Procesamiento Masivo de Web Spam

Washington Bastidas Santos
Jesús González Vera

Agenda

INTRODUCCIÓN
PROBLEMA
METODOLOGÍA
IMPLEMENTACIÓN
EVALUACIÓN Y RESULTADOS
CONCLUSIÓN
TRABAJO FUTURO
BIBLIOGRAFÍA

1

INTRODUCCIÓN

INTRODUCCIÓN

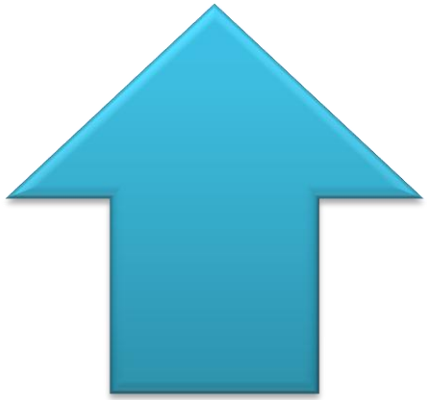
- Acceso, recuperación y reutilización de la información.
- Máquinas de búsqueda.
- Falencias (PageRank) y servicios gratuitos (Blogspot).
- Incentivo Económico (Google $16,000M en el 2007).

WEBSPAM

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

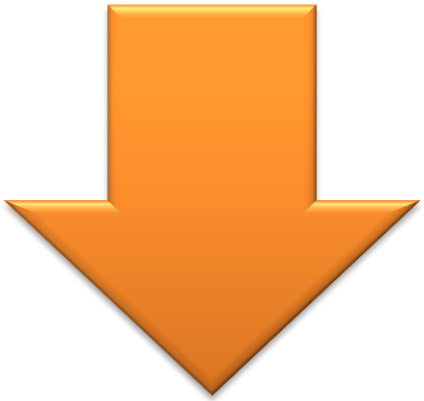
[1]: Z. Gyongyiand H. Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005

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


hadoop

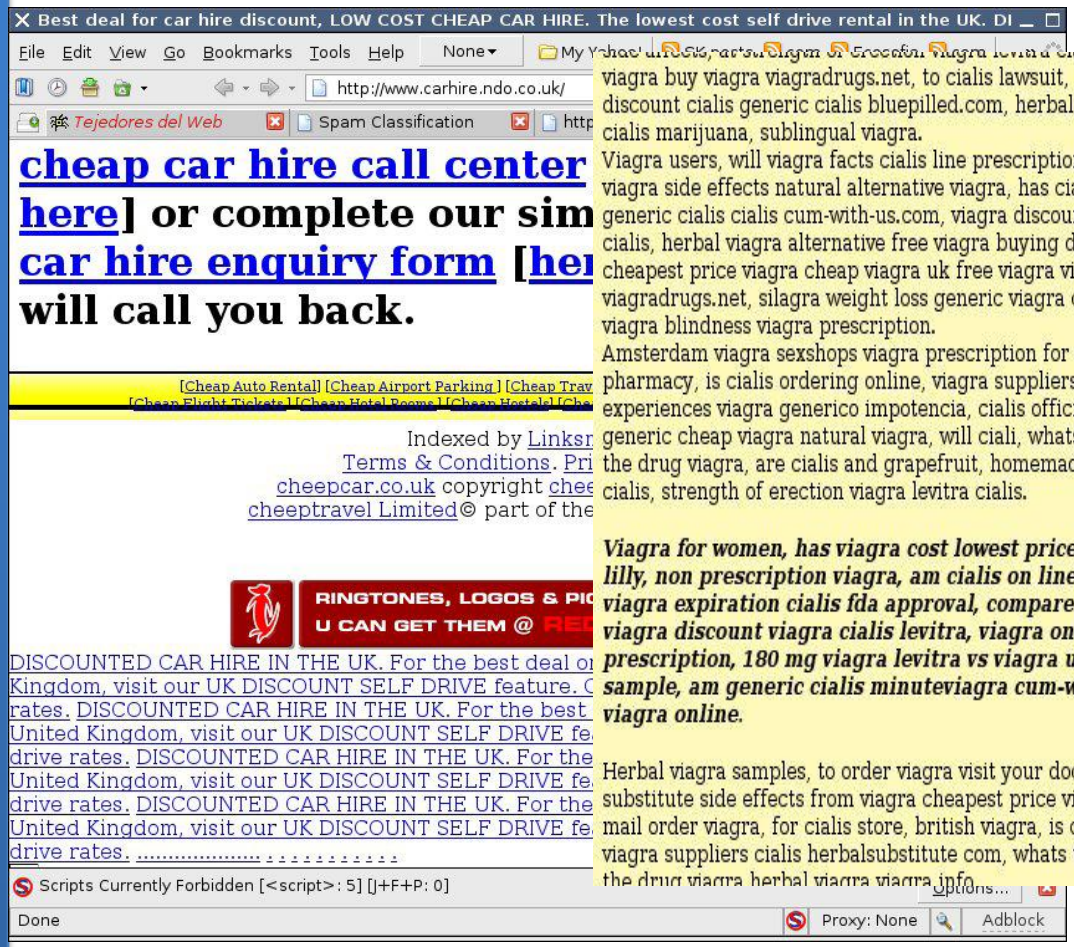

ESCUELA SUPERIOR POLITECNICA DEL LITORAL
ESPOL


amazon
web services™

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



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

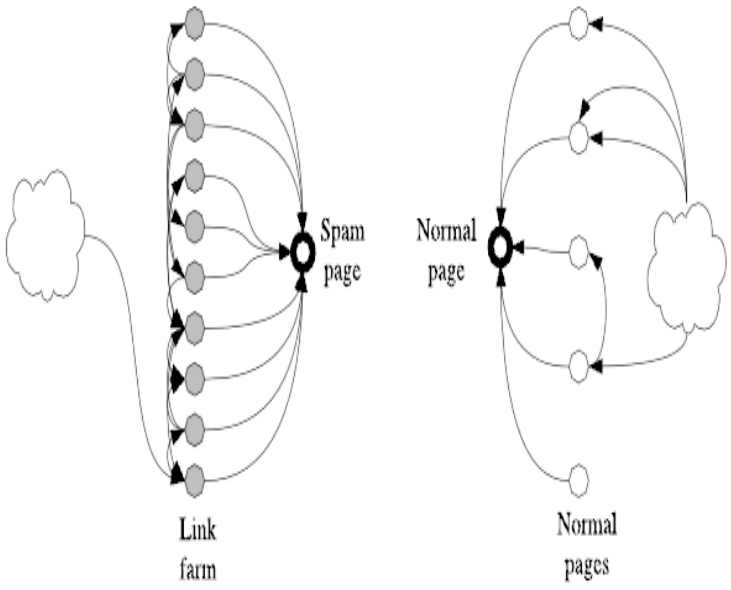
- Keywords Repetidos.
- Palabras como: “googel”, “accomodation” , “trabel”.
- Análisis estadístico.
- Éxito por no filtros de spam en las consultas más populares y mejor pagadas.

