la vigilancia de la fuerza pública.

Capacitate

Revisa la sexta clase

Otros Deportes

Nicolás empató la serie con Brasil

- Sin Nadal, España gana en la Davis
- Federer pone adelante a los suizos

Ranking

Más Leídas

- Policía confisca aretes de Maradona
- Llop: 'Clásico no es decisivo'
- Aguajes ponen en riesgo a 42 casas en Jambelí
- Con 29 y 8 puñaladas mataron a madre e hija

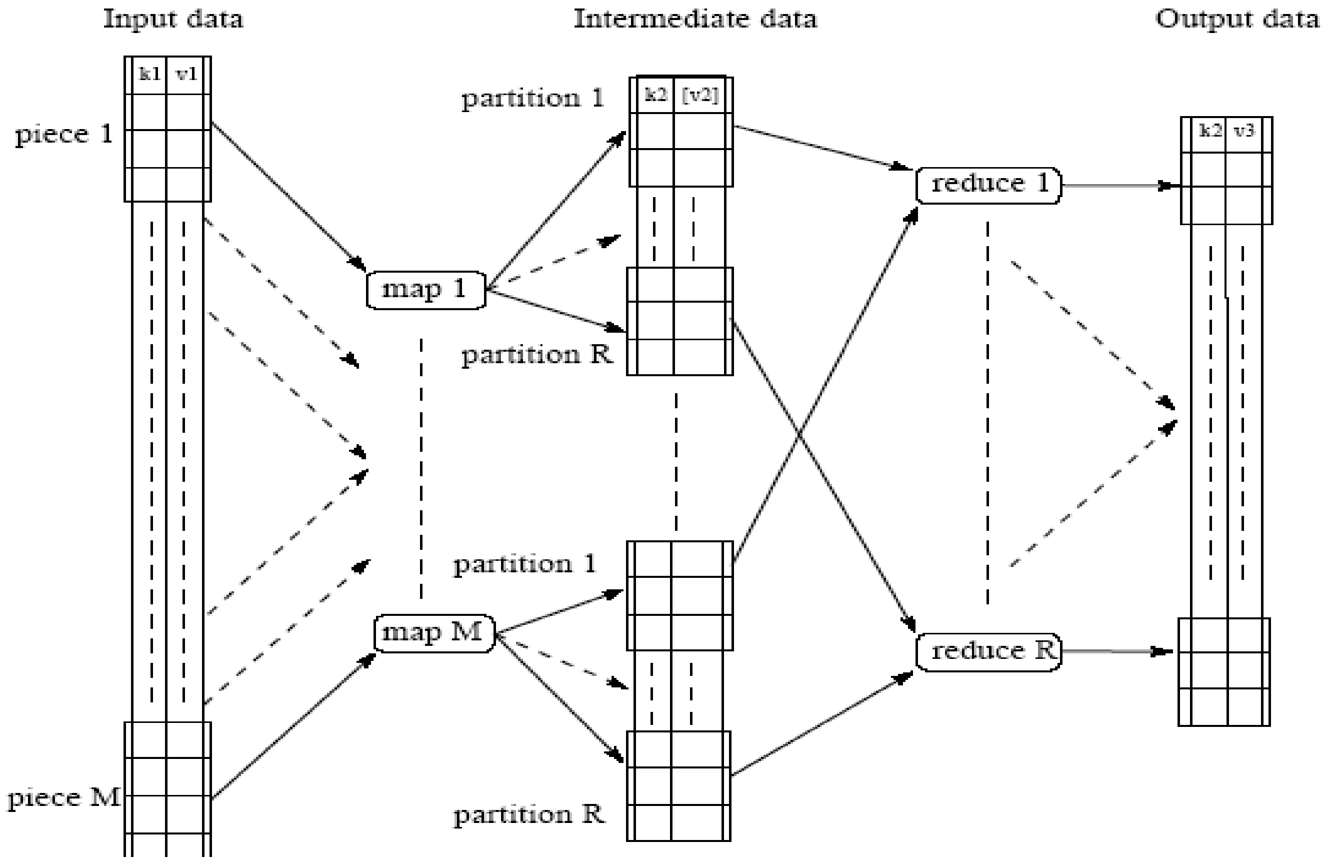
```
<a href="/deportes/" title="Deportes">
Deportes </a>
```

