

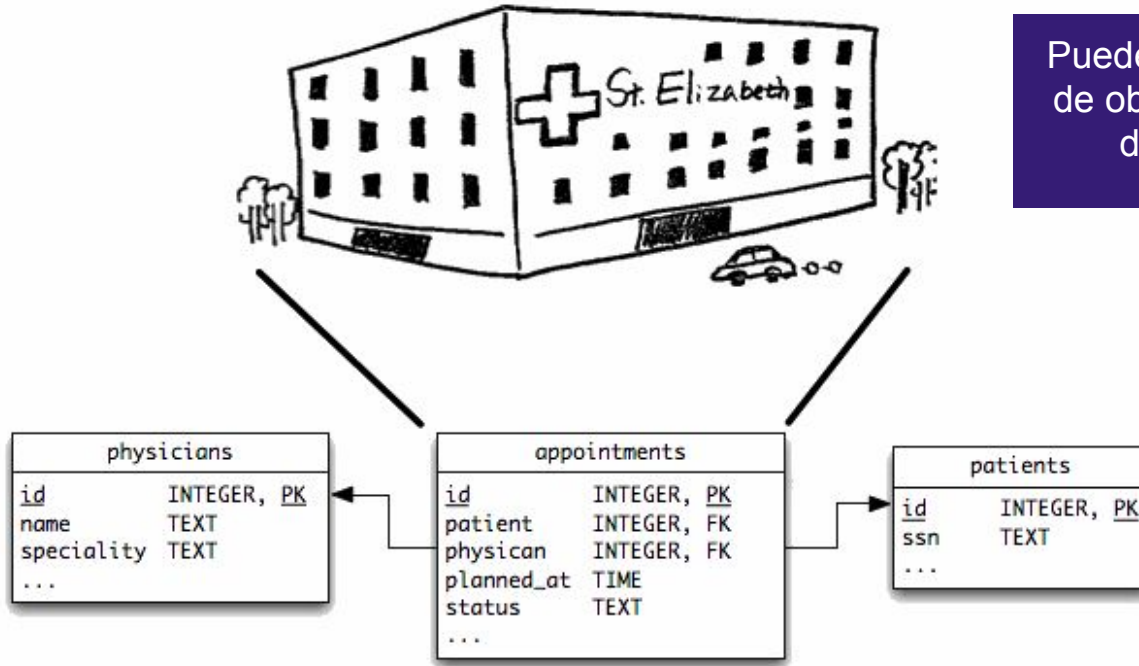
Data Mining



Datos no estructurados.
De bases de datos textuales a features

Datos Estructurados

Elementos bien definidos, relaciones entre elementos



Puede requerir mucha mano de obra para recopilar/curar datos estructurados

Image source: [Rvision-Zero](#)

Datos No Estructurados

Los datos no estructurados se refieren a datos que **no encajan perfectamente en la estructura tradicional de filas y columnas** de las bases de datos relacionales.

No hay un modelo predefinido.

Ejemplos:

- ☐ Textos
- ☐ Imágenes
- ☐ Videos
- ☐ Audios
- ☐ Mediciones numéricos

A menudo: es necesario usar datos heterogéneos para tomar decisiones.

Por supuesto, hay una estructura en estos datos, pero la estructura no está claramente explicada por nosotros. (HTML, Metadatos de multimedia, etc.)

Tenemos que extraer los elementos importantes y descubrir cómo se relacionan.

Actualmente, la mayoría de los datos que se crean **no están estructurados**, y se estima que representan **más del 95%** de todos los datos generados.

Extraer conocimiento desde textos

La extracción de información útil del texto con varios tipos de algoritmos estadísticos se conoce como **text mining**, **text analytics** o **machine learning from text**.

El análisis de texto se ha vuelto cada vez más popular en los últimos años debido a la ubicuidad de los datos de texto en la Web, las redes sociales, los correos electrónicos, las bibliotecas digitales y los sitios de chat.

- ❑ **Digital libraries:** Material de investigación y libros.
- ❑ **Electronic news:** Movimiento hacia la difusión de noticias electrónicas
- ❑ **Web and Web-enabled applications:** La Web es un vasto repositorio de documentos que se enriquece con enlaces y otros tipos de información secundaria.

