Modelo de espacio vectorial y recuperación de información

PhD. Carlos Fernando Montoya Cubas

- Modelo Vectorial

1. Recuperación de Información

- Modelo Vectorial

1. Recuperación de Información

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

- Modelo Vectorial

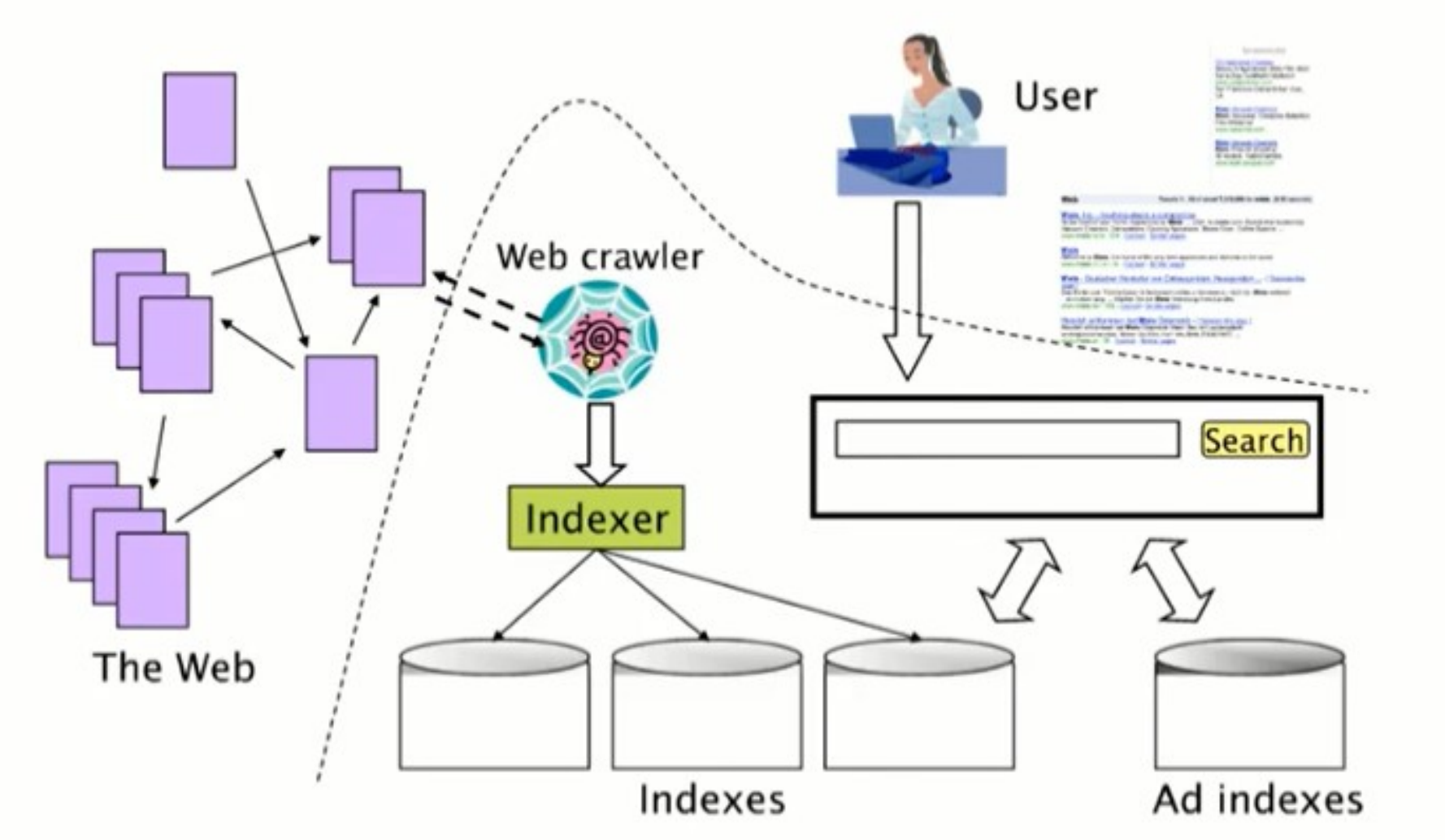
1. Recuperación de Información

- **¿**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:

- Modelo Vectorial

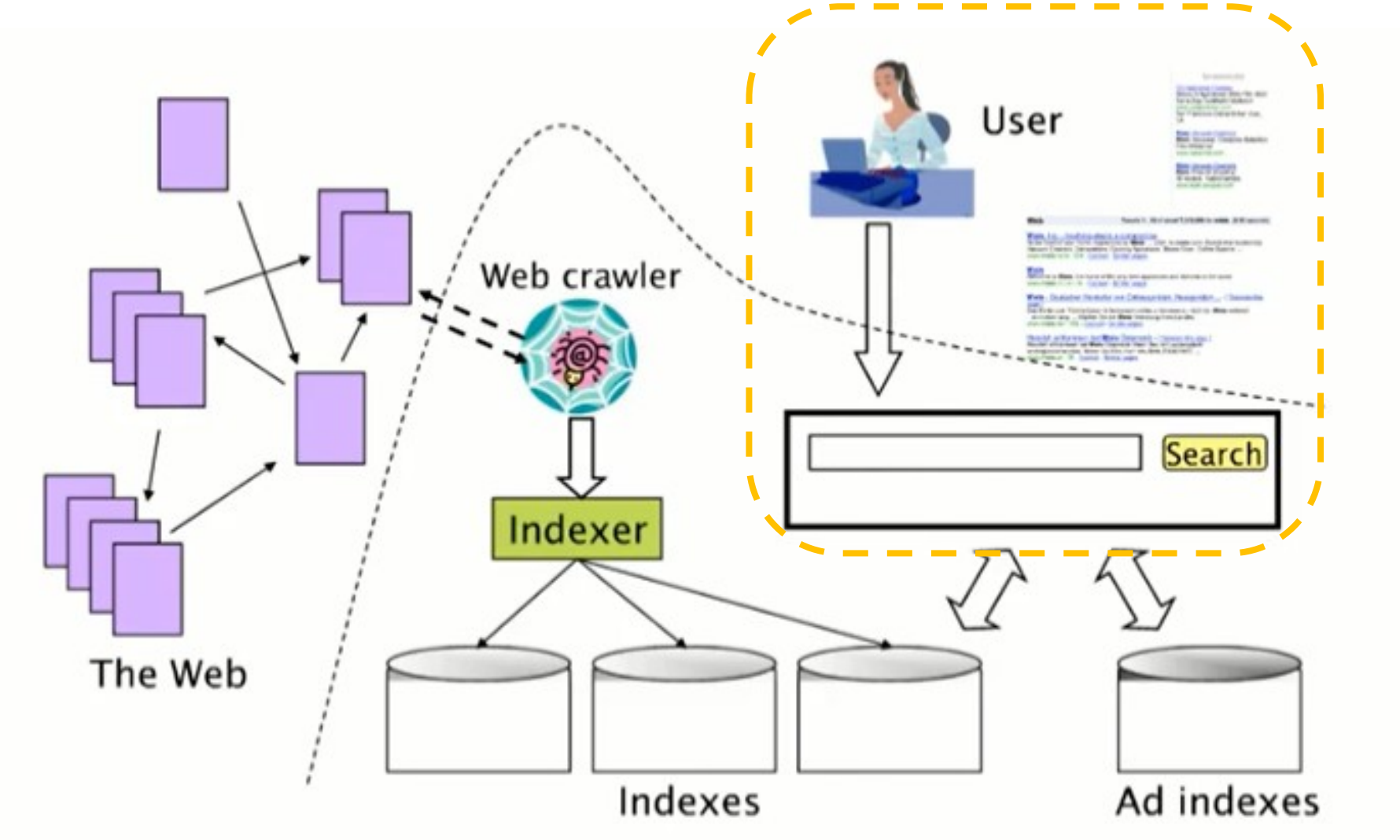
1. Recuperación de Información



- Modelo Vectorial

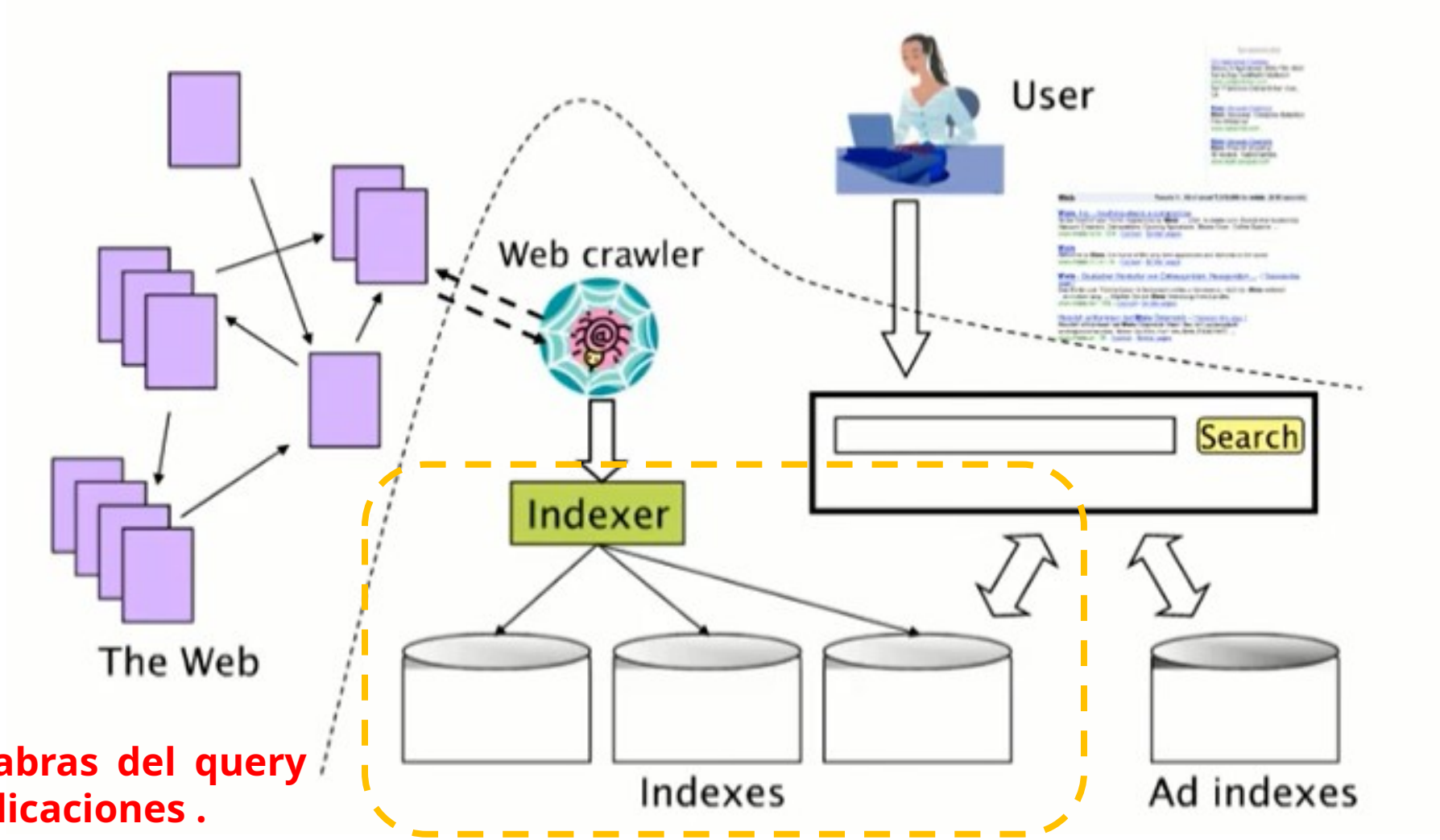
1. Recuperación de Información

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



- Modelo Vectorial

1. Recuperación de Información

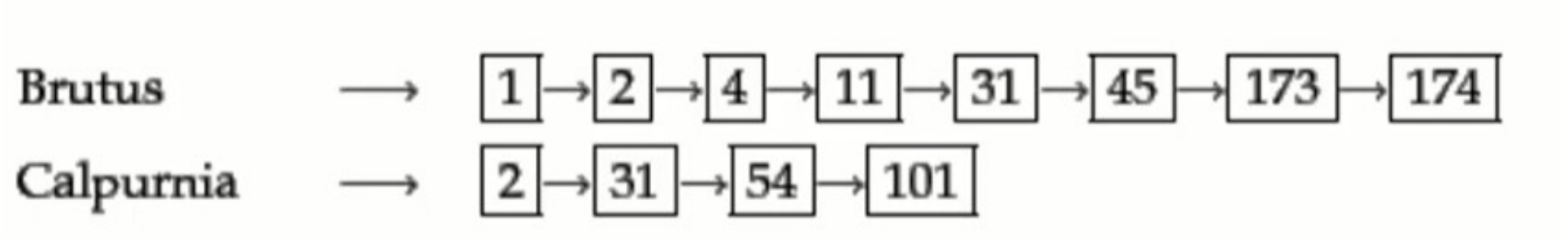


Mapear las palabras del query en listas de publicaciones .