Texto ancla

```
<input type="hidden" value="Internet">
```

MAPREDUCE

- Introducción
- Problema
- → **Metodología**
- Implementación
- Evaluación y resultados
- Conclusión
- Trabajo Futuro
- Bibliografía



Map: Toma las duplas de entrada y produce una dupla clave/valor intermedia.

Reduce: Acepta una clave intermedia y un set de valores para la clave.

Introducción
Problema
→ Metodología
Implementación
Evaluación y resultados
Conclusión
Trabajo Futuro
Bibliografía

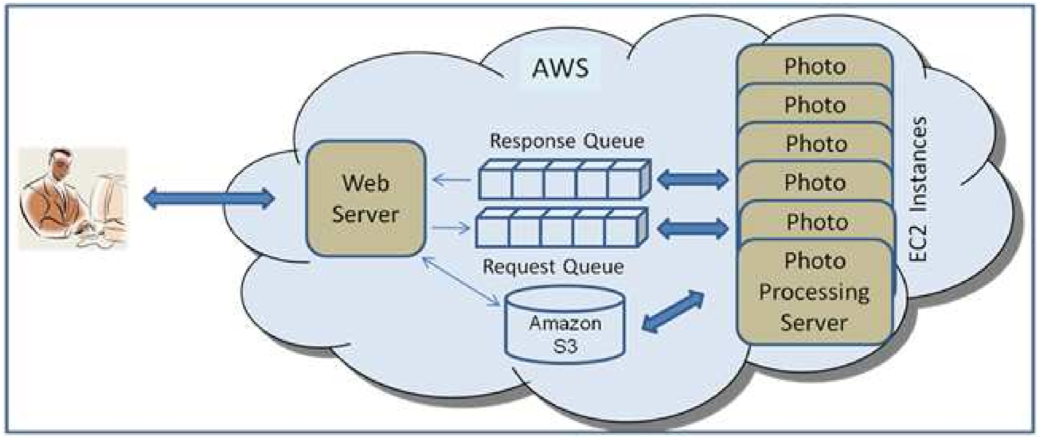
AMAZON WEB SERVICE

S3

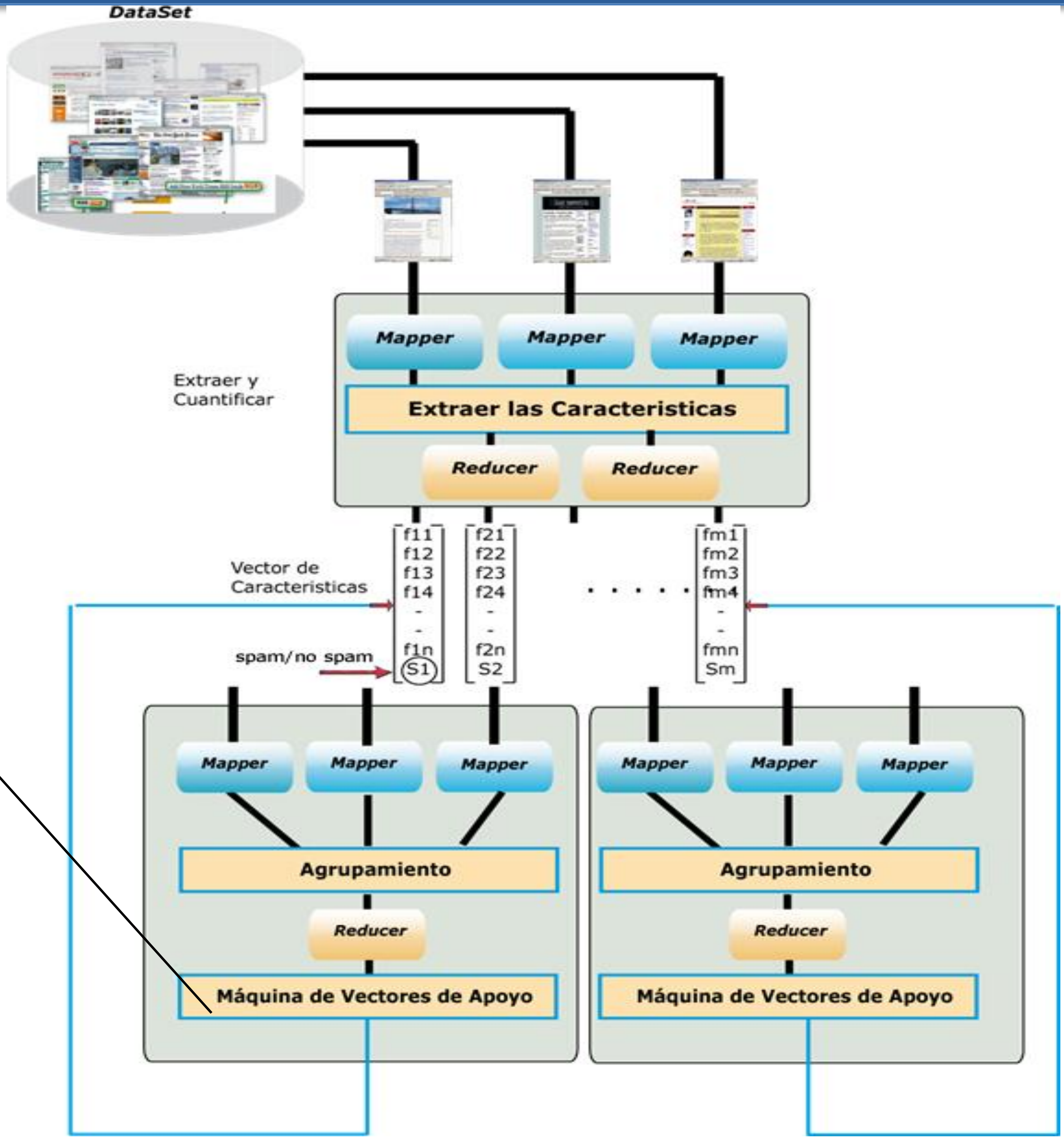
- Almacenar y recuperar cualquier cantidad de información.
- Computación web escalable más fácil para los desarrolladores

EC2

- Es un servicio web que provee capacidad computacional reajustable
- Provee un completo control de tus recursos computacionales



MODELO GENERAL



Kernel
RBF

4

IMPLEMENTACIÓN

DATASET

- WEBSPAM-UK2006 llenado con páginas de dominio ".UK" en el 2006.
- software UbiCrowler.
- Para el presente trabajo se utilizó un total de aproximadamente 120,000 páginas.