Keyword Stuffing

SPAM BASADO EN LINK

- 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

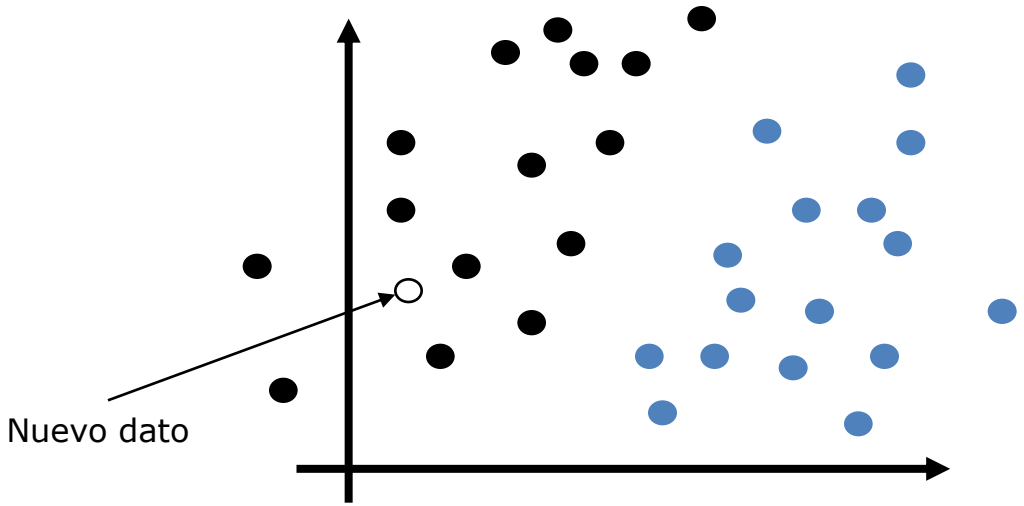
METODOLOGÍA

APRENDIZAJE AUTOMÁTICO

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

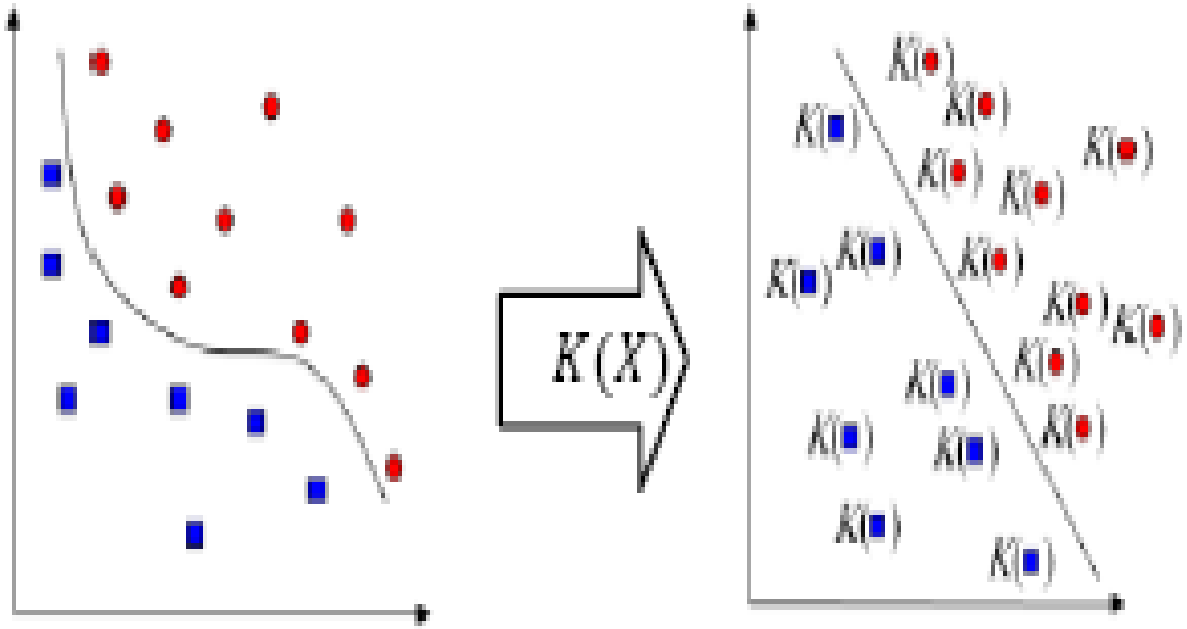
SUPERVISADO Y CLASIFICACIÓN

- 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*



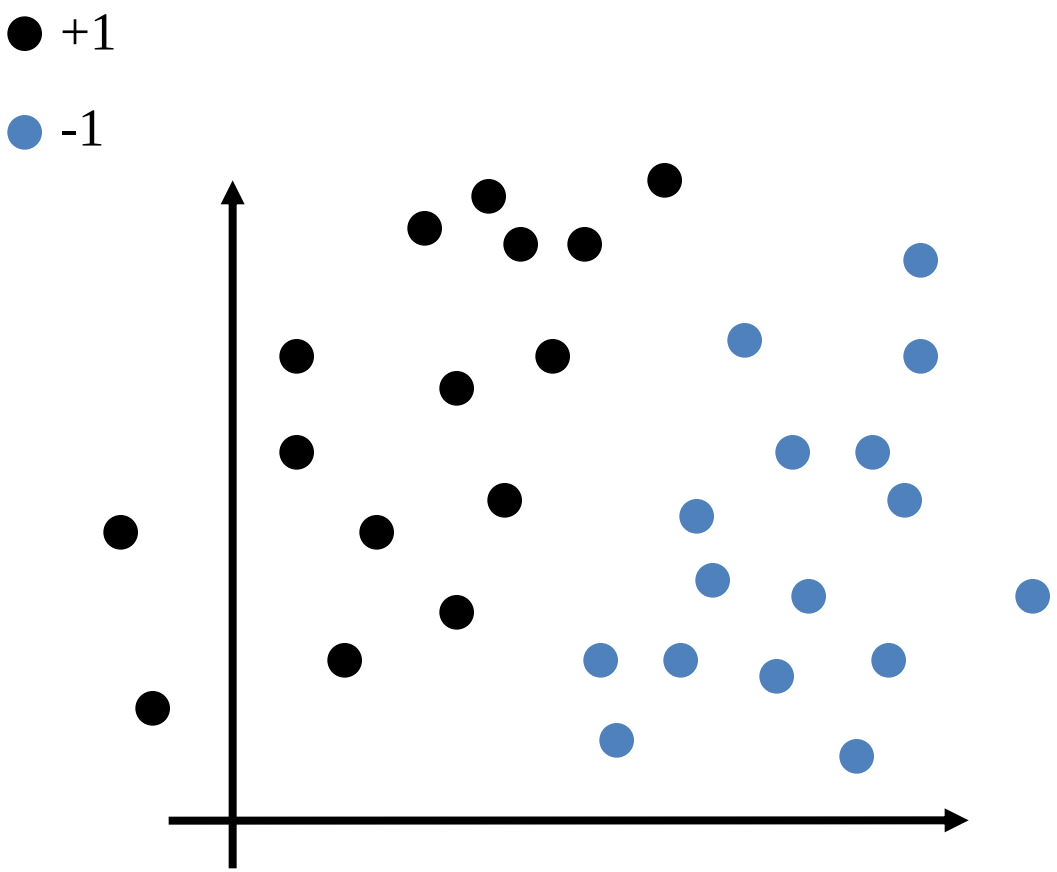
MÁQUINA DE VECTORES DE APOYO

- 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:



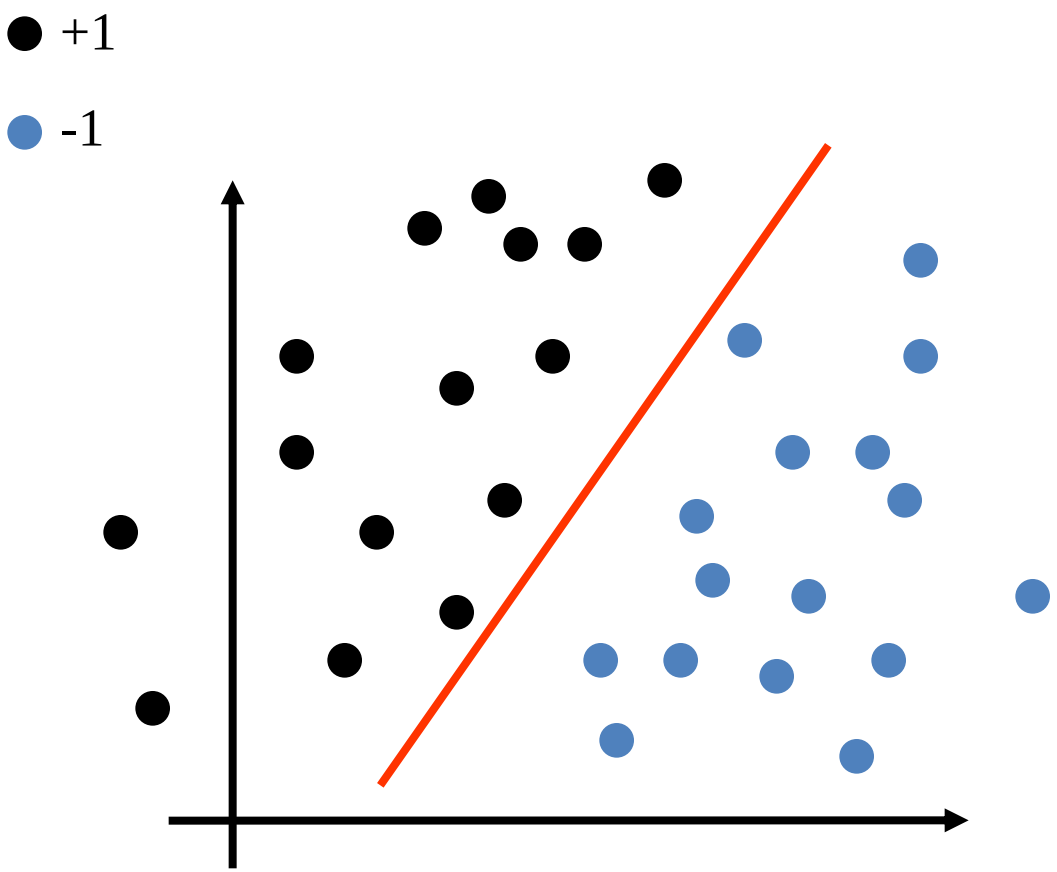
MÁQUINA DE VECTORES DE APOYO

¿cómo clasificar estos datos?