- Modelo Vectorial

1. Recuperación de Información

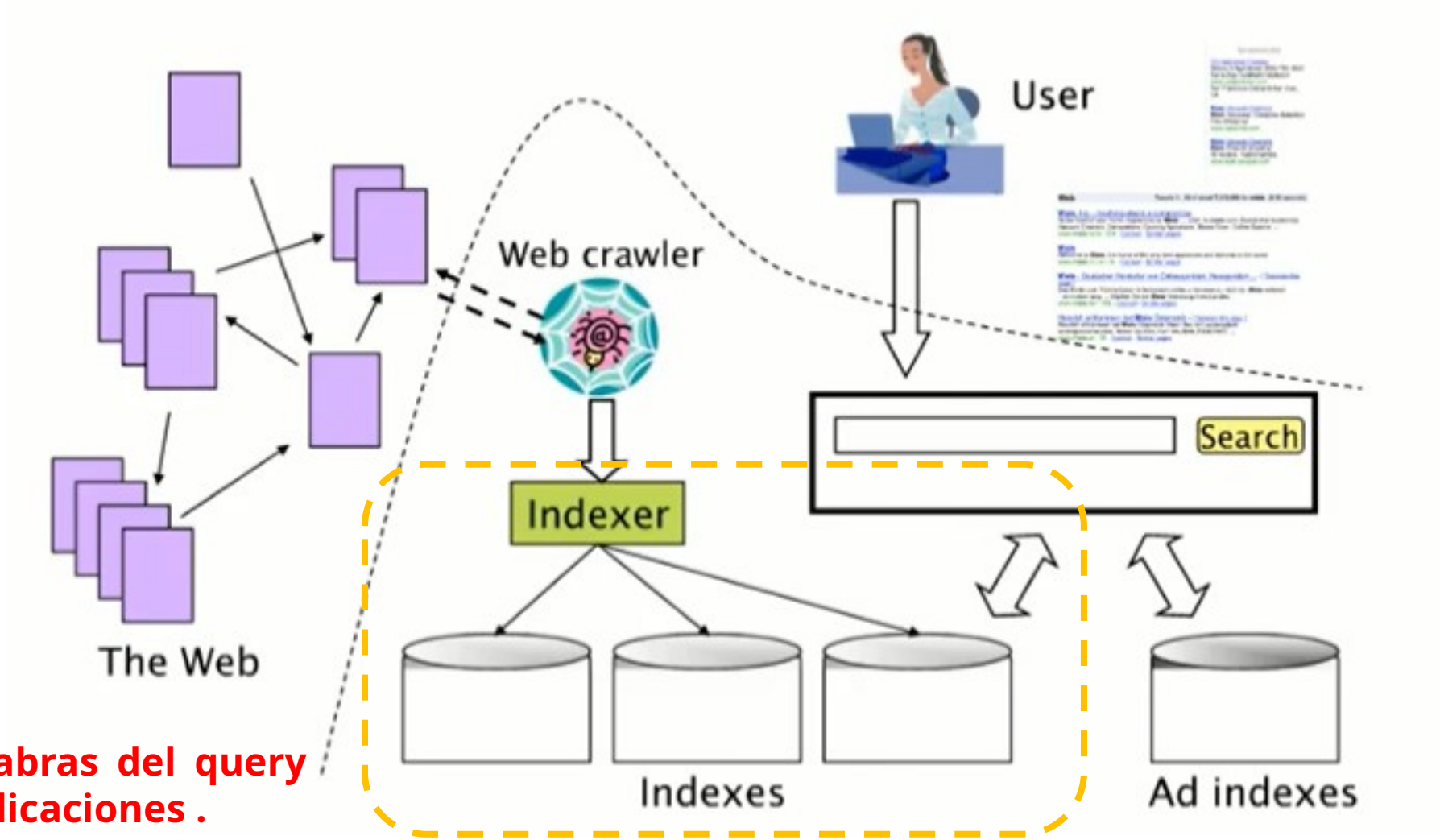
- Una lista de publicaciones de un termino es un lista de todos los documentos donde el termino aparece por lo menos una vez



Mapear las palabras del query en listas de publicaciones .

- Modelo Vectorial

1. Recuperación de Información

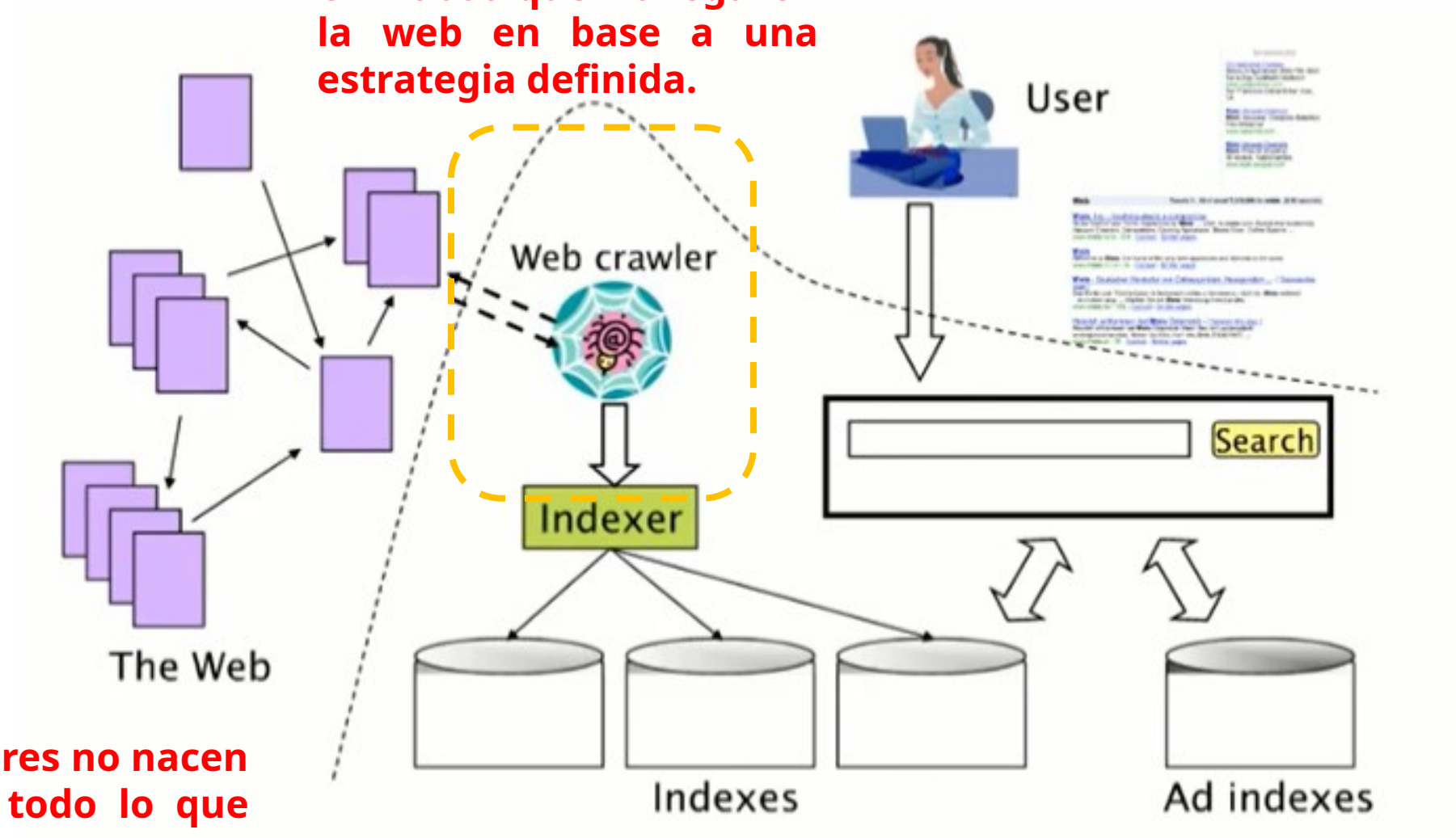


Mapear las palabras del query en listas de publicaciones .