- Tiene un tamaño aproximado de 1.7 GB.
- Sirve al subprocesos de *extracción*:

DATASET

Introducción
Problema
Metodología
→ Implementación
Evaluación y resultados
Conclusión
Trabajo Futuro
Bibliografía

- Los vectores de característica resultantes del subproceso de extracción, representados como un archivo de texto:

Dirección_página_web1 **\t** *f11 ; f12 ; f13 ; f14 ; f15 ; etiqueta1*
Dirección_página_web2 **\t** *f21 ; f22 ; f23 ; f24 ; f25 ; etiqueta2*

Dirección_página_web: *P*ágina Web
\t*:* Carácter de tabulación
fn1 ; fn2 ; fn3 ; fn4 ; fn5*:* Características de una página n cualquiera.
Etiqueta: ***-1*** spam y ***1*** no spam

LIBRERÍAS

Introducción

Problema

Metodología

→ Implementación

Evaluación y resultados

Conclusión

Trabajo Futuro

Bibliografía

Paradigma MapReduce	***Hadoop 1.8***
Distribución Linux	***Cloudera***
Software de virtualización	***VMware Player***
Gestión instancias de Amazon EC2	***Ec2 Api tools***
SVM	***LibSVM***
Gestionar Amazon S3	Pluggin ***S3 Organizer***
Gestionar EC2	Pluggin ***ElasticFox***

