



APLICACIONES DE DATA SCIENCE



¿Cuál es la utilidad del Procesamiento de Lenguaje Natural?



Logro Unidad 2

El estudiante podrá utilizar las habilidades que necesita para aplicar la ciencia de datos a las diferentes tareas de la minería de textos.



Contenido 3

Modelo de espacio vectorial y recuperación de información

- Recuperación de Información
- Tokens y Tipos
- Extracción de Vocabulario
- Modelo de Espacio Vectorial
- Frecuencia de Documento Invertida
- Similitudes entre vectores



Contenido 3

Modelo de espacio vectorial y recuperación de información

- Recuperación de Información
- Tokens y Tipos
- Extracción de Vocabulario
- Modelo de Espacio Vectorial
- Frecuencia de Documento Invertida
- Similitudes entre vectores

- Modelo Vectorial



1. Recuperación de Información



1. Recuperación de Información

- ¿Como un buscador como Google o Duckduckgo devuelven documentos (paginas web) relevantes a partir de una consulta (query)?





1. Recuperación de Información

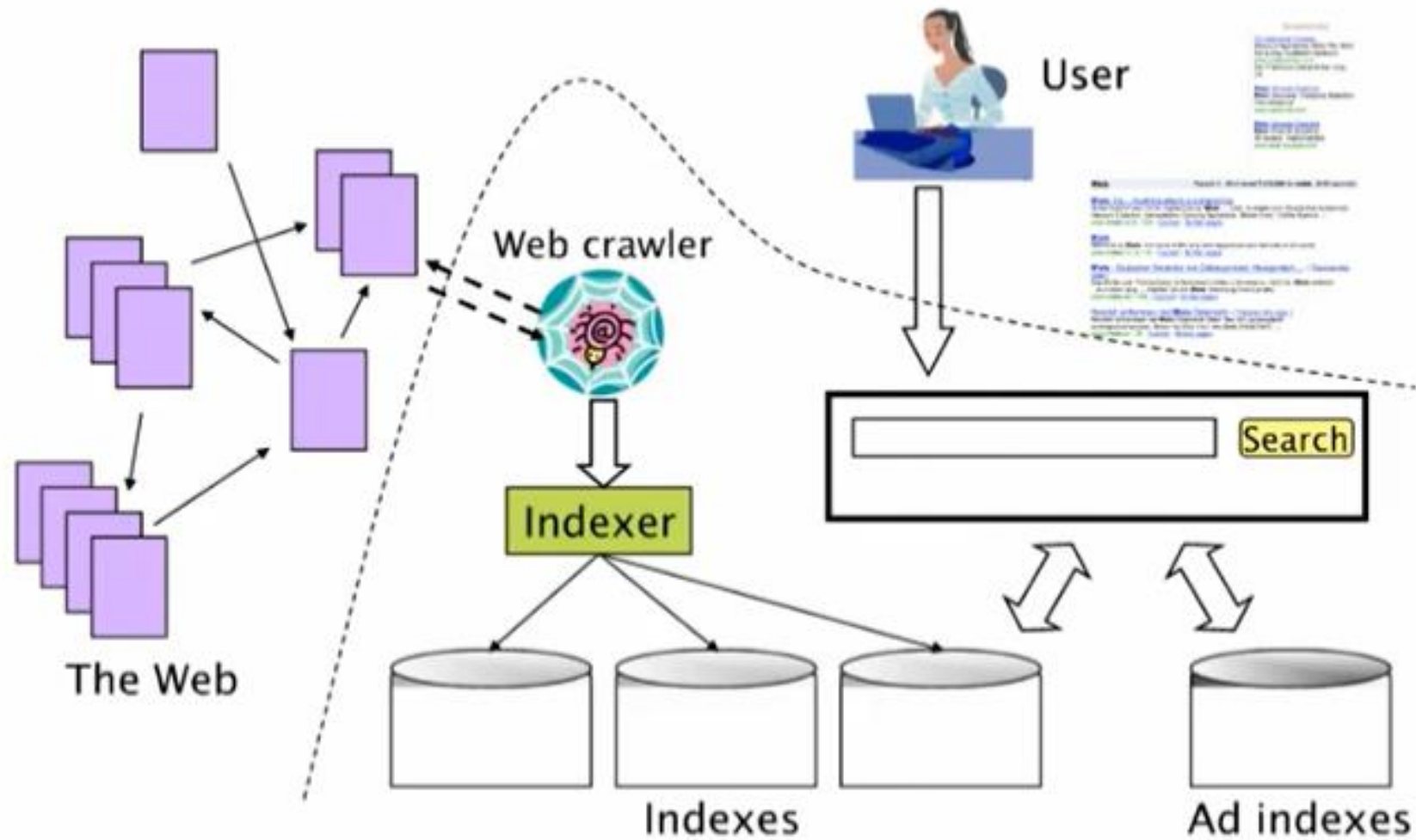
- ¿Como un buscador como Google o Duckduckgo devuelven documentos (paginas web) relevantes a partir de una consulta (query)?



- Existe un proceso de recuperación de información (Web Search Engines) y tiene los siguientes componentes:



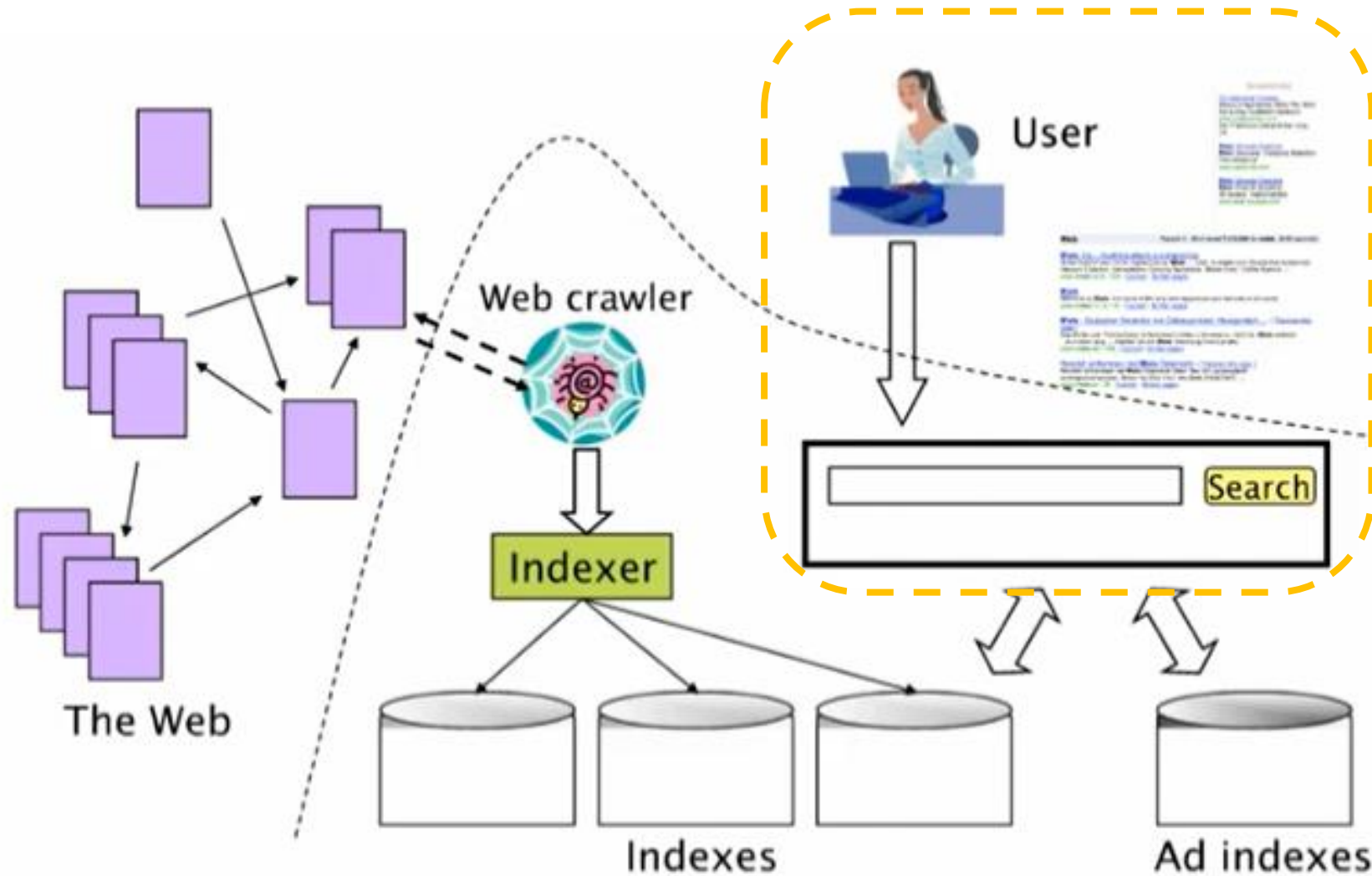
1. Recuperación de Información





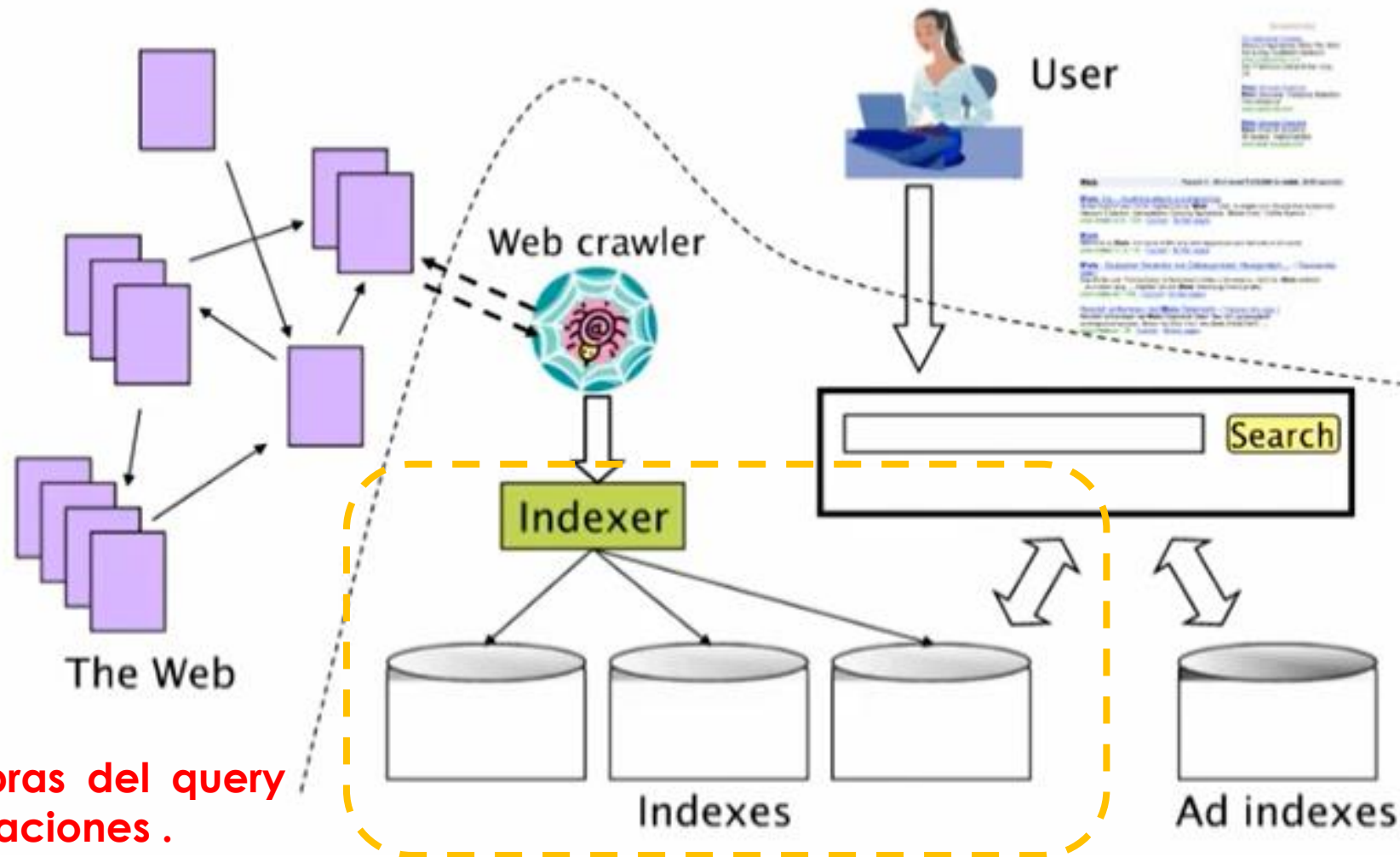
1. Recuperación de Información

Interface de usuario que recibe la consulta como entrada y retorno los documentos rankeados.





1. Recuperación de Información

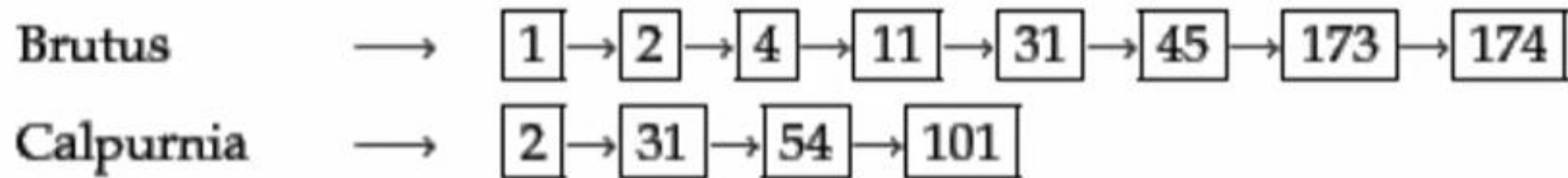


Mapear las palabras del query en listas de publicaciones .