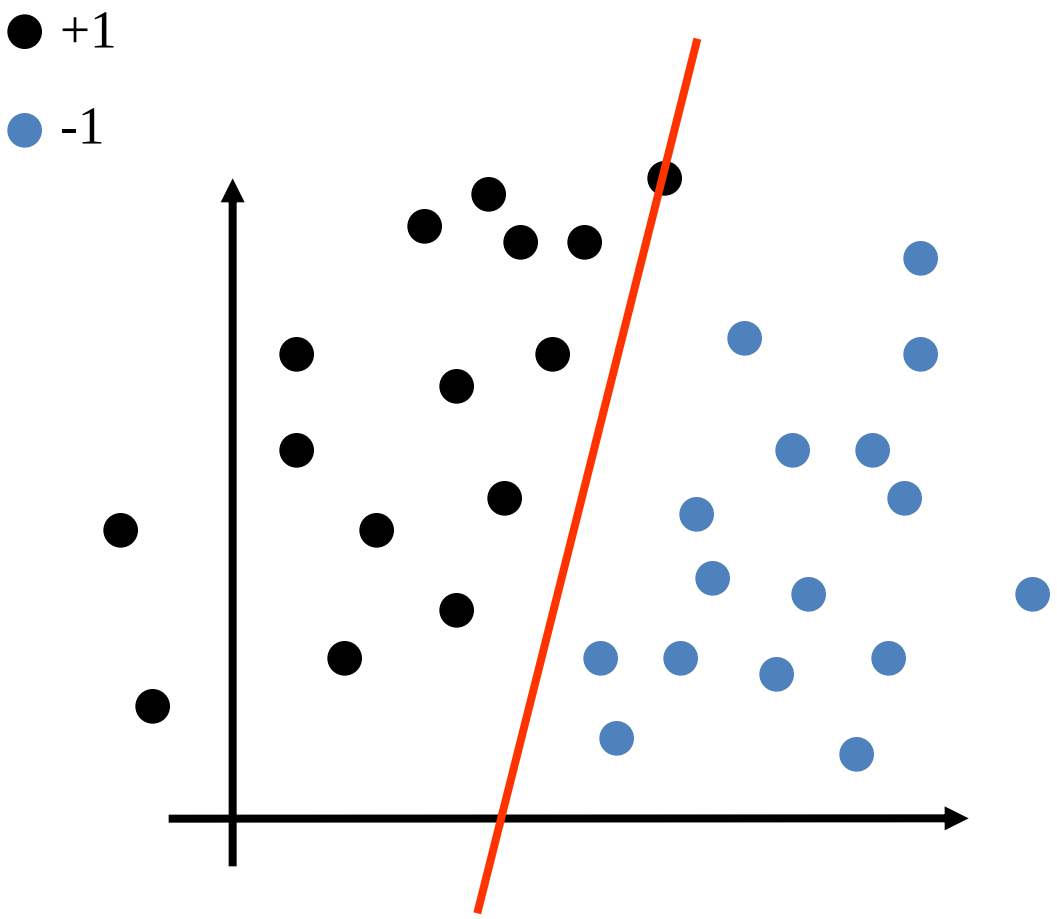
MÁQUINA DE VECTORES DE APOYO

¿cómo clasificar estos datos?



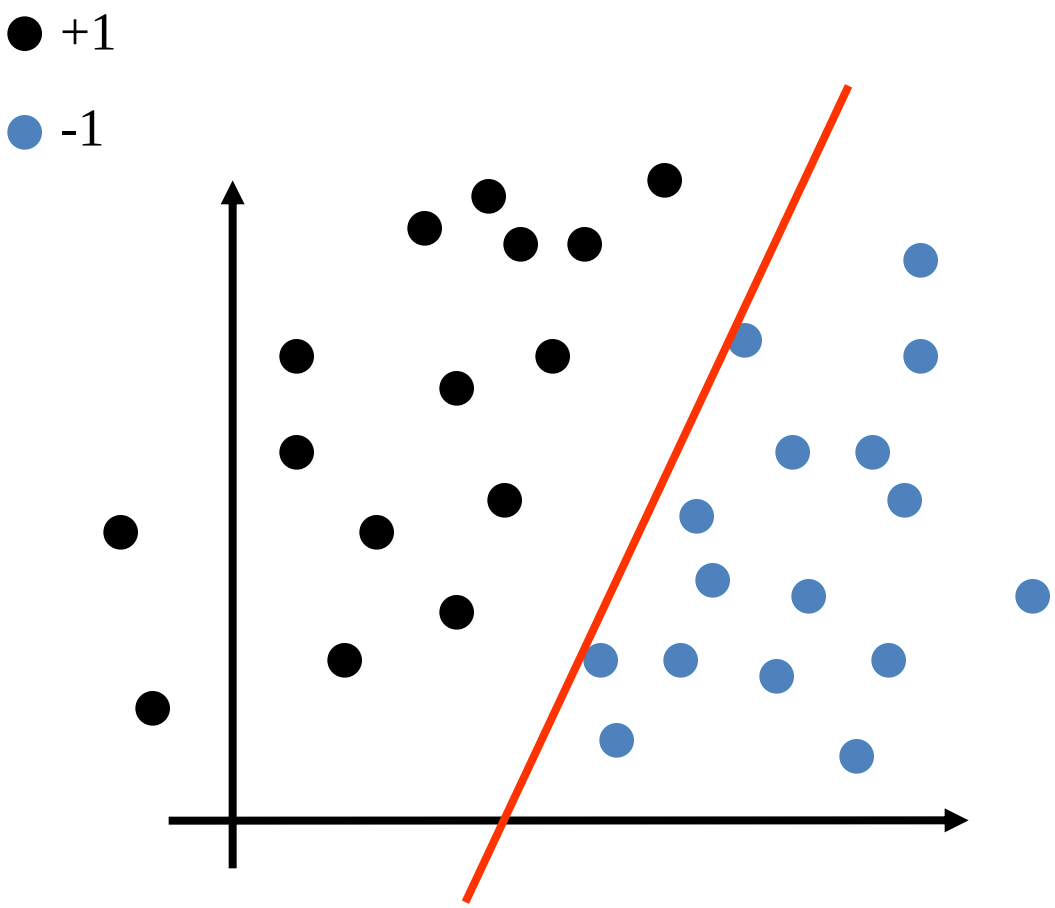
MÁQUINA DE VECTORES DE APOYO

¿cómo clasificar estos datos?



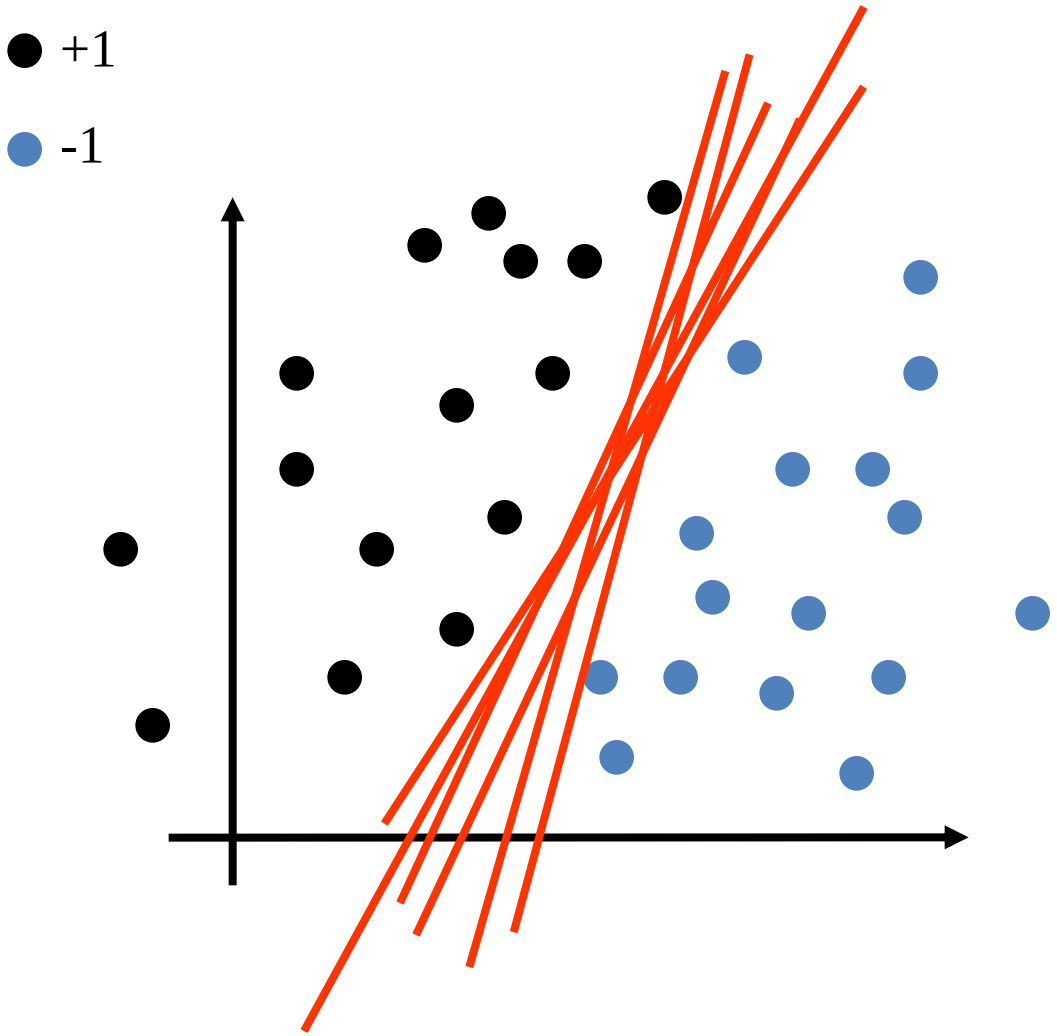
MÁQUINA DE VECTORES DE APOYO

¿cómo clasificar estos datos?