5

EVALUACIÓN Y RESULTADOS

ANÁLISIS DEL KERNEL

Introducción

Problema

Metodología

Implementación

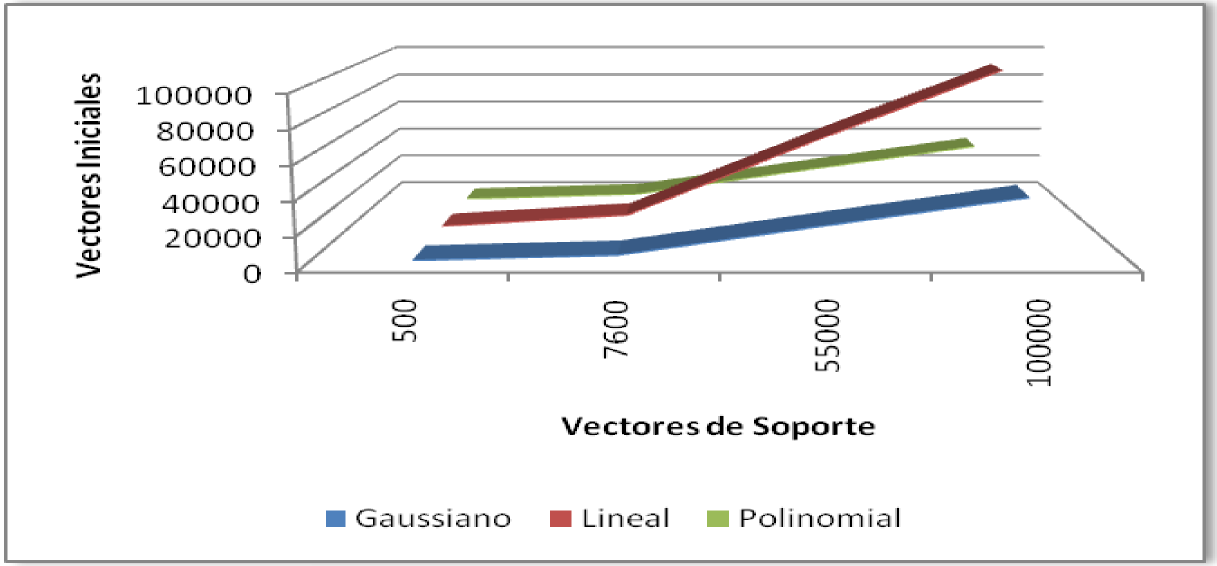
→ Evaluación y resultados

Conclusión

Trabajo Futuro

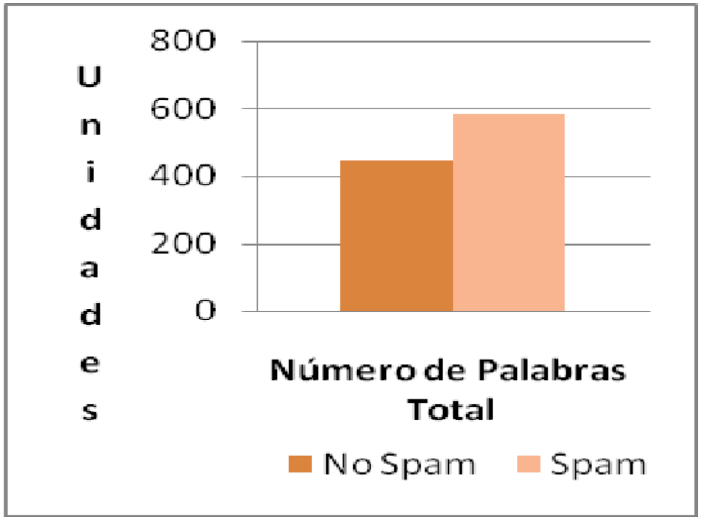
Bibliografía

Vectores Iniciales	**Vectores de Soporte**		
	Kernel Gaussiano	**Kernel Lineal**	**Kernel Polinomial**
500	156	500	156
7600	3300	7600	3300
55000	20370	55000	20370
100000	37037	100000	37037