1. Recuperación de Información

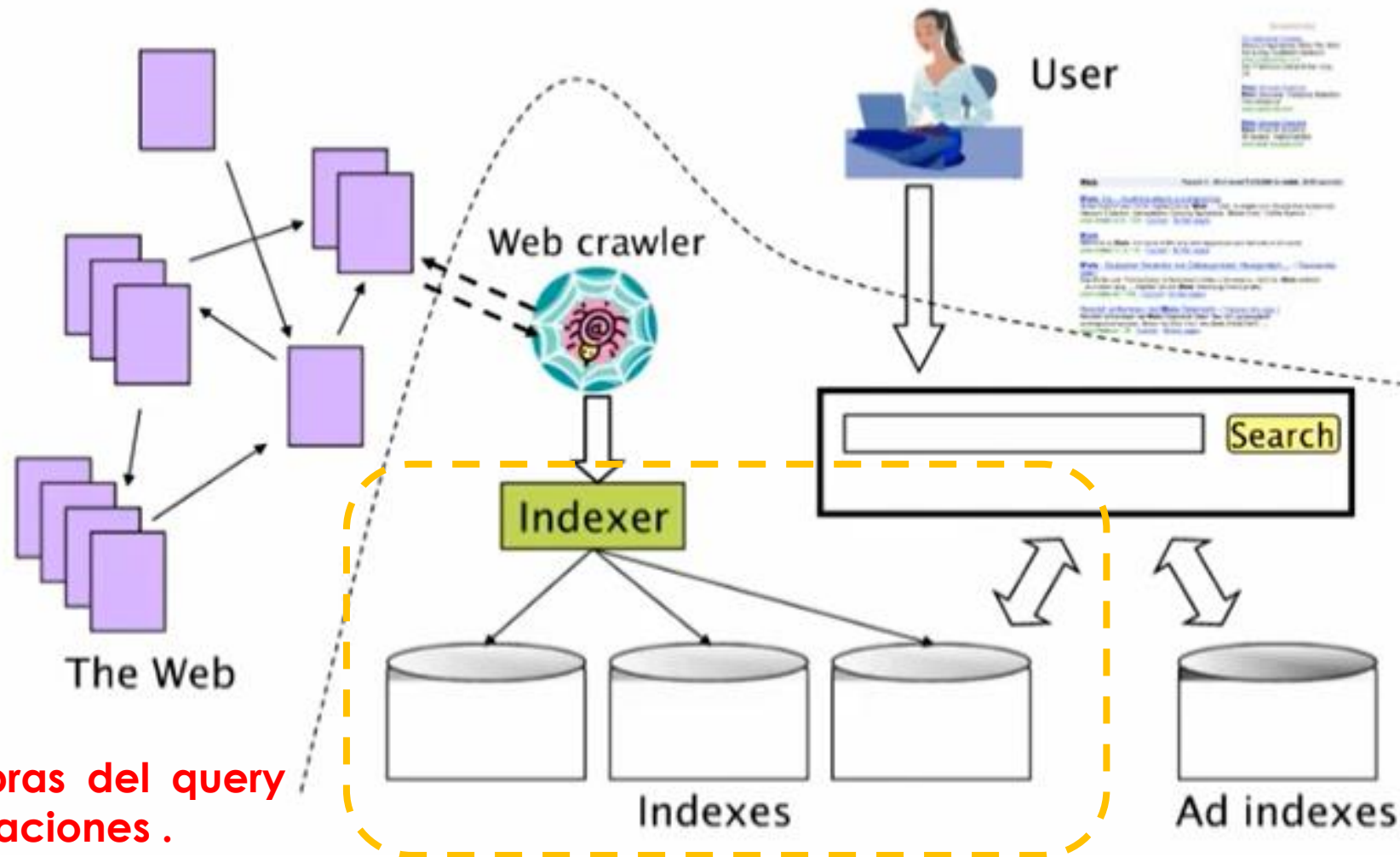
- Una lista de publicaciones de un término es un lista de todos los documentos donde el término aparece por lo menos una vez.



**Mapear las palabras del query
en listas de publicaciones .**



1. Recuperación de Información

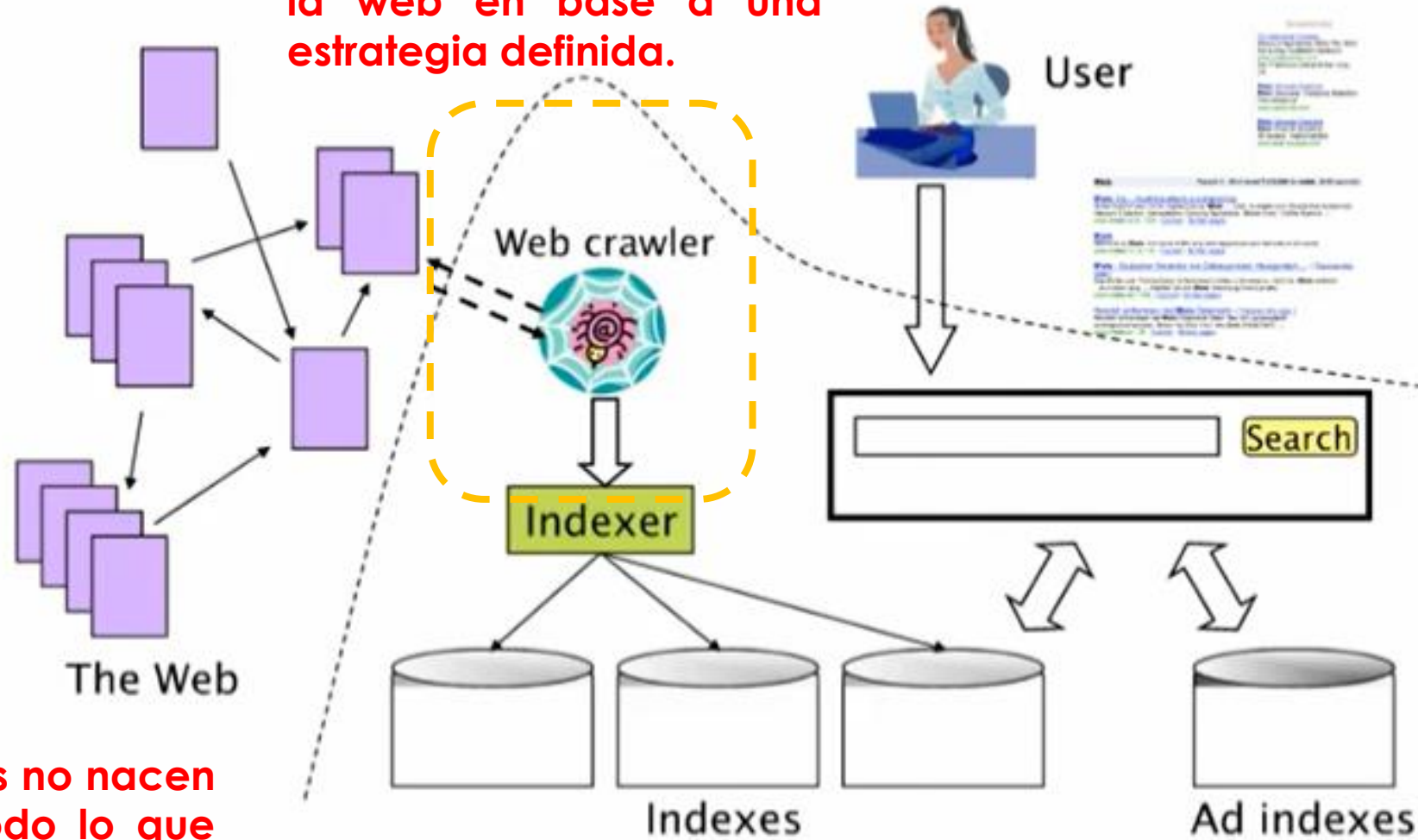


Mapear las palabras del query en listas de publicaciones .



1. Recuperación de Información

Un robot que navega en la web en base a una estrategia definida.



Los buscadores no nacen conociendo todo lo que esta la red!



Contenido 3

Modelo de espacio vectorial y recuperación de información

- Recuperación de Información
- **Tokens y Tipos**
- Extracción de Vocabulario
- Modelo de Espacio Vectorial
- Frecuencia de Documento Invertida
- Similitudes entre vectores



2. Modelo de Espacio Vectorial

¿Cómo pasar el texto a algo
procesable algorítmicamente?



2.1. Tokens y Tipos

- **Tokenización:** Es la tarea de dividir una oración o un documento en piezas llamadas **TOKENS** (*palabras*).



2.1. Tokens y Tipos

- **Tokenización:** Es la tarea de dividir una oración o un documento en piezas llamadas **TOKENS** (*palabras*).
- Ejemplo:

Input: I like human languages and programming languages.

Tokens: [I] [like] [human] [languages] [and] [programming] [languages]



2.1. Tokens y Tipos

- **Tokenización**: Es la tarea de dividir una oración o un documento en piezas llamadas **TOKENS** (*palabras*).

- Ejemplo:

Input: I like human languages and programming languages.