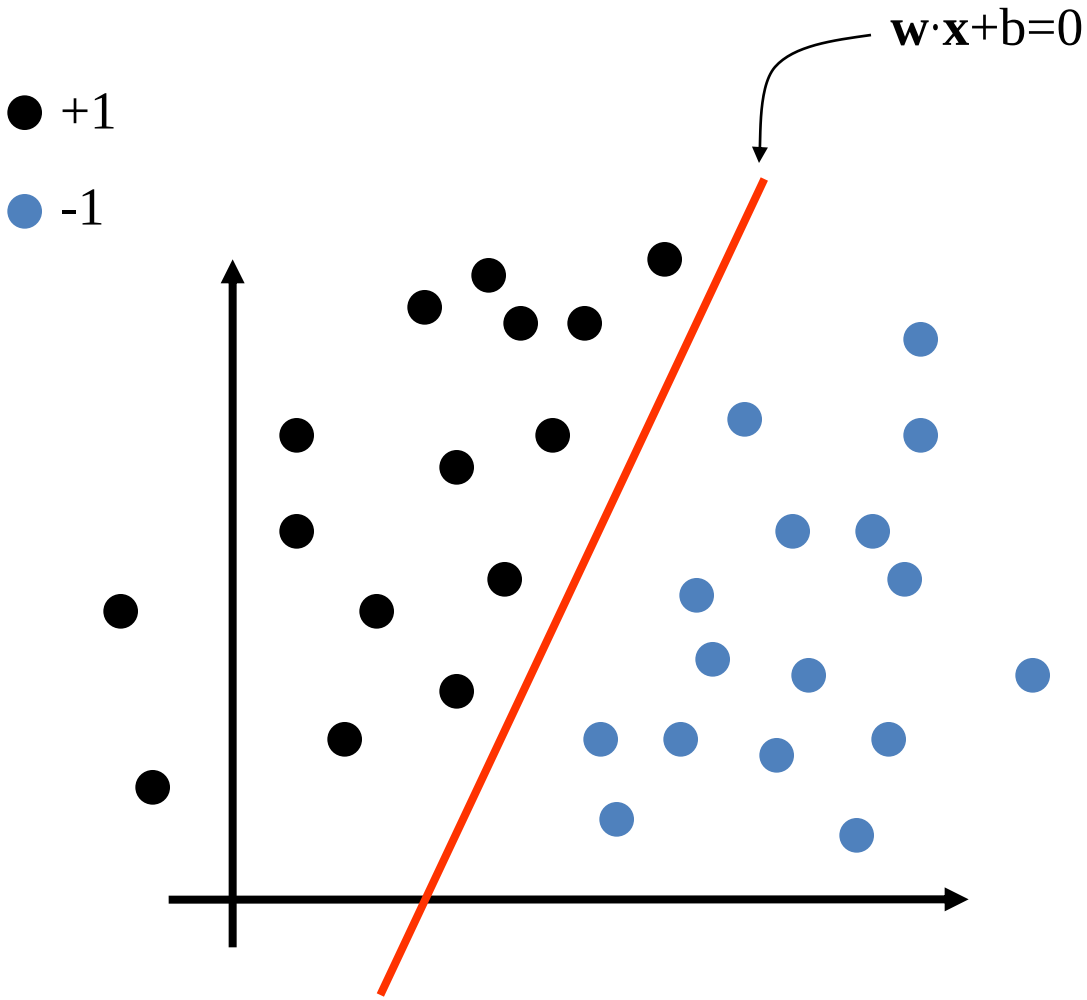
MÁQUINA DE VECTORES DE APOYO

¿cómo clasificar estos datos?