- Modelo Vectorial

1. Recuperación de Información

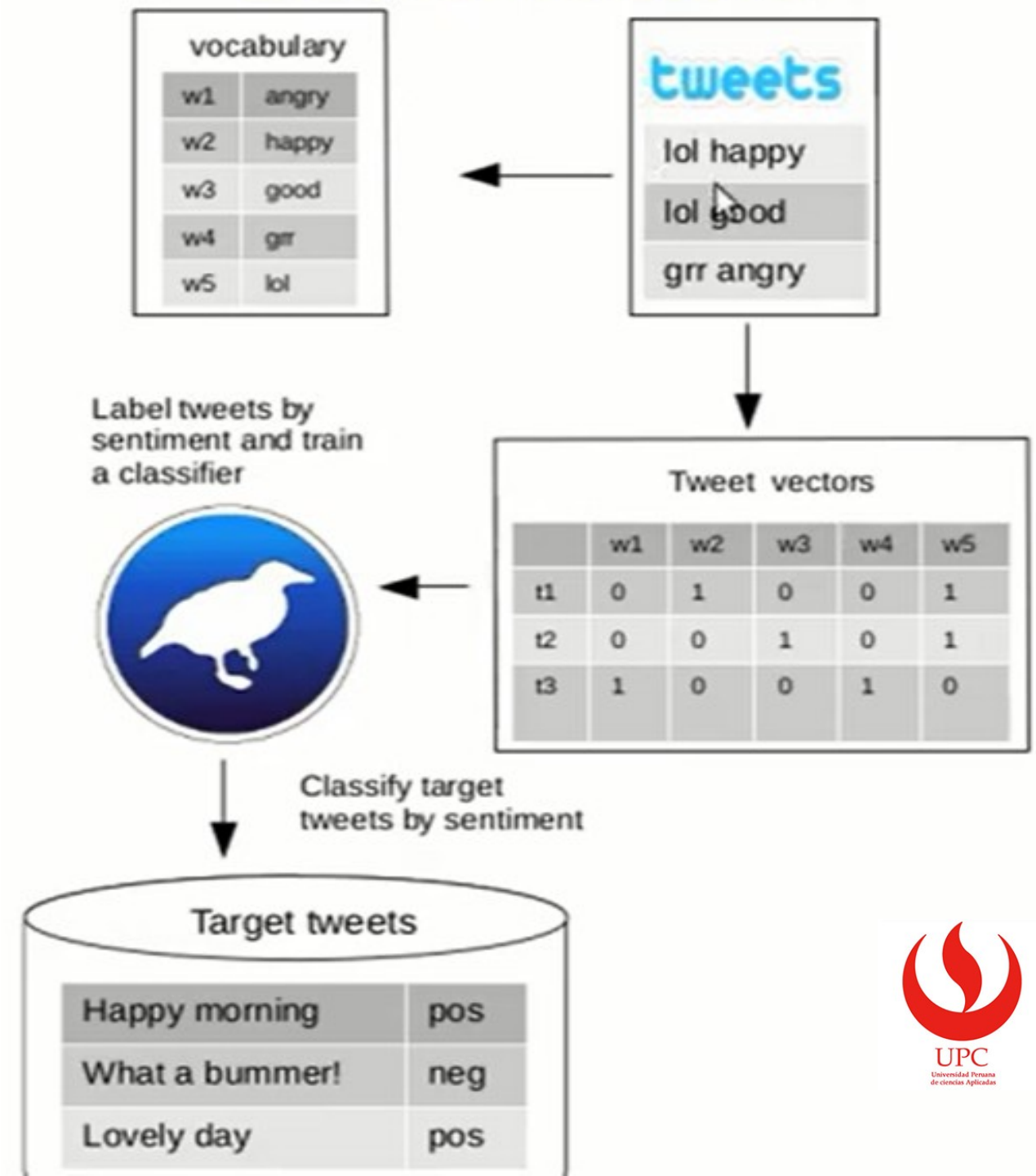
Un robot que navega en la web en base a una estrategia definida.



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

- Modelo Vectorial

2. Modelo de Espacio Vectorial



2. Modelo de Espacio Vectorial

¿Cómo pasar el texto a algo procesable algorítmicamente?

- Modelo Vectorial

2.1. Tokens y Tipos

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

- Modelo Vectorial

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]
```

- Modelo Vectorial

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.

- Modelo Vectorial

2.1. Tokens y Tipos

- Algunas librerías ya vienen tokenizadas: NLTK y Spyce.

Tokenize and tag some text:

```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged[0:6]
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'),
('Thursday', 'NNP'), ('morning', 'NN')]
```

- Modelo Vectorial

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.

- Modelo Vectorial

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]
```

- Modelo Vectorial

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]

- Modelo Vectorial

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:

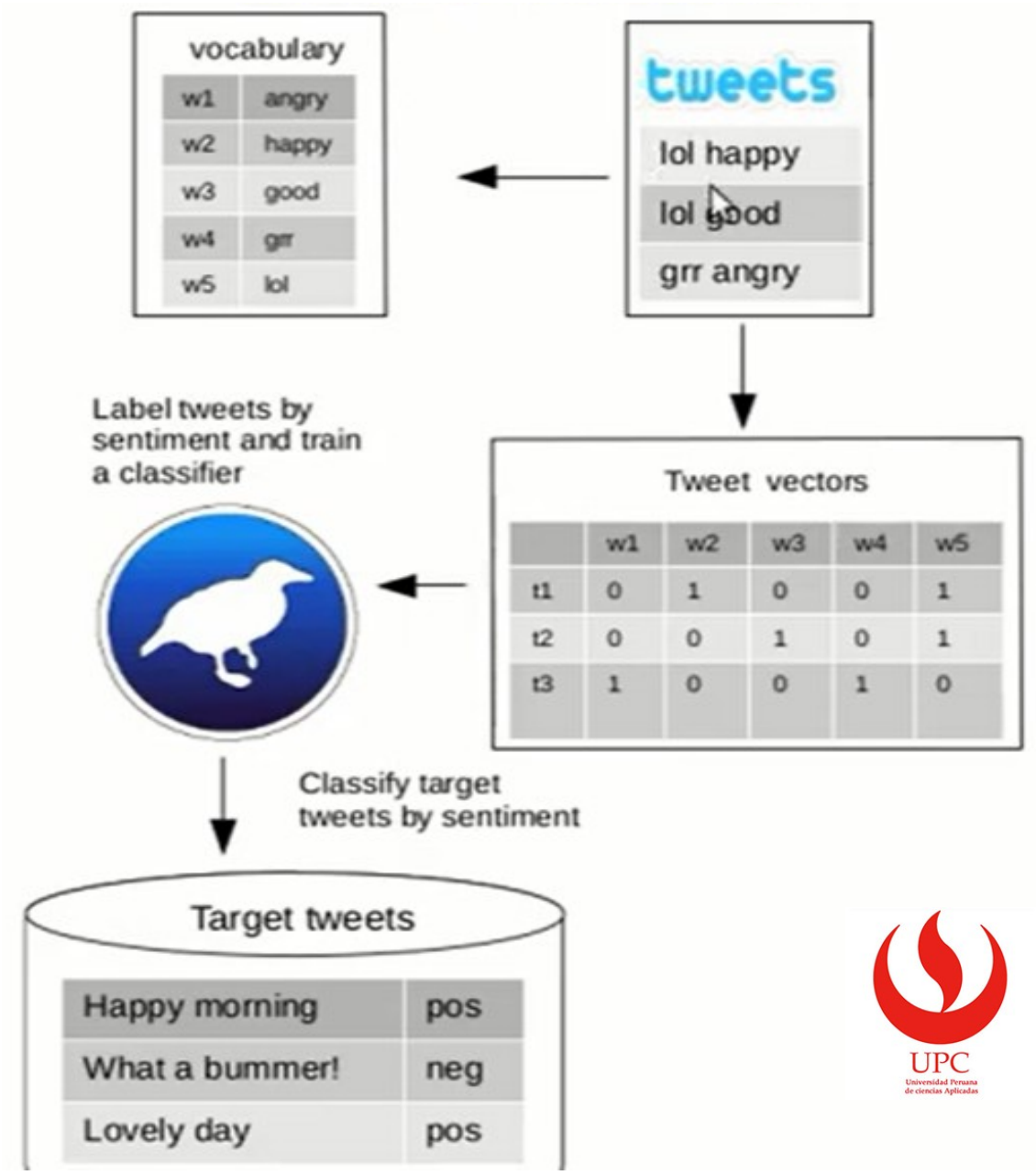
Input: I like human languages and programming languages.
Tokens: [I] [like] [human] [languages] [and] [programming] [languages]

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.