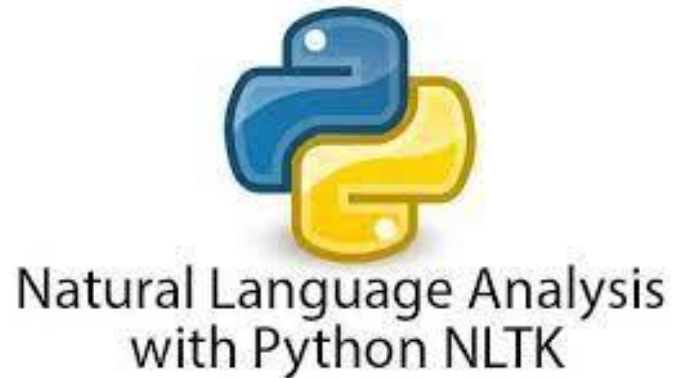
Tokens: [I] [like] [human] [languages] [and] [programming] [languages]

- Un tokenizador puede hacer ciertas transformaciones adicionales al texto como eliminación de caracteres especiales (ex.: signos de puntuación), transformar las palabras a minúsculas, etc.



2.1. Tokens y Tipos

- Algunas librerías ya vienen tokenizadas:



spaCy



2.1. Tokens y Tipos

- **Tipos:** Son una clase de tokens que contine una secuencia única de caracteres. Es decir, los tipos se obtienen identificando tokens únicos dentro del documento.



2.1. Tokens y Tipos

- **Tipos:** Son una clase de tokens que contine una secuencia única de caracteres. Es decir, los tipos se obtienen identificando tokens únicos dentro del documento.
- Ejemplo:

Input: I like human languages and programming languages.

Tokens: [I] [like] [human] [languages] [and] [programming] [languages]



2.1. Tokens y Tipos

- **Tipos:** Son una clase de tokens que contine una secuencia única de caracteres. Es decir, los tipos se obtienen identificando tokens únicos dentro del documento.
- Ejemplo:

Input: I like human languages and programming languages.

Tokens: [I] [like] [human] [languages] [and] [programming] [languages]



2.1. Tokens y Tipos

- Tipos:** Son una clase de tokens que contine una secuencia única de caracteres. Es decir, los tipos se obtienen identificando tokens únicos dentro del documento.
- Ejemplo:

Input: I like human languages and programming languages.

Tokens: [I] [like] [human] [languages] [and] [programming] [languages]

Types for the previous sentence: [I] [like] [human] [languages] [and] [programming]



2.1. Tokens y Tipos

- Tipos:** Son una clase de tokens que contine una secuencia única de caracteres. Es decir, los tipos se obtienen identificando tokens únicos dentro del documento.
- Ejemplo:

Input: I like human languages and programming languages.

Tokens: [I] [like] [human] [languages] [and] [programming] [languages]

Types for the previous sentence: [I] [like] [human] [languages] [and] [programming]

- La idea es recorrer el corpus y si encuentra una **palabra repetida** no la considero como una palabra nueva.



Contenido 3

Modelo de espacio vectorial y recuperación de información

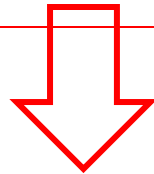
- Recuperación de Información
- Tokens y Tipos
- **Extracción de Vocabulario**
- Modelo de Espacio Vectorial
- Frecuencia de Documento Invertida
- Similitudes entre vectores



2.2. Extracción de Vocabulario

- El **vocabulario** es el conjunto de todos los **TIPOS** que salen de mi corpus. En otras palabras.

- Estos tipos pueden ser normalizados y obtener **TERMINOS**.



- La normalización es un proceso de crear clases equivalentes de diferentes tipos con el objetivo de que mi vocabulario sea más pequeño.
- Ejemplo: plurales y los singulares sean del mismo tipo.
PERSONA y PERSONAS sean el mismo objeto.



2.2. Extracción de Vocabulario

- El **vocabulario V** es el conjunto de **términos** (tokens únicos normalizados) dentro de mi colección de documentos o **corpus D**. [Manning et al., 2008]



2.2. Extracción de Vocabulario

- Existen algoritmos que realizan el proceso de normalización, por ejemplo el **Algoritmo de Porter's** en el cual los términos se transforman a su raíz para poder reducir el tamaño del vocabulario. Y se lleva a cabo aplicando reglas de reducción de palabras.

Example: Porter's Algorithm.

(F)	Rule	
	SSES	→ SS
	IES	→ I
	SS	→ SS
	S	→

Example	
caresses	→ caress
ponies	→ poni
caress	→ caress
cats	→ cat



2.2. Extracción de Vocabulario

Types for the previous sentence: [I] [like] [human] [languages] [and] [programming]

termId	value
t1	human
t2	languag
t3	program

- El tamaño de mi vocabulario influye en el tamaño de mis vectores.**



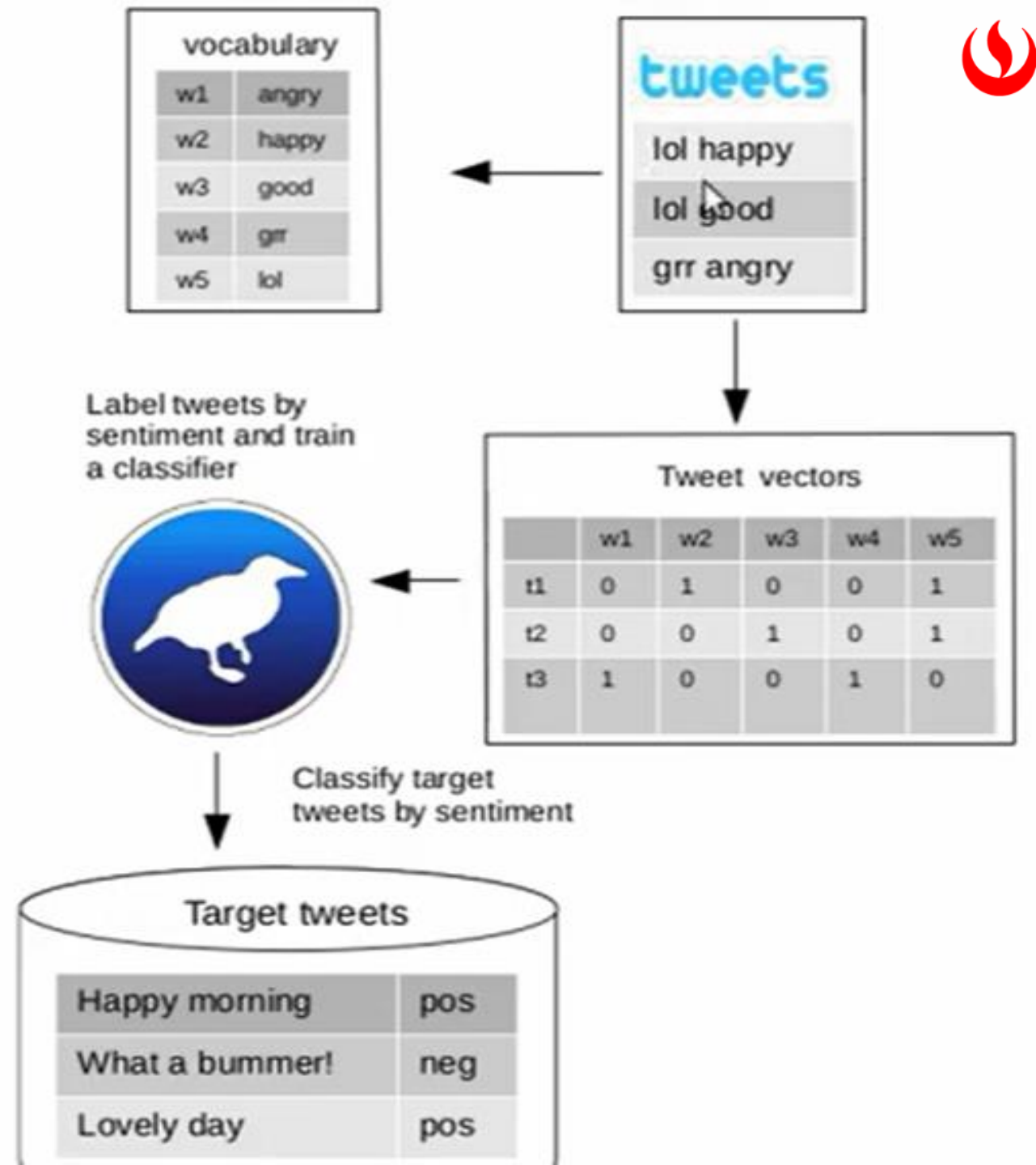
Contenido 3

Modelo de espacio vectorial y recuperación de información

- Recuperación de Información
- Tokens y Tipos
- Extracción de Vocabulario
- **Modelo de Espacio Vectorial**
- Frecuencia de Documento Invertida
- Similitudes entre vectores