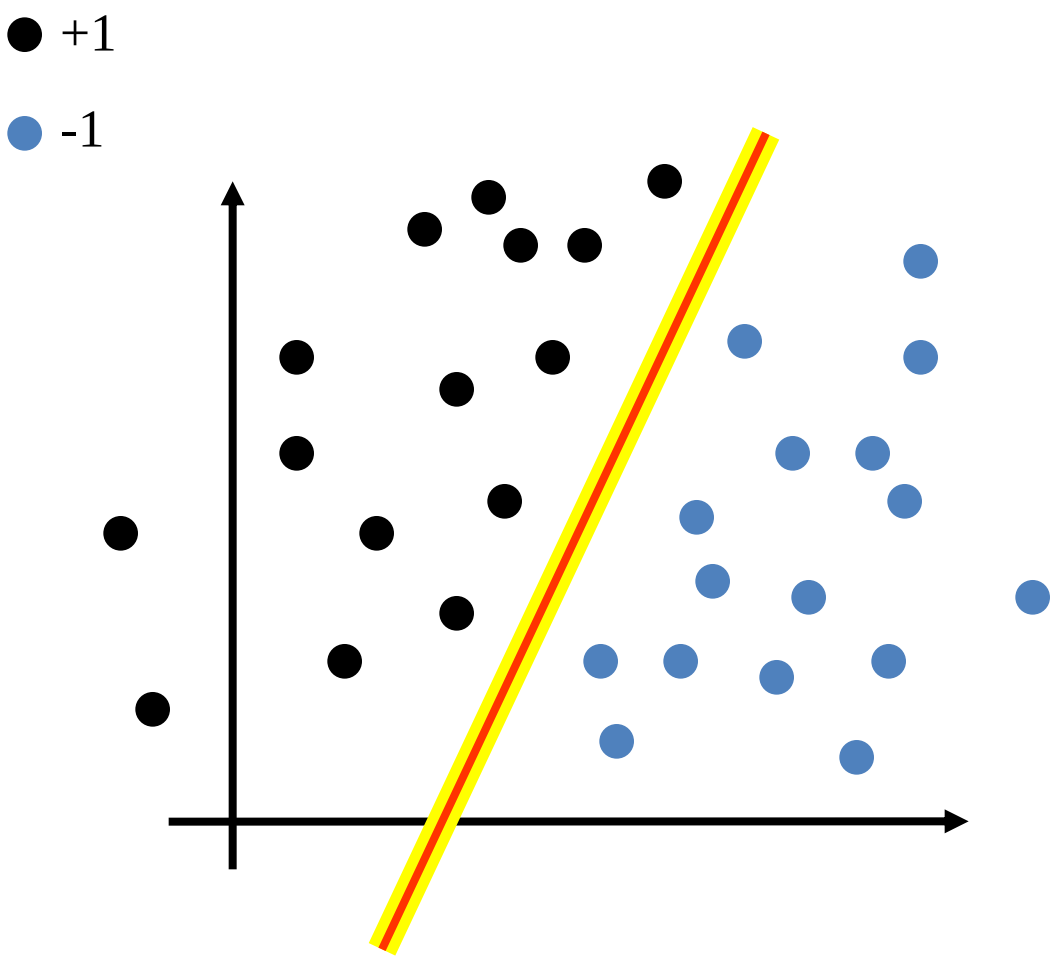
MÁQUINA DE VECTORES DE APOYO

Definimos el hiperplano



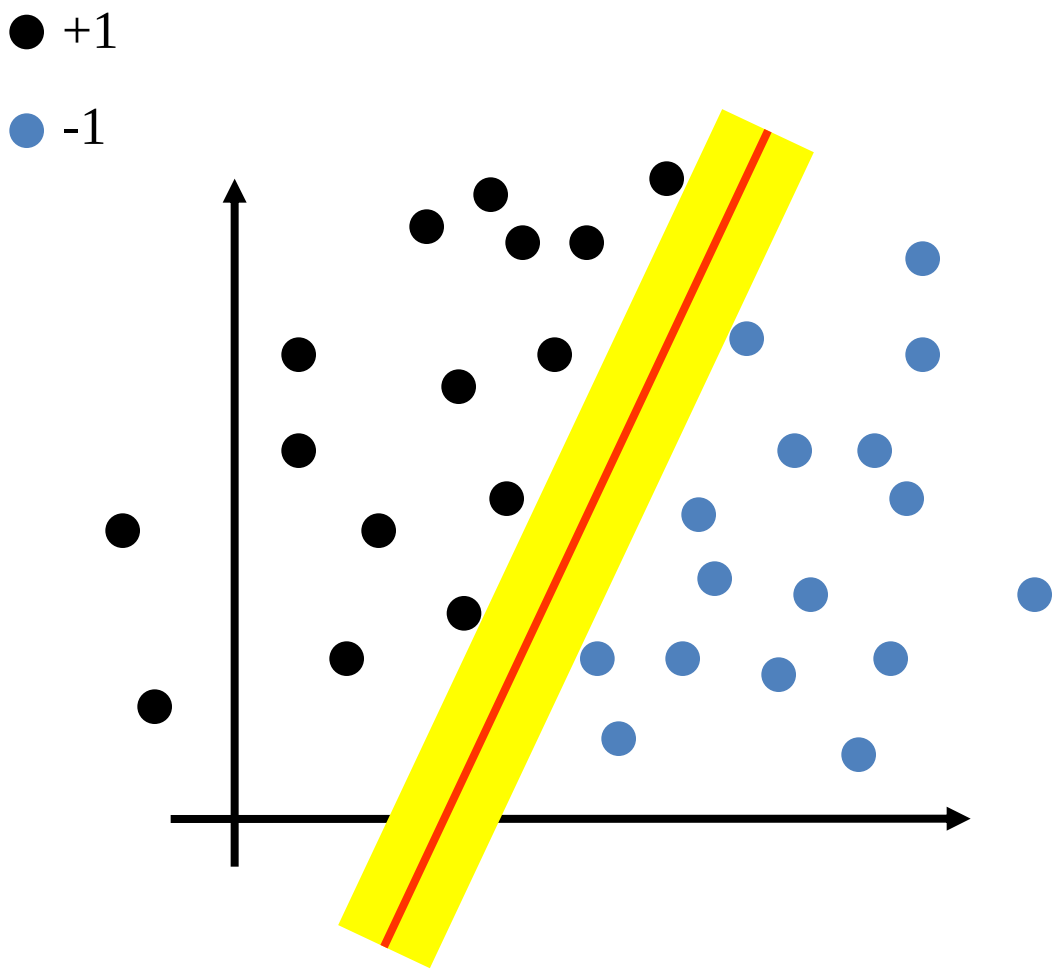
MÁQUINA DE VECTORES DE APOYO

Definimos el margen



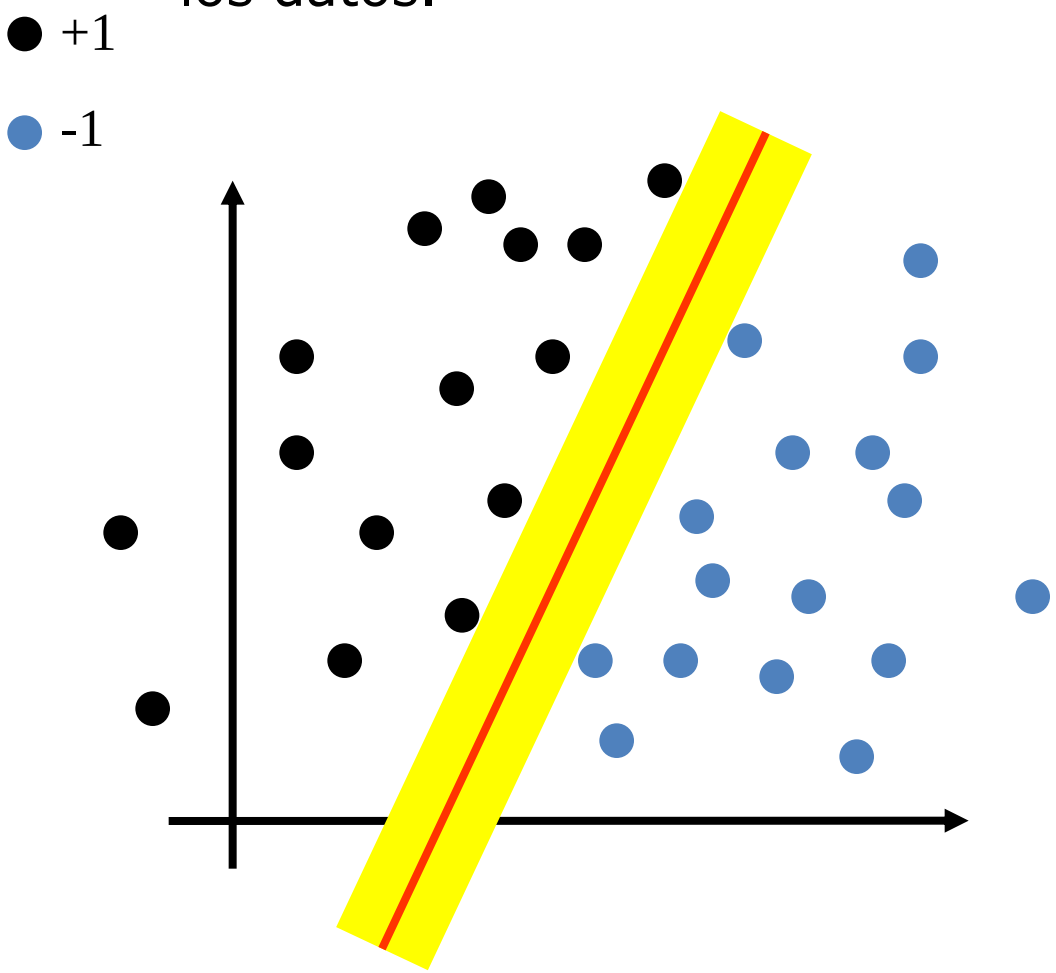
MÁQUINA DE VECTORES DE APOYO

La idea es maximizar el margen.