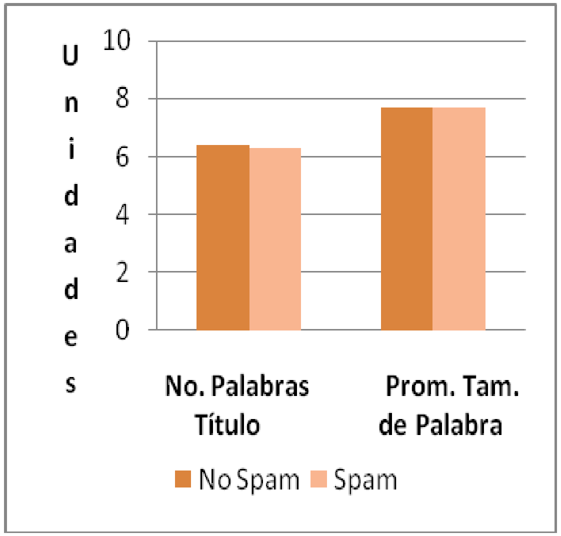


ANÁLISIS DE LAS CARACTERÍSTICAS

Tipo	Numero De Palabras
No Spam	444,79
Spam	583,3



	Características	
Tipo	Numero Palabras Título	Promedio Tamaño Palabra
No Spam	6,39	7,69
Spam	6,27	7,67