- Modelo Vectorial

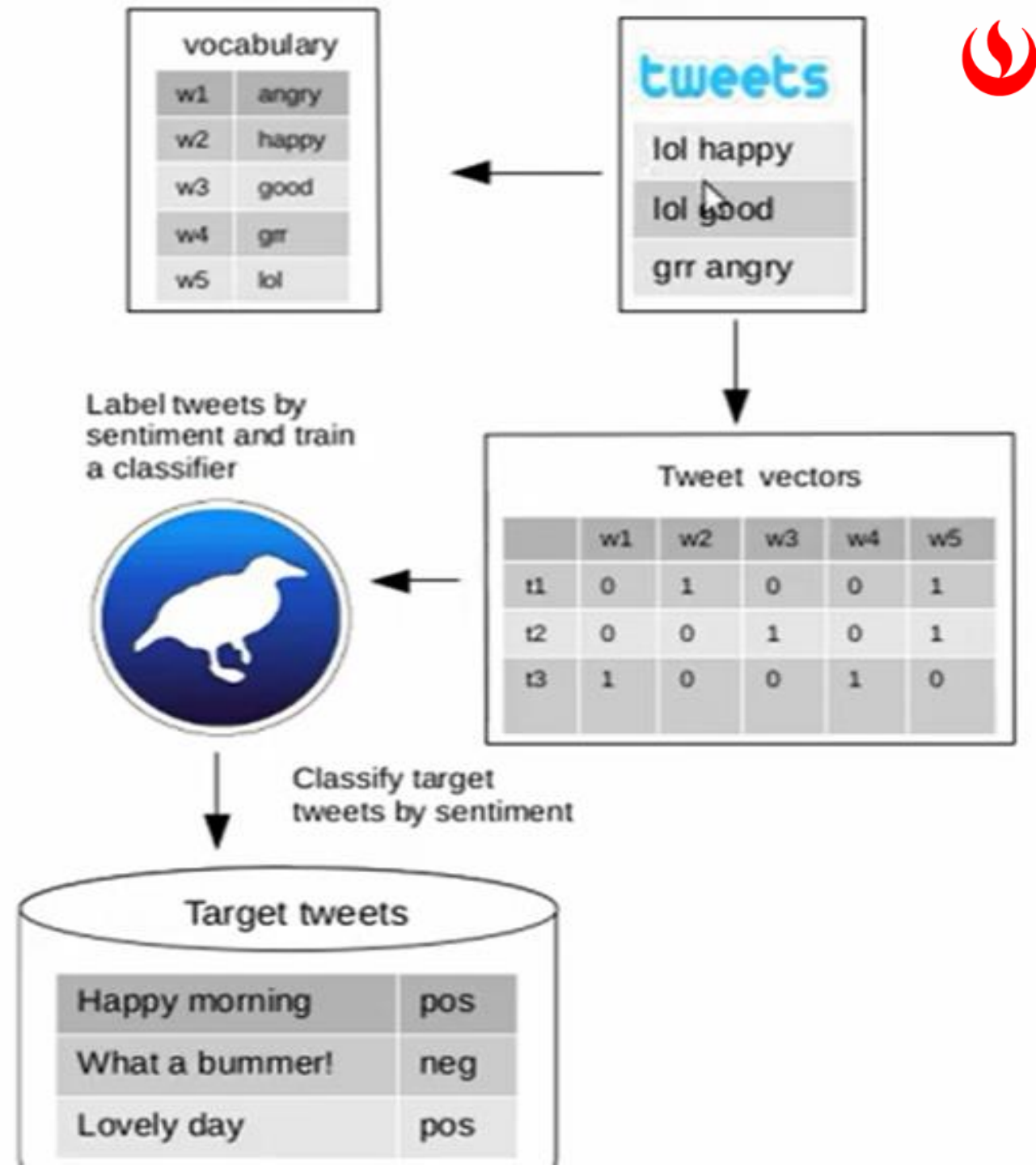
1. Modelo de Espacio Vectorial



- Modelo Vectorial

1. Modelo de Espacio Vectorial

CORPUS

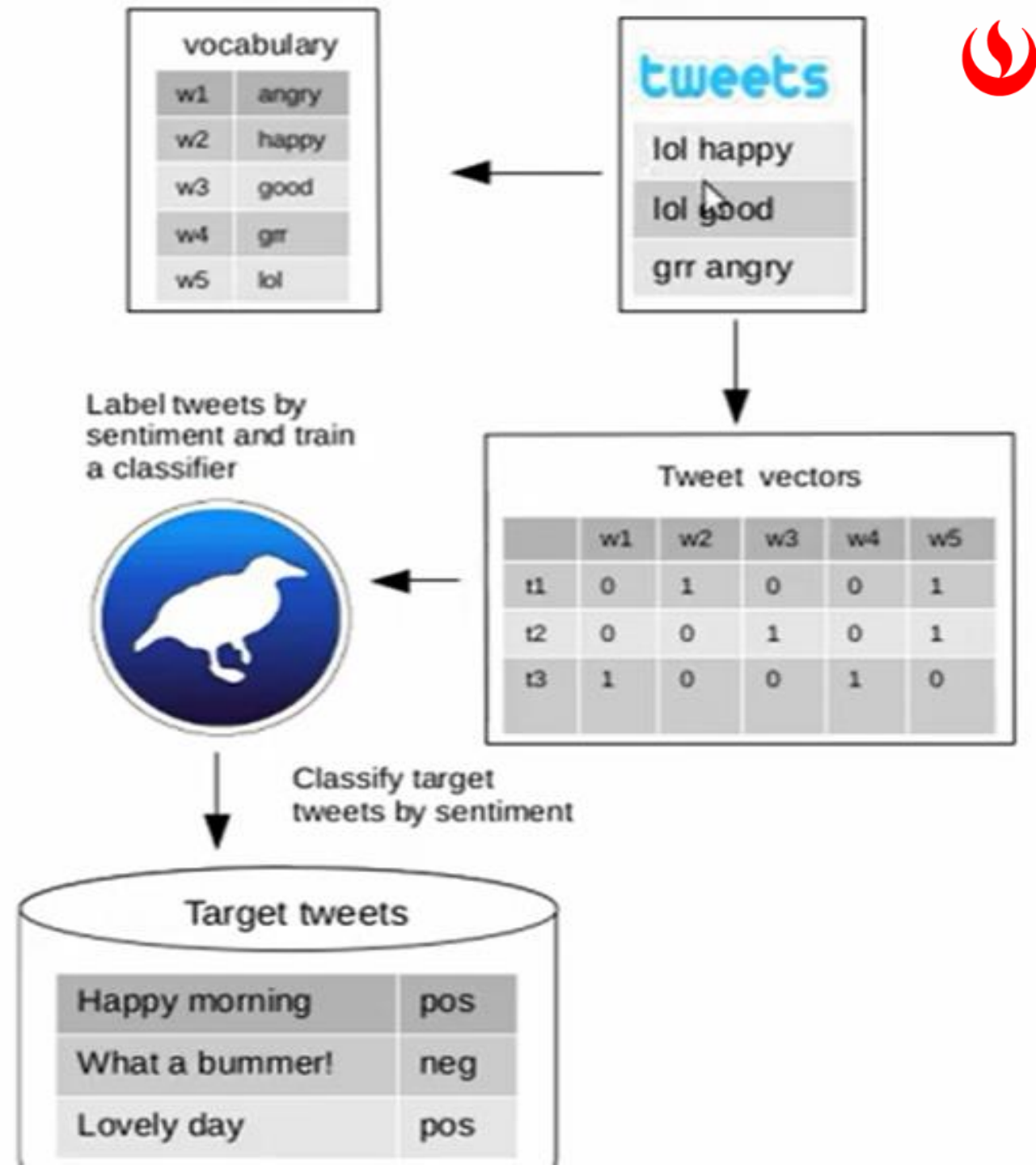


- Modelo Vectorial

1. Modelo de Espacio Vectorial

CORPUS

Conjunto de documentos.