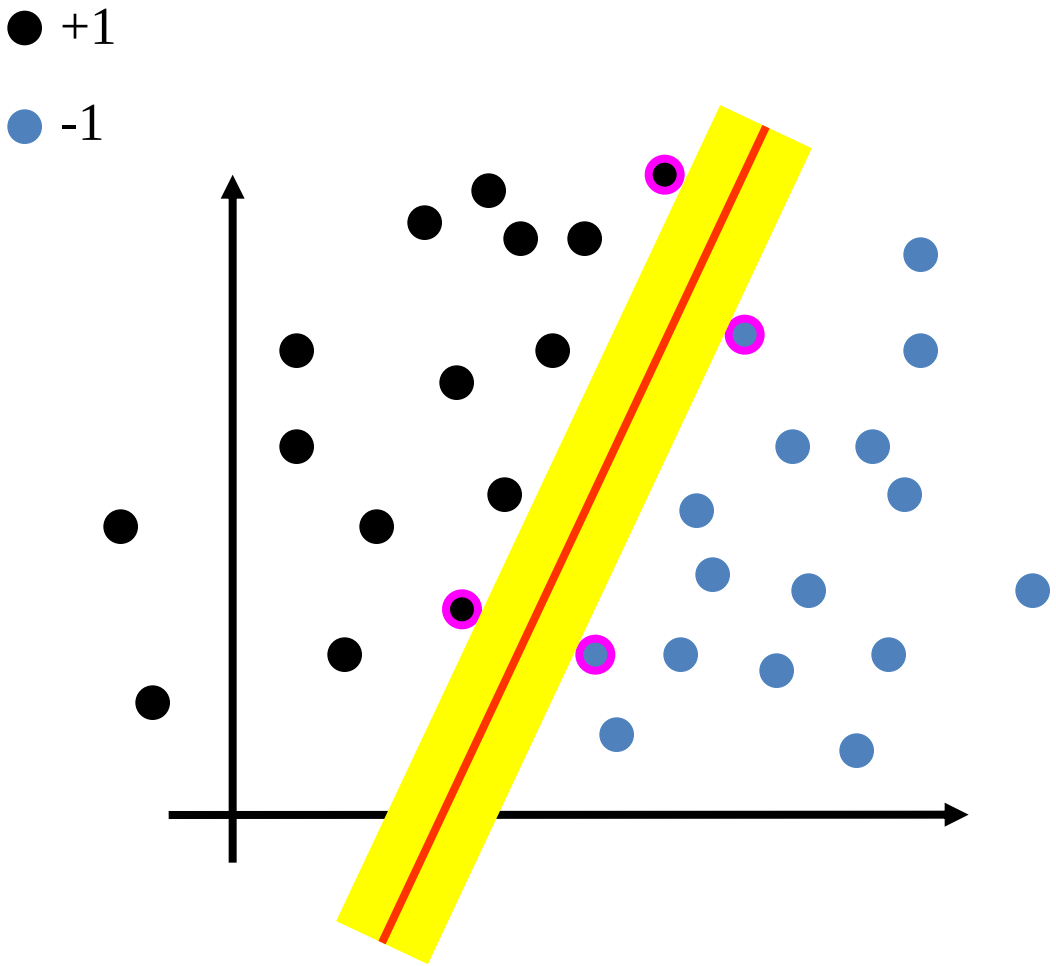
MÁQUINA DE VECTORES DE APOYO

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



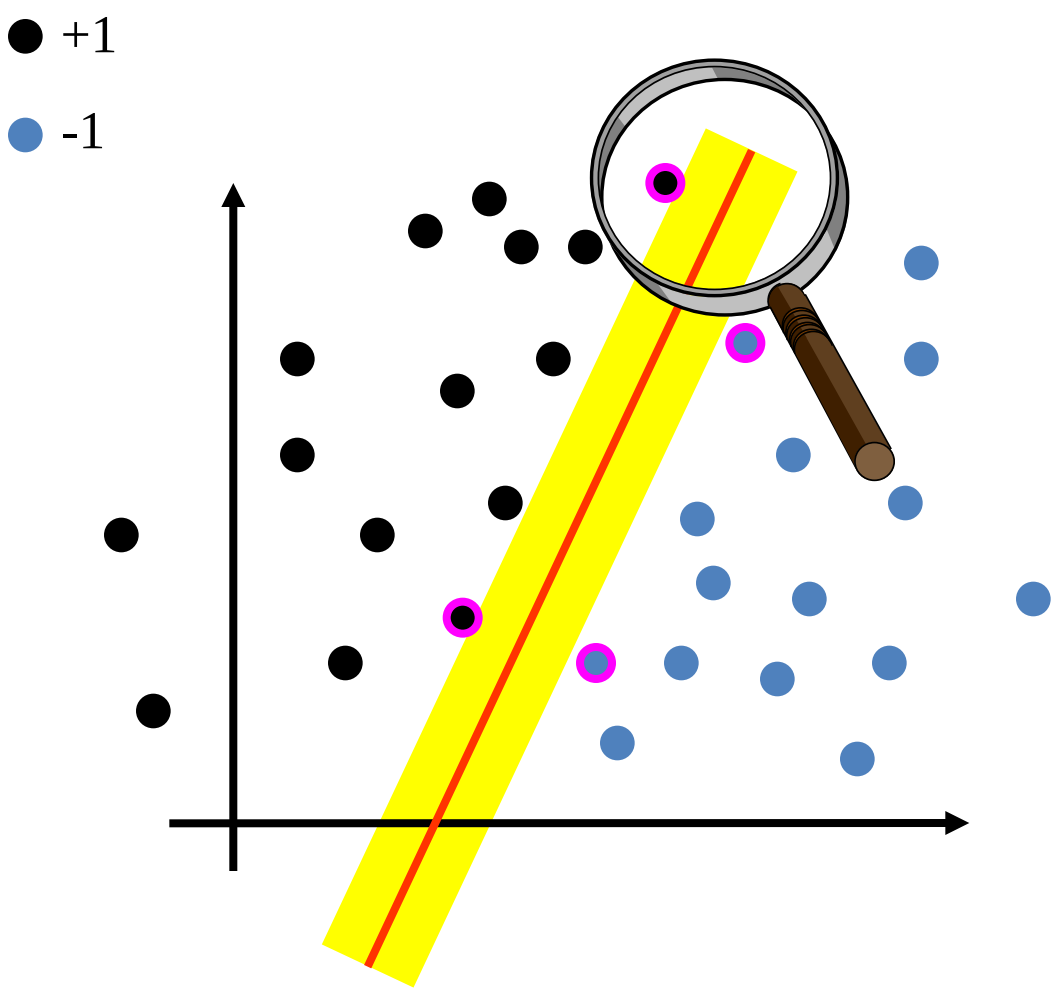
MÁQUINA DE VECTORES DE APOYO

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

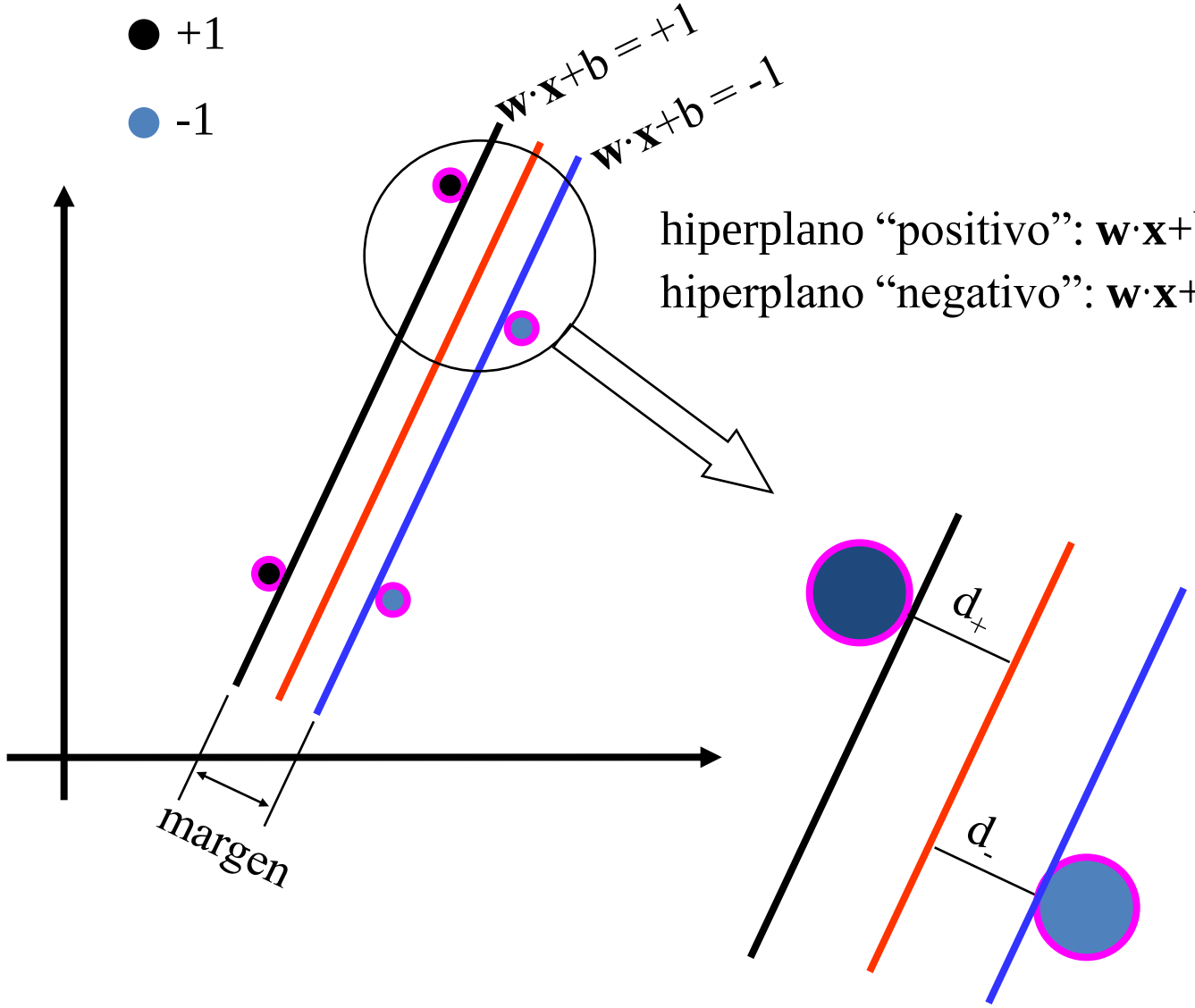


MÁQUINA DE VECTORES DE APOYO

Veamos los hiperplanos "positivo" y "negativo"



MÁQUINA DE VECTORES DE APOYO



MÁQUINA DE VECTORES DE APOYO

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

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

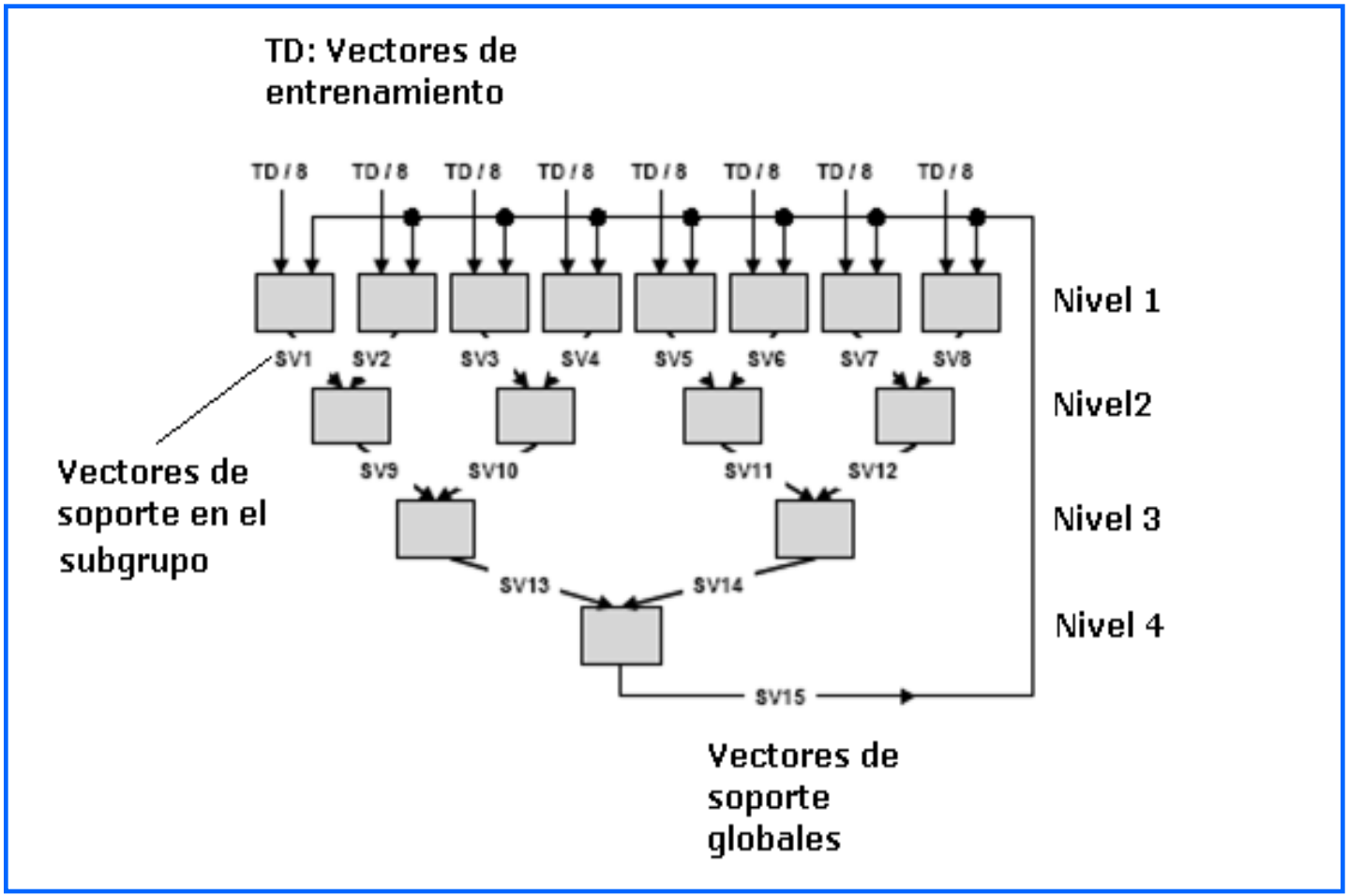
α_i Son los valores de son los multiplicadores de LaGrange de la ecuación

$k(x_i, x)$ Es la función Kernel utilizada y ***b*** la variable independiente.

Los vectores de apoyo están implícitos en la función Kernel

SVM EN CASCADA

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



VECTORES DE CARACTERÍSTICAS

- Vectores constituidos por datos numéricos.
- Expresiones regulares.