ANÁLISIS DE LAS CARACTERÍSTICAS

Introducción

Problema

Metodología

Implementación

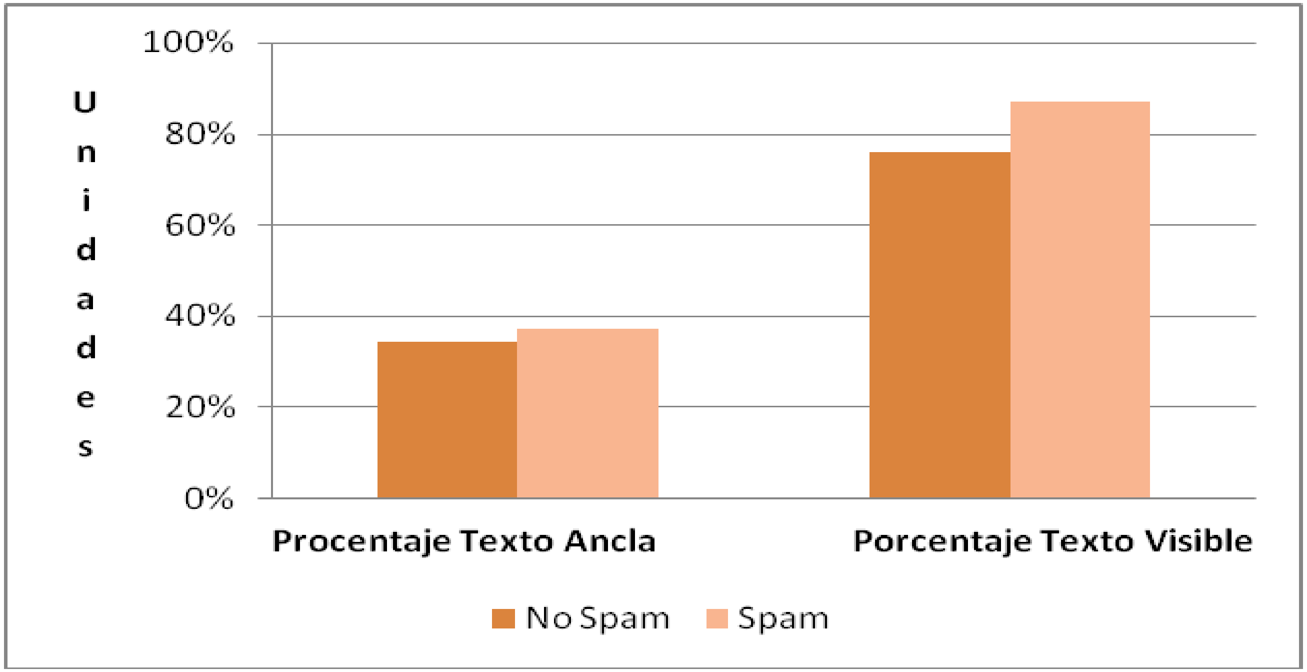
→ Evaluación y resultados

Conclusión

Trabajo Futuro

Bibliografía

	Características	
Tipo	Porcentaje Texto Ancla	Porcentaje Texto Visible
No Spam	34%	76%
Spam	37%	87%



MATRIZ DE CONFUSIÓN

Introducción
Problema
Metodología
Implementación
→ **Evaluación y resultados**
Conclusión
Trabajo Futuro
Bibliografía

		Clases Predichas	
		1	**-1**
Clases Conocidas	**1**	89,00%	0
	-1	10,30%	0

- Dataset compuesto mayormente por páginas no spam.
- Se predice con certeza cuando una página no es Web spam.

MEDICIONES EC2

Mediciones realizadas en los clusters de Amazon

Nodos EC2	Extracción Tiempo (seg.)	Entrenamiento Tiempo (seg.)
3	83.127	302
8	50.54	250

Kernel: RBF (Radial Basis Function)

Número de vectores: 67,577

Número de vectores de apoyo: 20,338

5

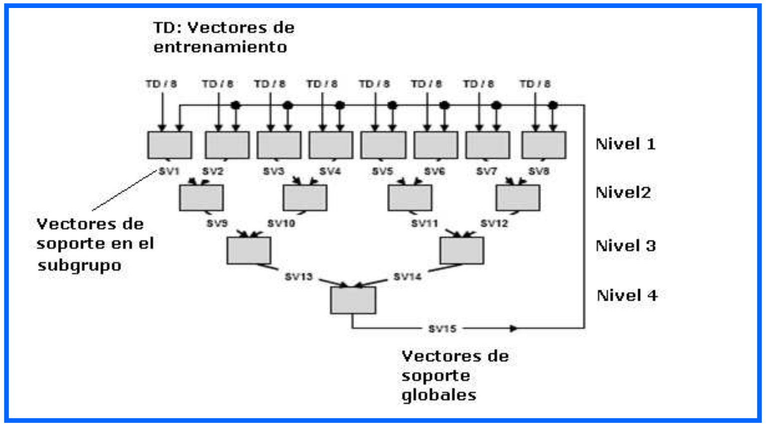
CONCLUSIONES Y TRABAJOS FUTUROS

CONCLUSIONES

Introducción
Problema
Metodología
Implementación
Evaluación y resultados
→ Conclusión
Trabajo Futuro
Bibliografía

- **Hadoop es una herramienta poderosa**
- **Servicios Web de Amazon ahorran costos**
- **SVM herramienta muy capaz de clasificación**
- **Solución Rendimiento SVM**