- Modelo Vectorial

1. Modelo de Espacio Vectorial

CORPUS

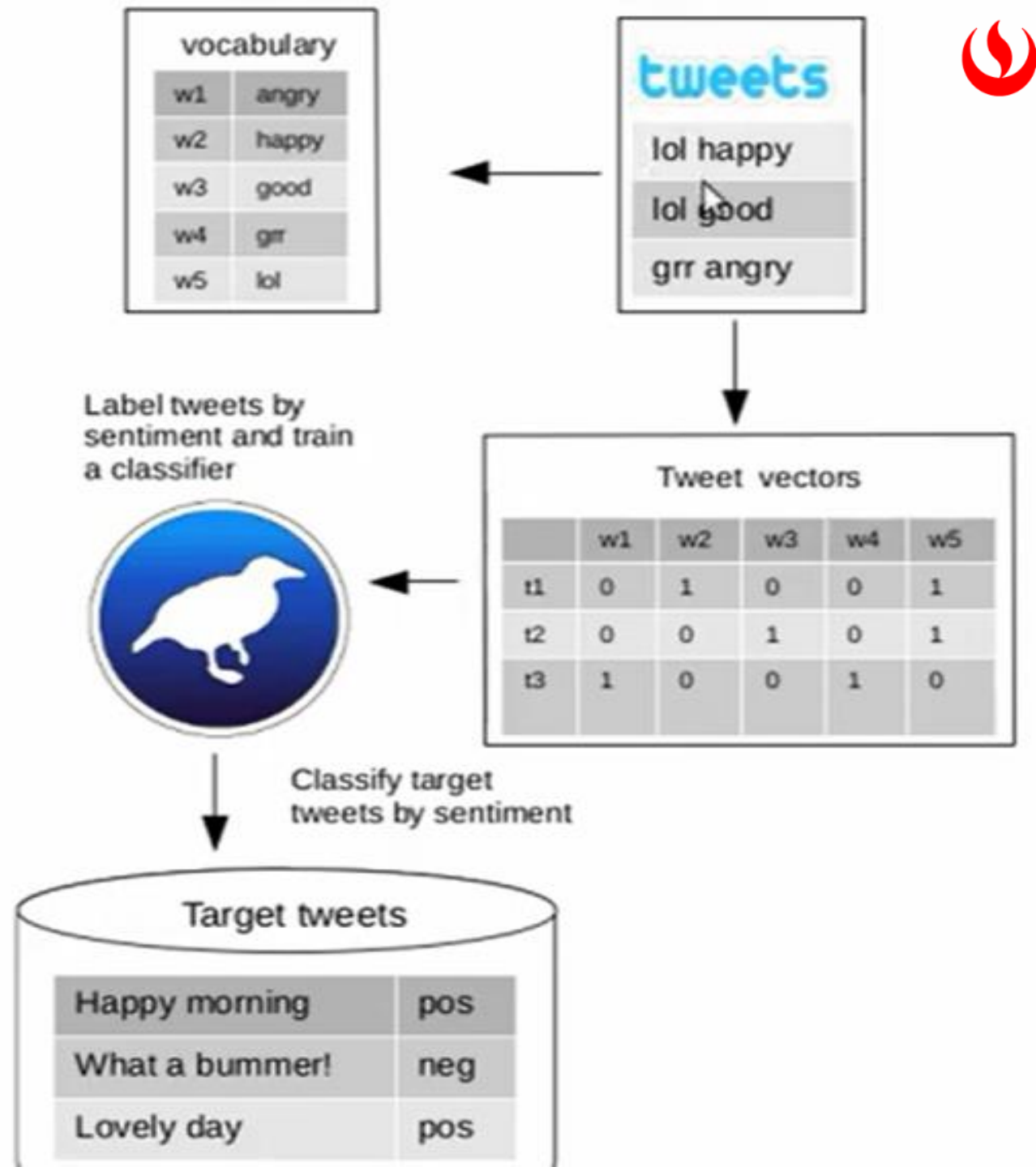
Conjunto de documentos.

$d_1 = \text{lol happy}$

$d_2 = \text{lol good}$

$d_3 = \text{grr angry}$

$\mathbf{D} = \{d_1, d_2, d_3\}$



- Modelo Vectorial

1. Modelo de Espacio Vectorial

CORPUS

Conjunto de documentos.