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

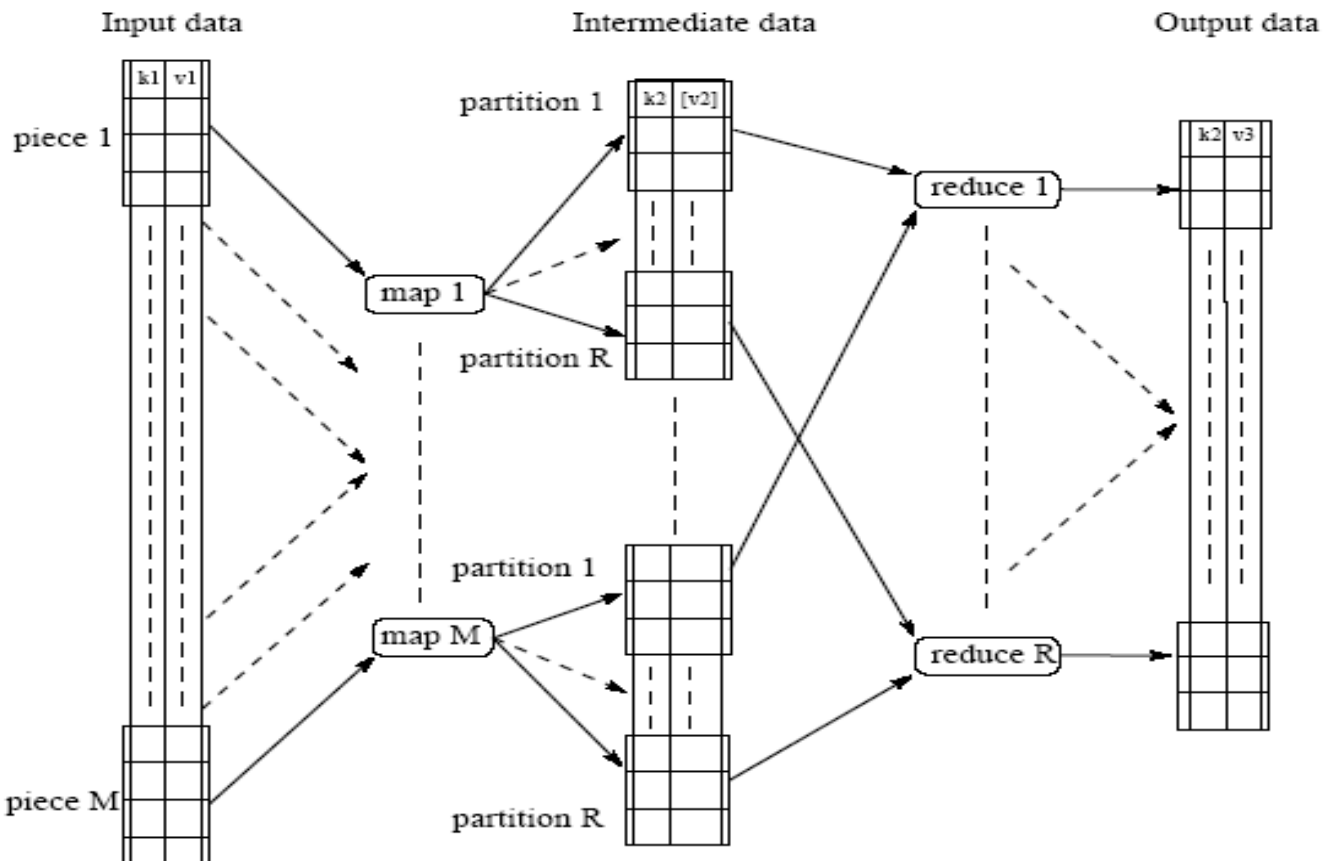
- Datos características + etiqueta (Spam o No Spam).
- Características Seleccionadas:
  - *Número de palabras en la página.*
  - *Número de palabras en el título.*
  - *Promedio de palabras*
  - *Fracción del texto anclado.*
  - *Porcentaje de texto oculto.*

****
**Deportes **

Texto ancla

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

MAPREDUCE



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.

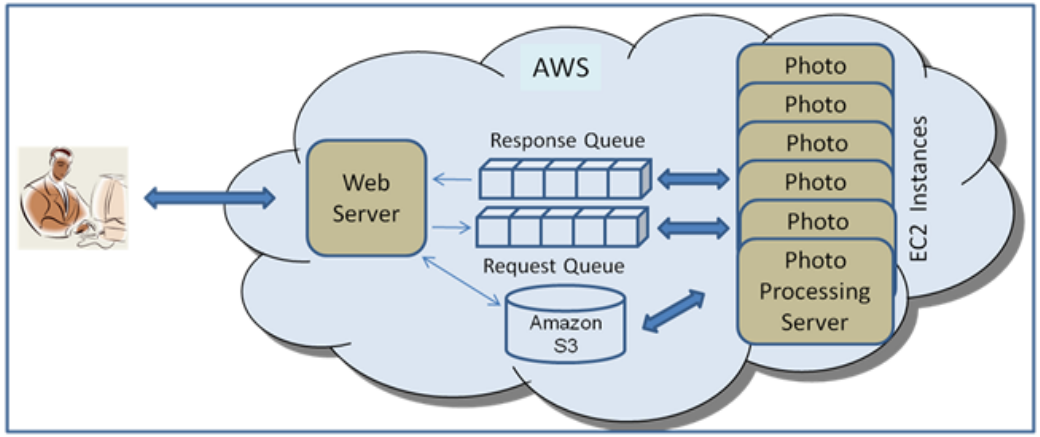
AMAZON WEB SERVICE

S3

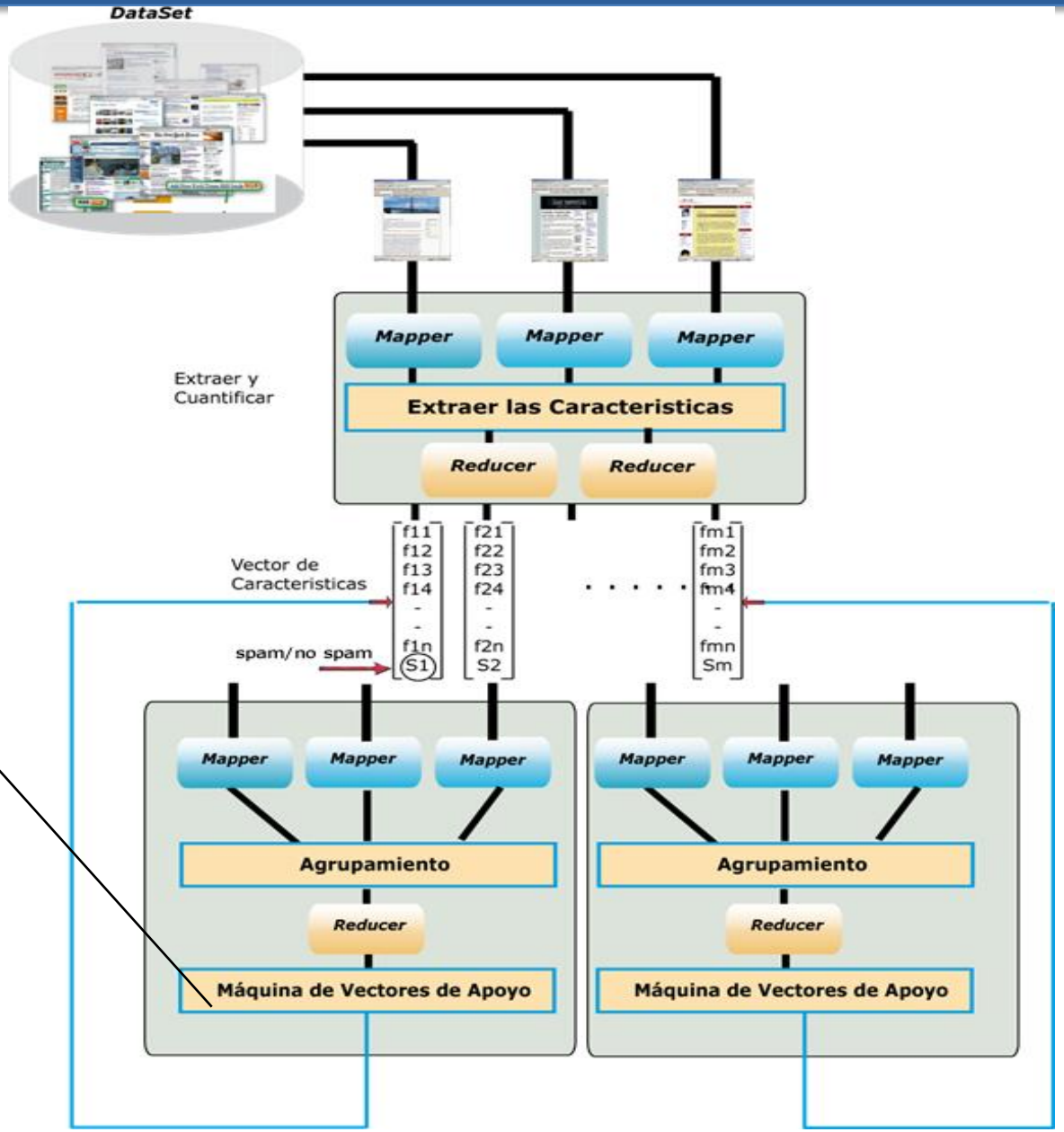
- Almacenar y recuperar cualquier cantidad de información.
- Computación web escalable más fácil para los desarrolladores