hadoop +



TRABAJO FUTURO

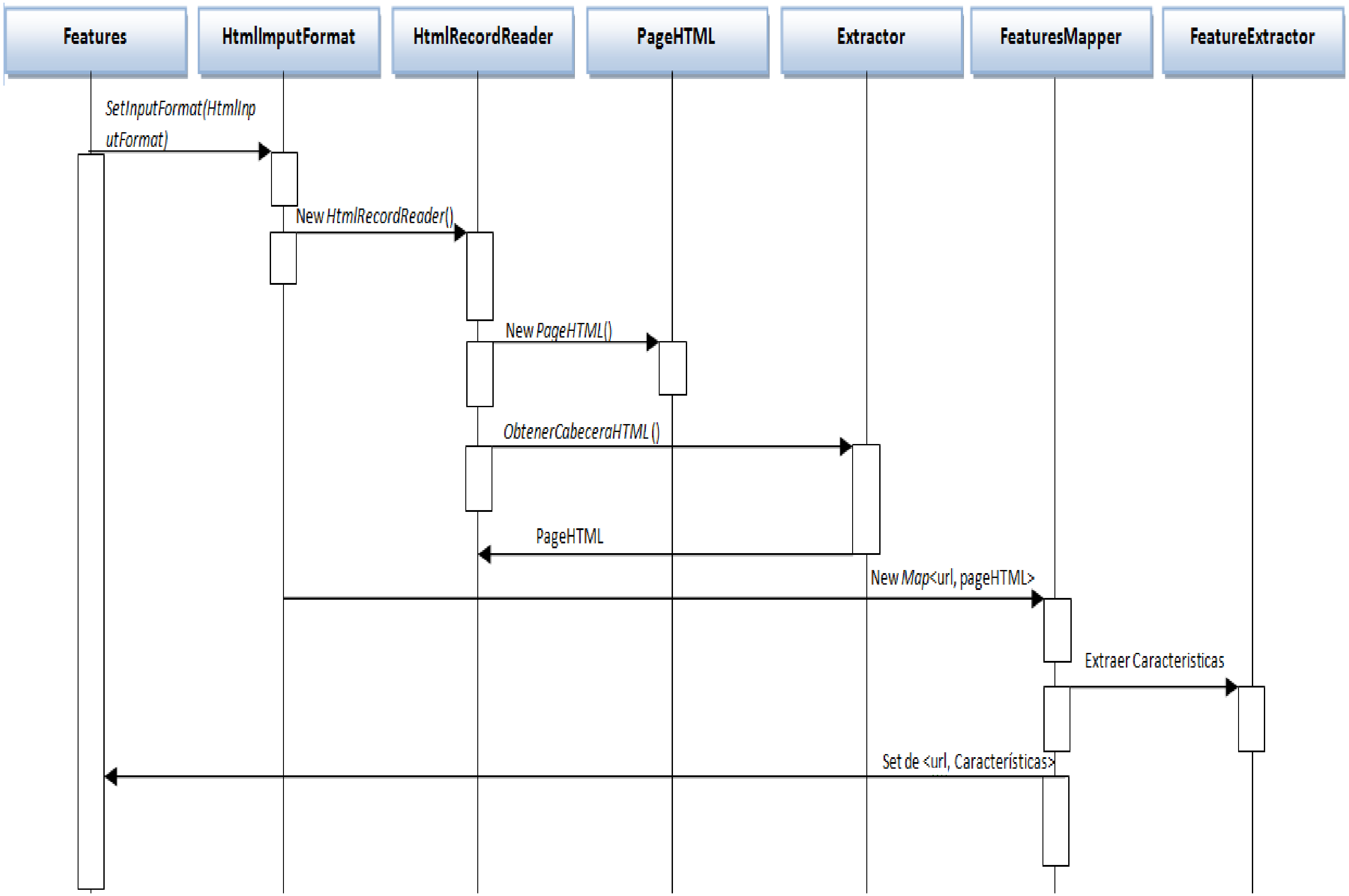
Introducción
Problema
Metodología
Implementación
Evaluación y resultados
Conclusión
→ Trabajo Futuro
Bibliografía

- **Otra solución en "Sub-problemas cuadráticos"**
- **Extender la cantidad de características.**
- **Mecanismos de validación cruzada para el ajuste de parámetros.**

Preguntas

EXTRAER Y CUANTIFICAR

- Introducción
- Problema
- → **Metodología**
- Implementación
- Evaluación y resultados
- Conclusión
- Trabajo Futuro
- Bibliografía



Features

HtmlImputFormat

HtmlRecordReader

PageHTML

Extractor

FeaturesMapper

FeatureExtractor

SetInputFormat(HtmlInputFormat)

New *HtmlRecordReader()*

New *PageHTML()*

ObtenerCabeceraHTML ()

PageHTML

New *Map*<url, pageHTML>

Extraer Caracteristicas

Set de <url, Características>