$d_1 = \text{lol happy}$

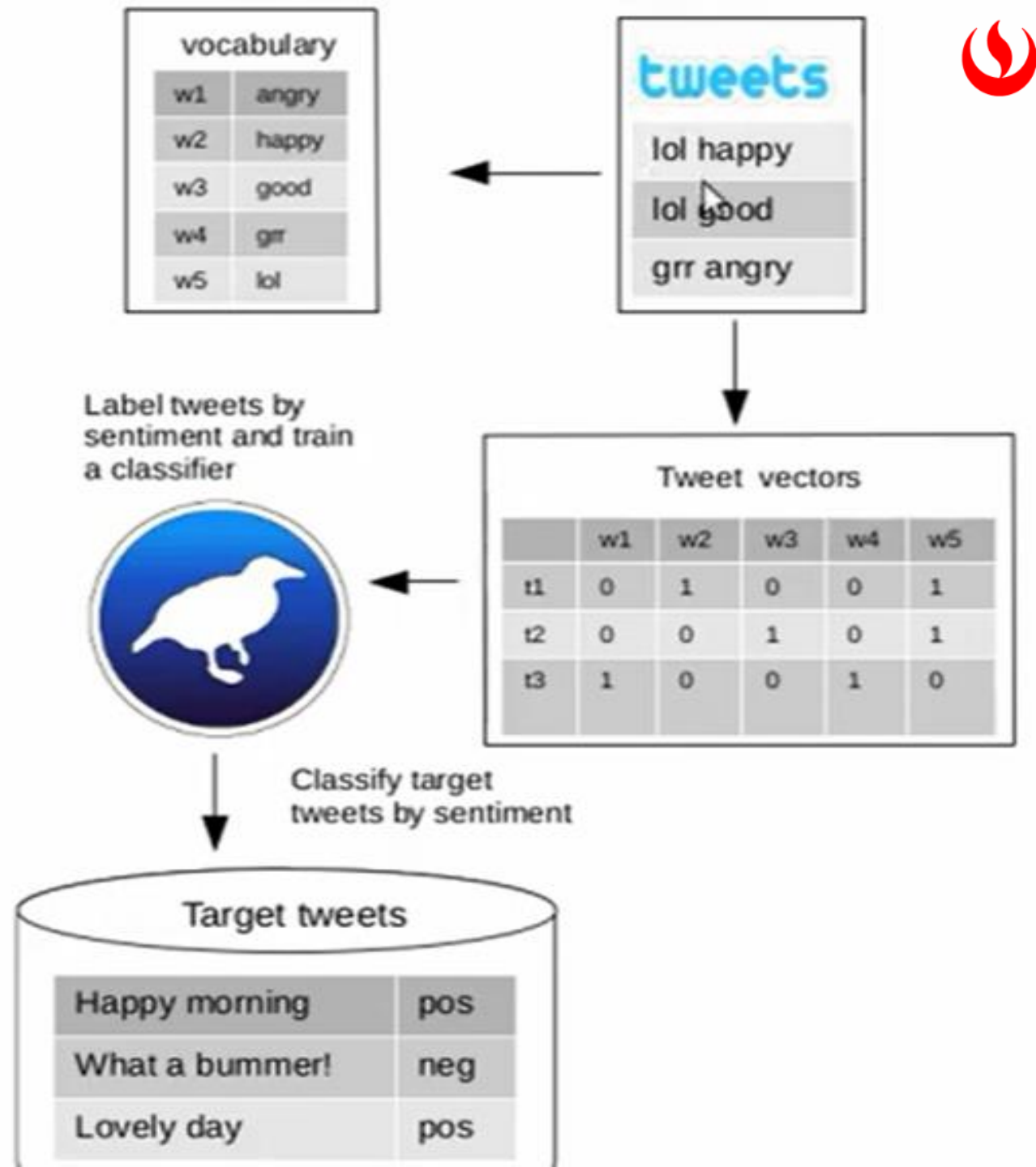
$d_2 = \text{lol good}$

$d_3 = \text{grr angry}$

$\mathbf{D} = \{d_1, d_2, d_3\}$

Vocabulario?

Notación?





1. Modelo de Espacio Vectorial

- Permite calcular **similitudes** o **distancias** entre documentos.
- Ejemplo: Si queremos ranquear las consultas que le hacen a un buscador, necesitamos medir las similitudes entre dos documentos.
- Se propone representar los documentos como vectores de términos donde cada termino va ser la dimensión del vector **[Salton et al., 1975]**
- Si tenemos documentos con longitudes de palabras distintas van a recibir el mismo espacio vectorial.



1. Modelo de Espacio Vectorial

- Ejemplo:

d1	perro gato
d2	gato casa elefante

V	
t1	perro
t2	gato
t3	casa
t4	elefante

	perro	gato	casa	elefante
d1	1	1	0	0
d2	0	1	1	1



1. Modelo de Espacio Vectorial

- Ejemplo:

d1	perro gato
d2	gato casa elefante

V	
t1	perro
t2	gato
t3	casa
t4	elefante

	perro	gato	casa	elefante
d1	1	1	0	0
d2	0	1	1	1



Estos dos documentos reciben el mismo espacio vectorial permitiéndome realizar operaciones vectoriales o compararlos en el espacio.



1. Modelo de Espacio Vectorial

- Ejemplo:

d1	perro gato
d2	gato casa elefante

V	
t1	perro
t2	gato
t3	casa
t4	elefante

¿Qué ocurre en el mundo real?

	perro	gato	casa	elefante
d1	1	1	0	0
d2	0	1	1	1



Estos dos documentos reciben el mismo espacio vectorial permitiéndome realizar operaciones vectoriales o compararlos en el espacio.



1. Modelo de Espacio Vectorial

- Ejemplo:

d1	perro gato
d2	gato casa elefante

V	
t1	perro
t2	gato
t3	casa
t4	elefante

Los vocabularios son cientos de miles. Entonces generalmente los vectores **SPARSE**.

¿Qué ocurre en el mundo real?

	perro	gato	casa	elefante
d1	1	1	0	0
d2	0	1	1	1



Estos dos documentos reciben el mismo espacio vectorial permitiéndome realizar operaciones vectoriales o compararlos en el espacio.



1. Modelo de Espacio Vectorial

- Este tipo de representaciones de se llaman **BOLSA DE PALABRAS**.
- ¿Por qué es llamado así?
- En este modelo se pierde la estructura lingüística o lo que se llama el orden de palabras de la oración. Eso lógicamente puede hacer mucho daño a la resolución del problema.

Por ejemplo:

- New York
- No quiero (análisis de sentimiento)
- Es un modelo muy simple que generalmente sirve para recuperación de información o *text mining* (minería de textos).



1. Modelo de Espacio Vectorial

¿Qué representa el valor dentro de cada cuadro/dimensión?



1. Modelo de Espacio Vectorial

- El valor de cada dimensión es el peso que representa la relevancia del termino t_i en el documento d .
- Tratamos de modelar que tan informativo es el termino para el documento.
- En nuestro ejemplo: pusimos valor booleano pero se pueden hacer cosas mejores.



1. Modelo de Espacio Vectorial

¿Cómo podemos modelar cuan informativo es un termino en un documento?



Contenido 3

Modelo de espacio vectorial y recuperación de información

- Recuperación de Información
- Tokens y Tipos
- Extracción de Vocabulario
- Modelo de Espacio Vectorial
- **Frecuencia de Documento Invertida**
- Similitudes entre vectores



2. Frecuencia de Documento Invertida

- Para ir mas haya de colocar dimensiones booleanas es considerar la **FRECUENCIA** (las veces que aparece en el termino en el texto).
- Dado $tf_{i,j}$ representa la frecuencia del termino t_i en el documento d_j .
- ¿De que sirve saber la frecuencia?
- Un término que aparece 10 veces en el documento proporciona más información que un termino que aparece una sola vez.
- Ejemplo:
Si aparece el termino “Barack Obama” aparece 50 veces en un documento debería recibir más peso que si ese mismo término aparece 2 veces en otro documento.



2. Frecuencia de Documento Invertida

d1	perro gato gato
d2	gato casa elefante

V	
t1	perro
t2	gato
t3	casa
t4	elefante

	perro	gato	casa	elefante
d1	1	2	0	0
d2	0	1	1	1



2. Frecuencia de Documento Invertida

- ¿Qué pasa cuando se tiene documento mas grandes que otros?
- Ejemplo: Mi corpus esta forado por documento: el primero un texto de la bibliografía de Barack Obama y el segundo un tweet "*I love Barack Obama*".
- Se puede normalizar por el término de frecuencia máxima en el documento.

$$ntf_{i,j} = \frac{tf_{i,j}}{\max_i(tf_{i,j})}$$



2. Frecuencia de Documento Invertida

d1	perro gato gato
d2	gato casa elefante

V	
t1	perro
t2	gato
t3	casa
t4	elefante

	perro	gato	casa	elefante
d1	0,5	1	0	0
d2	0	1	1	1



2. Frecuencia de Documento Invertida

- **¿Un término que aparece pocas veces en el documento proporciona más o menos información que un término que aparece más veces?**

Ejemplo: El respetado alcalde de **Lima**. El alcalde realiza labores necesarias.

- El término Lima ocurre menos veces que alcalde pero es más descriptivo.
- Si alguien hace la consulta a Lima este documento debería tener un peso más importante.



2. Frecuencia de Documento Invertida

- Entre más escaso un término en general (CORPUS) más importancia le dan.
- Definimos **N** como el tamaño del corpus (número de documentos) y **n_i** el número de documentos que contienen el termino **t_i**, entonces definimos **idf** (frecuencia del documento invertida) como:

$$idf_{t_i} = \log_{10}\left(\frac{N}{n_i}\right)$$

- Un término que aparece en todos los documentos debería tener $idf = 0$.



2. Frecuencia de Documento Invertida

- ¿Cómo podemos colocar el peso (relevancia) del término?
- Es representado por el **modelo de pesos tf-idf**:

$$w(t_i, d_j) = tf_i \times \log_{10}\left(\frac{N}{n_i}\right)$$

- Que tan frecuente es el término en el documento y si eso afecta positiva o negativamente.