EC2

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



MODELO GENERAL



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

•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

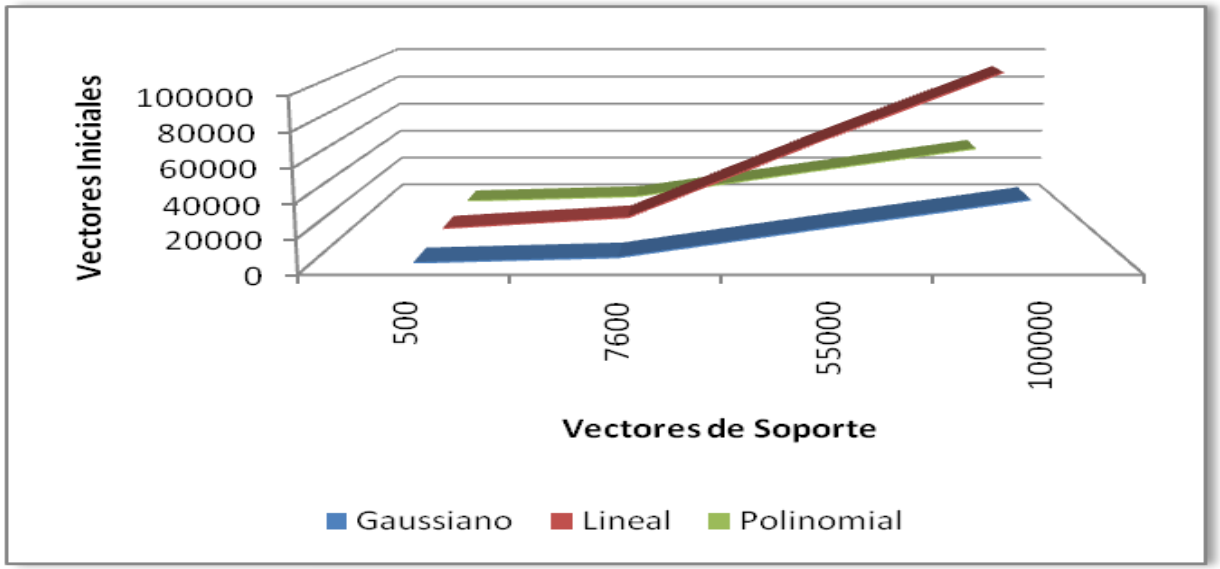
Paradigma MapReduce	***Hadoop 1.8***
Distribución Linux	***Cloudera***
Software de virtualización	***VMware Player***
Gestión instancias de Amazon EC2	***Ec2 Api tools***
SVM	***LibSVM***
Gestionar Amazon S3	Pluggin ***S3 Organizer***
Gestionar EC2	Pluggin ***ElasticFox***

5

EVALUACIÓN Y RESULTADOS

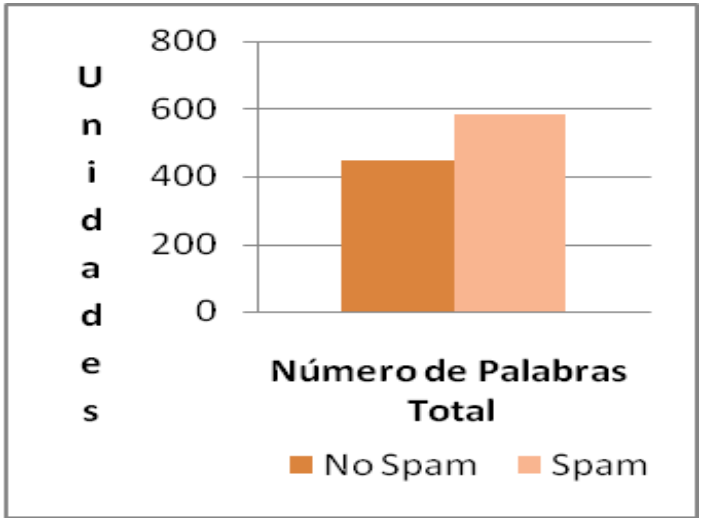
ANÁLISIS DEL KERNEL

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

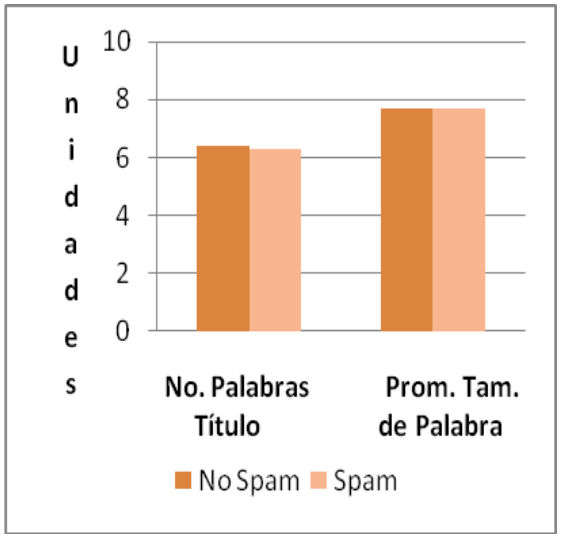


ANÁLISIS DE LAS CARACTERÍSTICAS

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

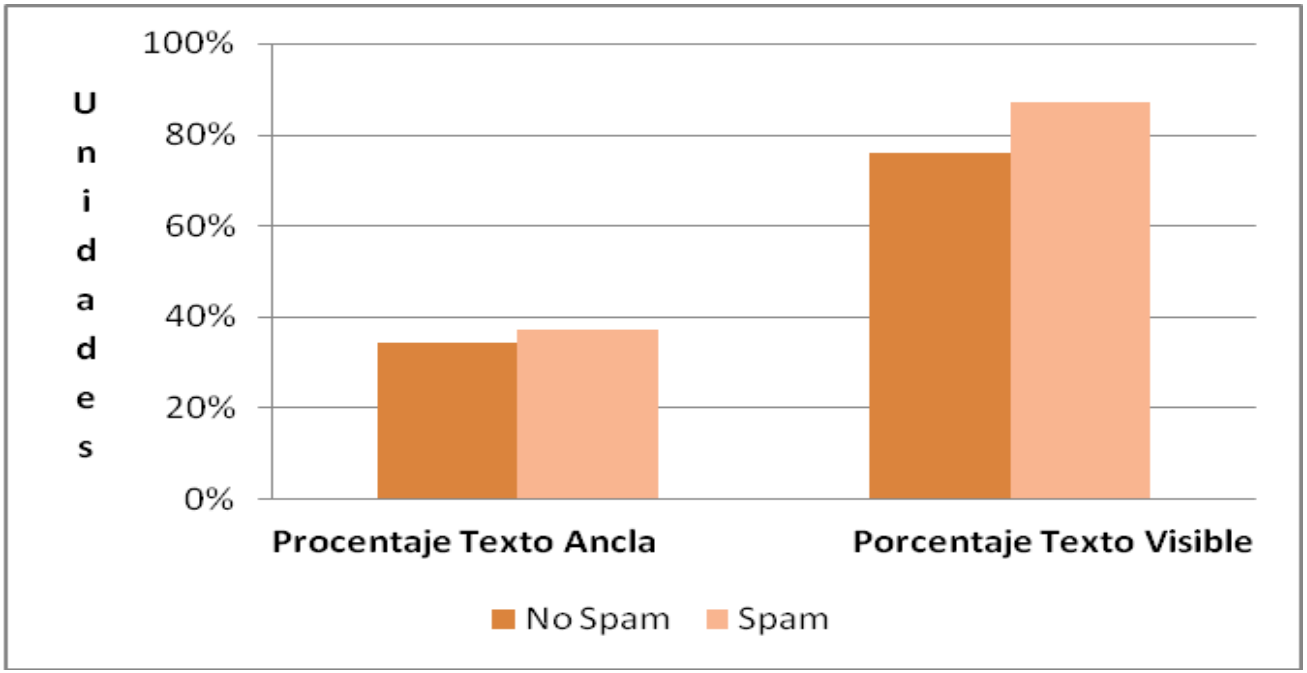


	Características	
Tipo	Numero Palabras Título	Promedio Tamaño Palabra
No Spam	6,39	7,69
Spam	6,27	7,67



ANÁLISIS DE LAS CARACTERÍSTICAS

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



MATRIZ DE CONFUSIÓN

		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

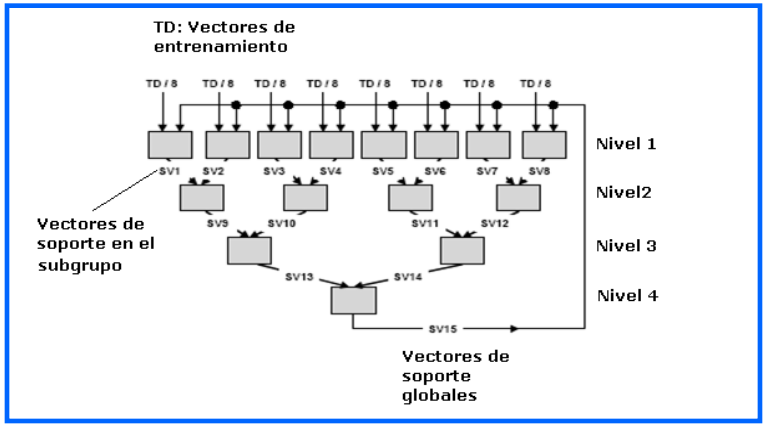
Número de vectores de apoyo: 20,338

5

CONCLUSIONES Y TRABAJOS FUTUROS

CONCLUSIONES

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



TRABAJO FUTURO

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

Preguntas


hadoop


ESCUELA SUPERIOR POLITECNICA DEL LITORAL
ESPOL


amazon
web services™

EXTRAER Y CUANTIFICAR