- Modelo Vectorial

2.2. Extracción de Vocabulario



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 mas 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]

- Modelo Vectorial

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			Example		
	SSES	→	SS	caresses	→	caress
	IES	→	I	ponies	→	poni
	SS	→	SS	caress	→	caress
	S	→		cats	→	cat

- Modelo Vectorial

2.2. Extracción de Vocabulario

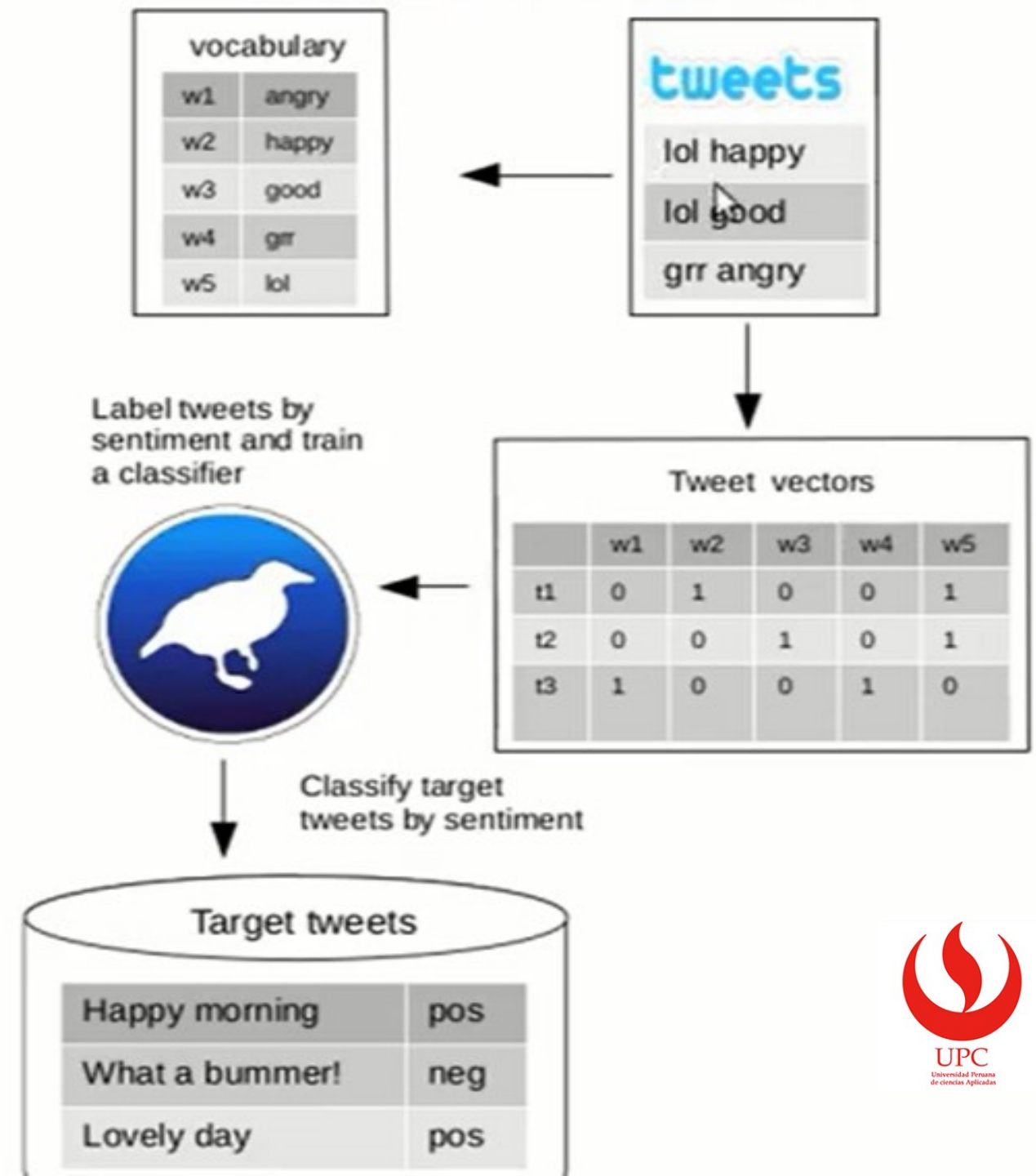
Types for the previous sentence: [I] [like] [human] [languages] [and] [programming]