2. Frecuencia de Documento Invertida

d1	perro gato gato
d2	gato casa elefante

V	
t1	perro
t2	gato
t3	casa
t4	elefante

	perro	gato	casa	elefante
d1	0,5	1	0	0
d2	0	1	1	1

idf (t1) =	0,30103
idf (t2) =	0
idf (t3) =	0,30103
idf (t4) =	0,30103



1. Modelo de Espacio Vectorial

Google



Turismo en Perú





1. Modelo de Espacio Vectorial

d_1

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.

d_2

El turismo en el Perú comprende vestigios maravillosos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.

- Modelo Vectorial



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

- Modelo Vectorial



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú



1. Modelo de Espacio Vectorial

d_0

Turismo en Perú

d_1

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.

d_2

El turismo en el Perú comprende vestigios maravillos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1		
d_2		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		

d₁

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		

d₁

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

	Turismo	Perú
	t ₁	t ₂
d ₁	1	
d ₂		

d₁

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	1	0
d_2		

d_1

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	1	0
d_2		

d_2

El turismo en el Perú comprende vestigios maravillosos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	1	0
d_2	1	

d_2

El turismo en el Perú comprende vestigios maravillosos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	1	0
d_2	1	1

d_2

El turismo en el Perú comprende vestigios maravillosos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.




1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	1	0
d_2	1	1





1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1		
d_2		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

$tf_{i,j}$	Turismo	Perú
	t_1	t_2
d_1		
d_2		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		

d₁

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		

d₁

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	
d ₂		

d₁

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂		

d₁

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

$tf_{i,j}$	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2		

d_2

El turismo en el Perú comprende vestigios maravillosos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂		

d₂

El turismo en el Perú comprende vestigios maravillosos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	

d₂

El turismo en el Perú comprende vestigios maravillosos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1

d₂

El turismo en el Perú comprende vestigios maravillosos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

$tf_{i,j}$	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1





1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}\left(\frac{N}{n_i}\right)$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t1}

idf_{t2}

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}\left(\frac{N}{n_i}\right)$$

$$idf_{t_1} = \boxed{= \text{LOG10}(2/2)}$$

$$idf_{t_2} = \boxed{= \text{LOG10}(2/1)}$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

	Turismo	Perú
	t ₁	t ₂
d ₁		
d ₂		

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1

Frecuencia de Documento Invertida

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

	Turismo	Perú
	t ₁	t ₂
d ₁	=4*0	
d ₂		

tf_{i,j}

	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁	0	
d ₂		

tf_{i,j}

	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

	Turismo	Perú
	t ₁	t ₂
d ₁	0	
d ₂		

tf_{i,j}

	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁	0	=0*0,30103
d ₂		

tf_{i,j}

	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

	Turismo	Perú
	t ₁	t ₂
d ₁	0	0
d ₂		

tf_{i,j}

	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1

Frecuencia de Documento Invertida

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁	0	0
d ₂		

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t1} = 0

idf_{t2} = 0,30103

	Turismo	Perú
	t ₁	t ₂
d ₁	0	0
d ₂	=2*0	

tf_{i,j}

	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁	0	0
d ₂	0	

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t1} = 0

idf_{t2} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁	0	0
d ₂	0	

	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁	0	0
d ₂	0	=1*0,30103

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$

idf_{t₁} = 0

idf_{t₂} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁	0	0
d ₂	0	0,30103

<i>tf_{i,j}</i>	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1



1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t ₁	Turismo
t ₂	Perú

$$idf_{t_i} = \log_{10}(\frac{N}{n_i})$$


idf_{t₁} = 0

idf_{t₂} = 0,30103

$$w(t_i, d_j) = tf_i \times \log_{10}(\frac{N}{n_i})$$

Frecuencia de Documento Invertida

	Turismo	Perú
	t ₁	t ₂
d ₁	0	0
d ₂	0	0,30103



	Turismo	Perú
	t ₁	t ₂
d ₁	4	0
d ₂	2	1



1. Modelo de Espacio Vectorial

- Permite calcular **similitudes** o **distancias** entre documentos.
- Ejemplo: Si queremos ranquear las consultas que le hacen a un buscador, necesitamos medir las similitudes entre dos documentos.
- Se propone representar los documentos como vectores de términos donde cada término va a ser la dimensión del vector **[Salton et al., 1975]**
- Si tenemos documentos con longitudes de palabras distintas van a recibir el mismo espacio vectorial.



Contenido 3

Modelo de espacio vectorial y recuperación de información

- Recuperación de Información
- Tokens y Tipos
- Extracción de Vocabulario
- Modelo de Espacio Vectorial
- Frecuencia de Documento Invertida
- **Similitudes entre vectores**



3. Similitudes entre vectores

- ¿Cómo podemos comparar los documentos?
- ¿Cómo podemos ranquear los documentos de una búsqueda?



3. Similitudes entre vectores

d_0

Turismo en Perú

d_1

El turismo es un fenomeno social y cultural. El turismo supone asocia muchos factores. El turismo comprende las acividades que realizan esas personas. El turismo es bueno.

d_2

El turismo en el Perú comprende vestigios maravillos que sorprenden al mundo como Machu Picchu. El turismo nacional es muy popular en el mundo.



3. Similitudes entre vectores

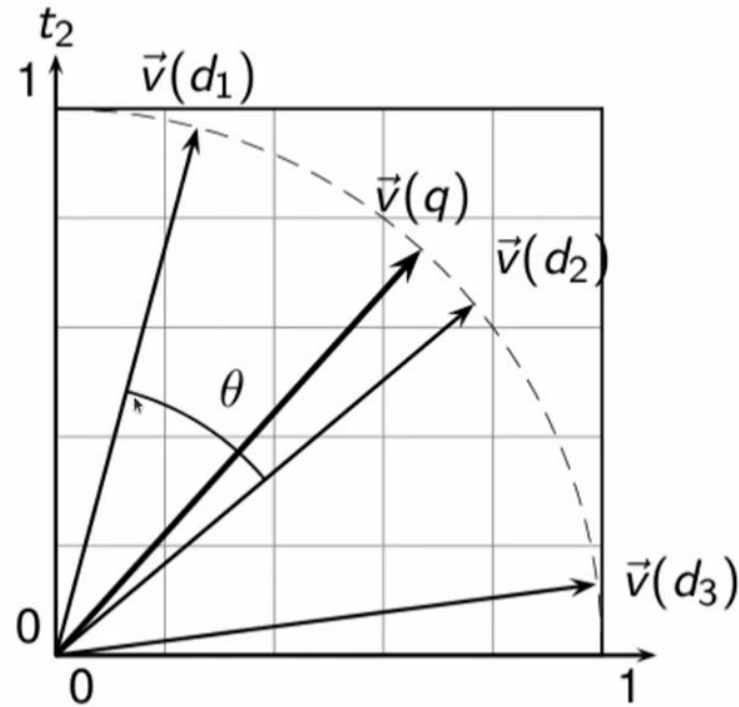
- ¿Cómo podemos comparar los documentos?
- ¿Cómo podemos ranquear los documentos de una búsqueda?
- Convertir la consulta a documento y compararlos con los demás documentos
- Podemos usar la distancia Euclidiana

d1	0,5	1	0	0
d2	0	1	1	1



2. Frecuencia de Documento Invertida

Similitud de Cosenos



$$\cos(\vec{d}_1, \vec{d}_2) = \frac{\vec{d}_1 \cdot \vec{d}_2}{|\vec{d}_1| \times |\vec{d}_2|} = \frac{\sum_{i=1}^{|V|} (w(t_i, d_1) \times w(t_i, d_2))}{\sqrt{\sum_{i=1}^{|V|} w(t_i, d_1)^2} \times \sqrt{\sum_{i=1}^{|V|} w(t_i, d_2)^2}}$$



CIERRE



Tokenización

Lematización

Frecuencia de
términos



CONSULTAS

pcsirife@upc.edu.pe