termId	value
t1	human
t2	languag
t3	program

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

- Modelo Vectorial

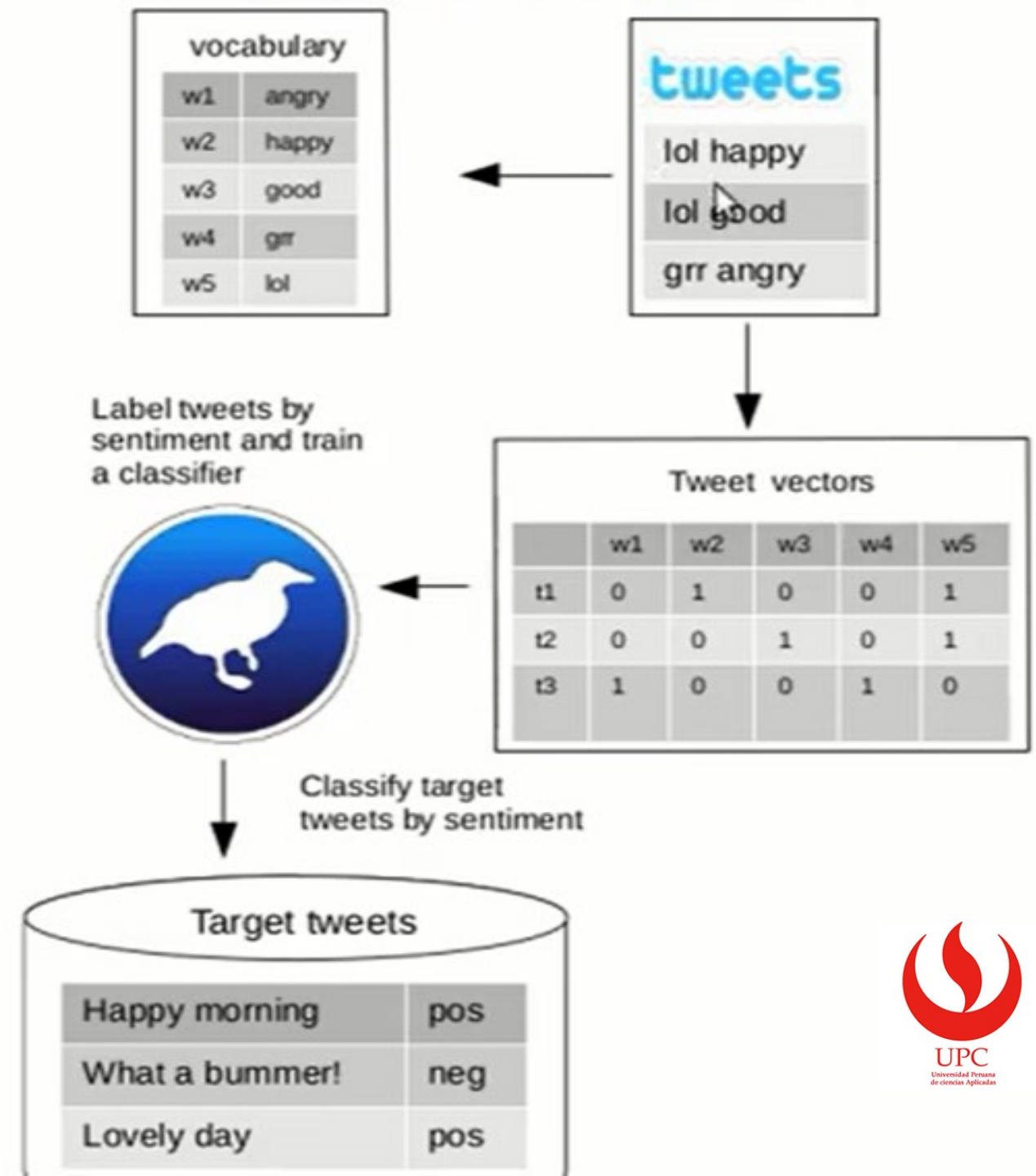
1. Modelo de Espacio Vectorial



- Modelo Vectorial

1. Modelo de Espacio Vectorial

CORPUS

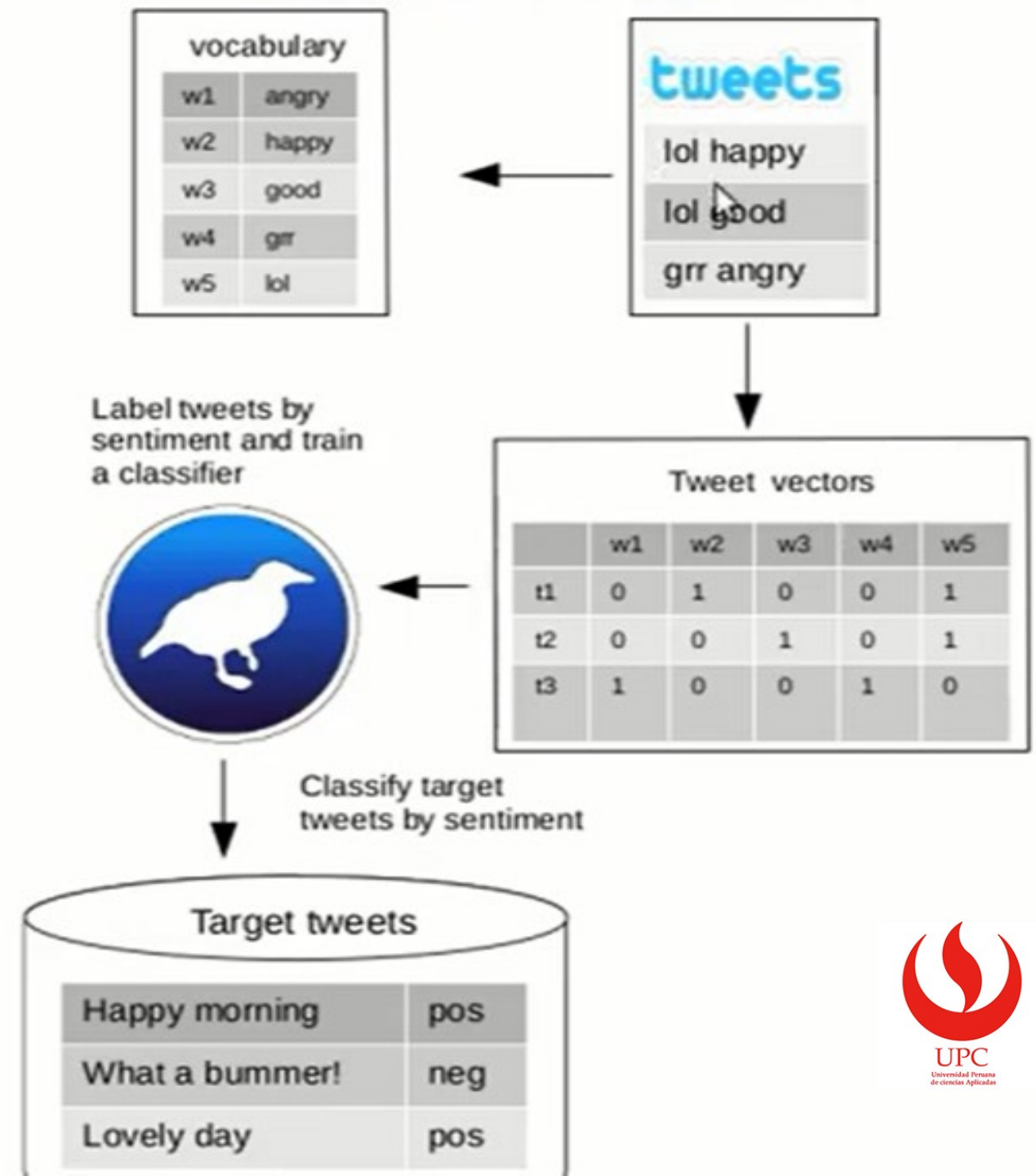


- Modelo Vectorial

1. Modelo de Espacio Vectorial

CORPUS

Conjunto de documentos.



- Modelo Vectorial

1. Modelo de Espacio Vectorial

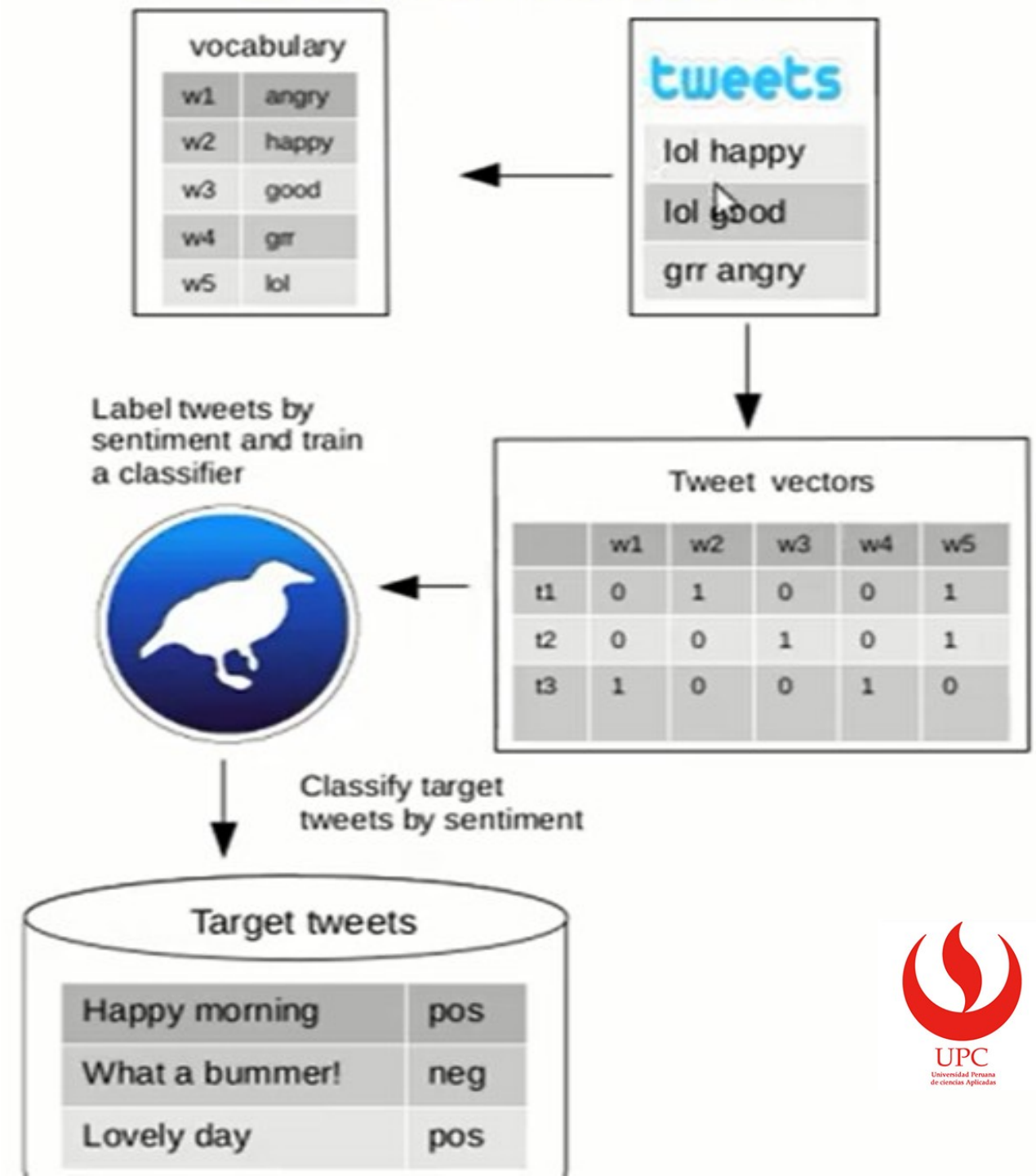
CORPUS

Conjunto de documentos.

d_1 = lol happy

d_2 = lol good

d_3 = grr angry **D** = {d_1, d_2, d_3}



- Modelo Vectorial

1. Modelo de Espacio Vectorial

CORPUS

Conjunto de documentos.

d_1 = lol happy

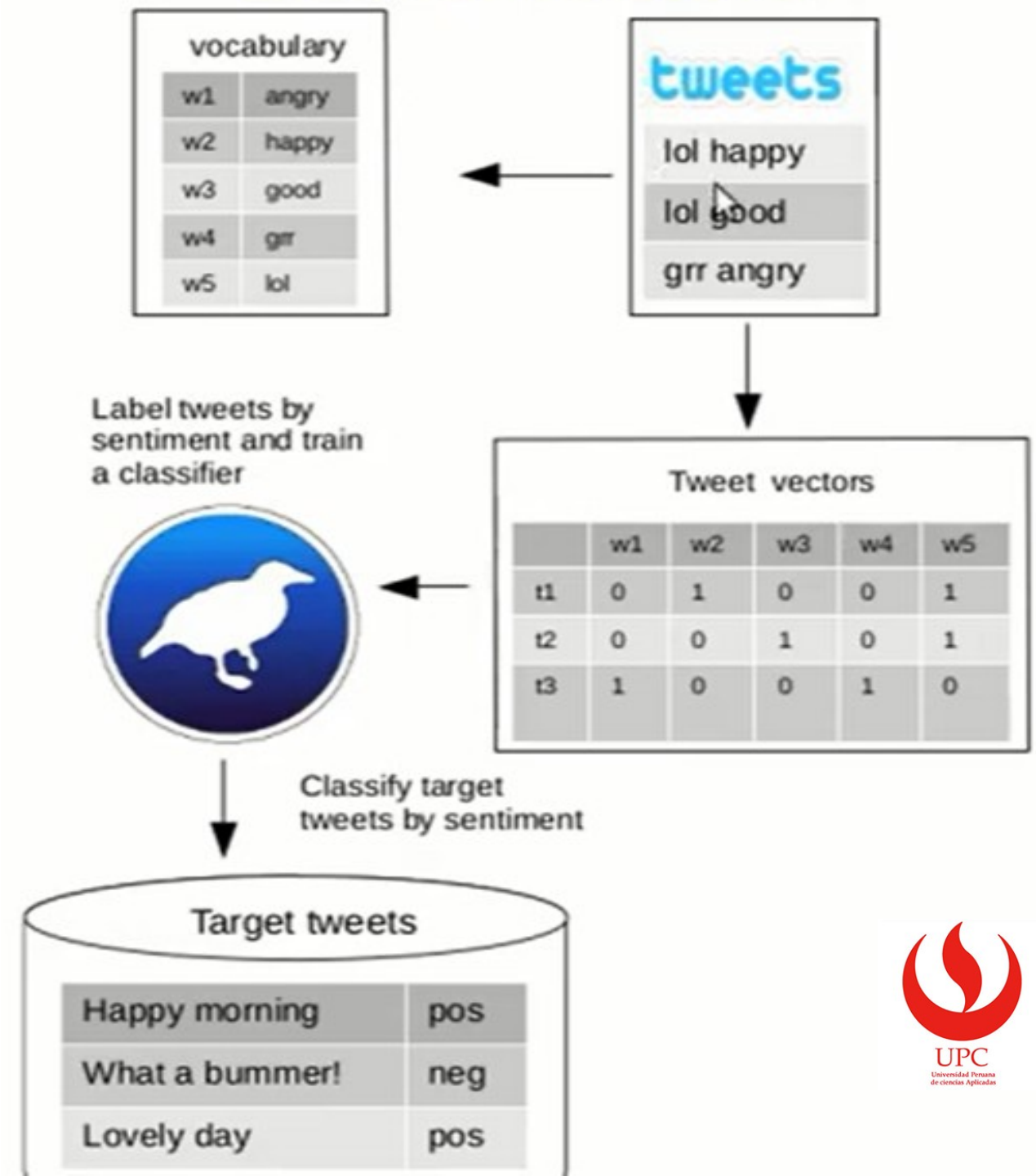
d_2 = lol good

d_3 = grr angry

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.

- Modelo 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

- Modelo 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

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

- Modelo Vectorial

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.

- Modelo Vectorial

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.

- Modelo Vectorial

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?

- Modelo Vectorial

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?

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 $\mathbf{tf}_{i,j}$ representa la frecuencia del termino $\mathbf{t}_i$ en el documento $\mathbf{d}_j$.
- ¿De que sirve saber la frecuencia?.
- Un termino que aparece 10 veces en el documento proporciona mas información que un termino que aparece una sola vez.
- Ejemplo:

Si aparece el termino “Barack Obama” aparece 50 veces en un documento debería recibir mas peso que si ese mismo termino aparece 2 veces en otro documento.

- Modelo Vectorial

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

- Modelo Vectorial

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 twit *“I love Barack Obama”.*
- Se puede normalizar por el termino de frecuencia máxima en el documento.

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

- Modelo Vectorial

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

- Modelo Vectorial

2. Frecuencia de Documento Invertida

- **¿Un termino que aparece pocas veces en el documento proporciona mas o menos información que un termino que aparece mas veces?.**

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

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

- Modelo Vectorial

2. Frecuencia de Documento Invertida

- Entre mas escaso un termino en general (CORPUS) mas importancia le dan.

- Definimos **N** como el tamaño del corpus (número de documentos) y $\mathbf{n_i}$ el número de documentos que contienen el termino $\mathbf{t_i}$, entonces definimos **idf** (frecuencia del documento invertida) como:

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

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

- Modelo Vectorial

2. Frecuencia de Documento Invertida

- ¿Cómo podemos colocar el peso (relevancia) del termino?

- Es representado por el **modelo de pesos** **tf-idf:**

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

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

- Modelo Vectorial

2. Frecuencia de Documento Invertida

d1	perro gato gato
d2	gato casa elefante

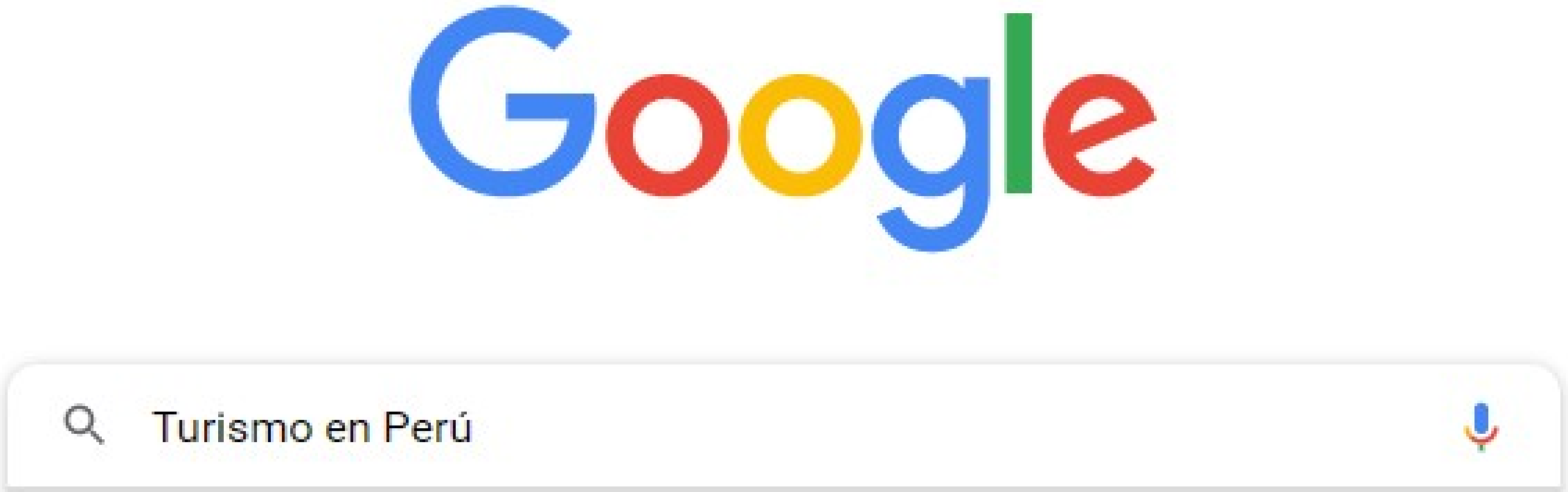
V	
t1	perro
t2	gato
t3	casa
t4	elefante

	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

- Modelo Vectorial

1. Modelo de Espacio Vectorial



- Modelo Vectorial

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 maravillos 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

- Modelo Vectorial

1. Modelo de Espacio Vectorial

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

- Modelo Vectorial

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.

- Modelo Vectorial

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		

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		
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	
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_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 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

	Turismo	Perú
	t_1	t_2
d_1	1	0
d_2	1	

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

	Turismo	Perú
	t_1	t_2
d_1	1	0
d_2	1	1

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.

- Modelo Vectorial

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		

- Modelo Vectorial

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		

- Modelo Vectorial

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		

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

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

- Modelo Vectorial

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	
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

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

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

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

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

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

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

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

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

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.

- Modelo Vectorial

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		

- Modelo Vectorial

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		

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

- Modelo Vectorial

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		

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

idf_{t1}

idf_{t2}

- Modelo Vectorial

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		

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

idf_{t1} = =LOG10(2/2)

idf_{t2} = =LOG10(2/1)

- Modelo Vectorial

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		

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

$idf_{t1} =$ 0

$idf_{t2} =$ 0,30103

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1		
d_2		

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

$idf_{t1} =$ 0

$idf_{t2} =$ 0,30103

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1		
d_2		

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

$idf_{t1} = 0$

$idf_{t2} = 0{,}30103$

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	=4*0	
d_2		

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

$idf_{t1} =$ 0

$idf_{t2} =$ 0,30103

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	0	
d_2		

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

$idf_{t1} =$ 0

$idf_{t2} =$ 0,30103

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	0	
d_2		

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

$idf_{t1} =$ 0

$idf_{t2} =$ 0,30103

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿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*0,30103
d_2		

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿Cuál es el primer paso?

Vocabulario	
t_1	Turismo
t_2	Perú

	Turismo	Perú
	t_1	t_2
d_1	0	0
d_2		

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

$idf_{t1} =$ 0

$idf_{t2} =$ 0,30103

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿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

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿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

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿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	0	

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿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	0	

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿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	0	=1*0,30103

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿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	0	0,30103

$tf_{i,j}$

	Turismo	Perú
	t_1	t_2
d_1	4	0
d_2	2	1

- Modelo Vectorial

1. Modelo de Espacio Vectorial

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

Frecuencia de Documento Invertida

¿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	0	0,30103

$tf_{i,j}$

	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 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.

3. Similitudes entre vectores

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

- Modelo Vectorial

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

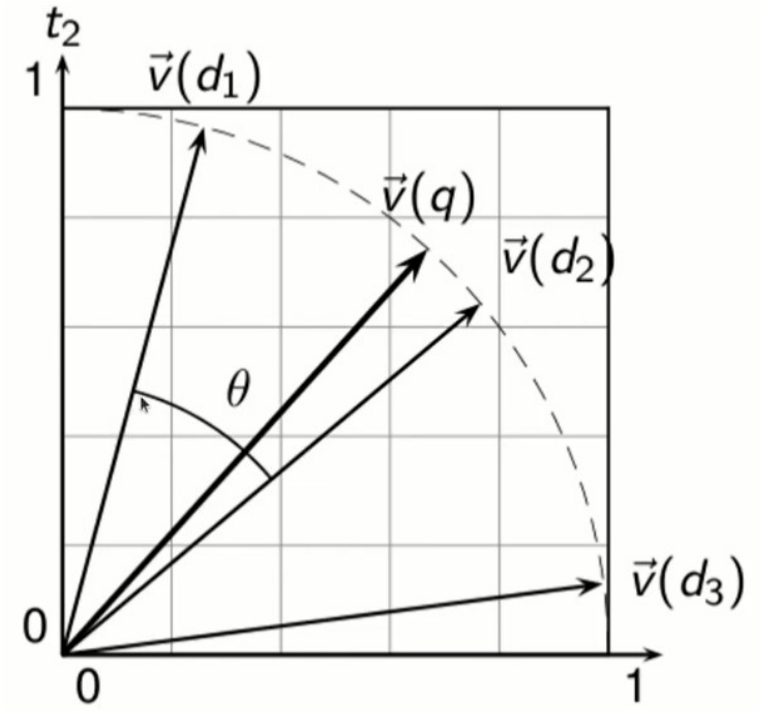
d1	0,5	1	0	0

d2	0	1	1	1

- Modelo Vectorial

2. Frecuencia de Documento Invertida

Similitud de Cosenos



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