¿Cómo
representar
documentos de
texto?

Representación en Texto Libre de un sitio Web

Tango

Véase también: [Tango \(baile\)](#)

El **tango** es un **género musical** y una **danza**, característica de la región del **Río de la Plata** y su zona de influencia, principalmente de las ciudades de **Buenos Aires** (en **Argentina**) y **Montevideo** (en **Uruguay**). El escritor [Ernesto Sabato](#) destacó la condición de "híbrido" del tango.² El poeta [Eduardo Giorlandini](#) destaca sus raíces **afrorioplatenses**, con la **cultura gauchesca**, **hispana**, **aficana**, **italiana** y la enorme diversidad étnica de la gran ola inmigratoria llegada principalmente de **Europa**.³ La investigadora Beatriz Crisorio dice que "el tango es deudor de aportes multiétnicos, gracias a nuestro pasado colonial (indígena, africano y criollo) y al sucesivo aporte inmigratorio".⁴ Desde entonces se ha mantenido como uno de los géneros musicales cuya presencia se ha vuelto familiar en todo el mundo, así como uno de los más populares.^{5 6}

Distintas investigaciones señalan seis estilos musicales principales que dejaron su impronta en el tango: el **tango andaluz**, la **habanera cubana**, el **candombe**, la **milonga**, la **mazurca** y la **polka** europea.^{7 8}

El tango revolucionó el baile popular introduciendo una danza sensual con pareja abrazada que propone una profunda relación emocional de cada persona con su propio cuerpo y de los cuerpos de los bailarines entre sí. Refiriéndose a esa relación, [Enrique Santos Discépolo](#), uno de sus máximos poetas, definió al tango como «un pensamiento triste que se baila».⁹



Lexicon

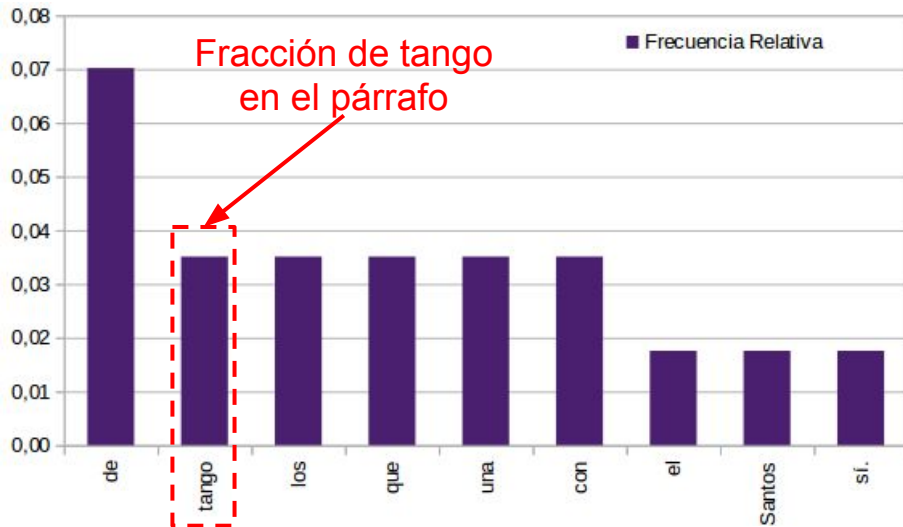
El conjunto completo y de las distintas palabras usadas para definir el corpus se conoce como **lexicón**.

el Santos sí danza relación al cada pensamiento El sensual
Refiriéndose baile baila tango máximos los relación sus esa
Discépolo persona abrazada de triste propone pareja cuerpo y que
revolucionó como emocional poetas bailarines a introduciendo entre
propio su definió una un cuerpos popular Enrique uno profunda se
con

Bolsa de palabras

El tango revolucionó el baile popular introduciendo una danza sensual con pareja abrazada que propone una profunda relación emocional de cada persona con su propio cuerpo y de los cuerpos de los bailarines entre sí. Refiriéndose a esa relación, Enrique Santos Discépolo, uno de sus máximos poetas, definió al tango como «un pensamiento triste que se baila»

57 palabras
2 oraciones



Términos	Frecuencia
de	4 /57
tango	2
los	2
que	2
una	2
con	2
el	1
Santos	1
sí.	1
danza	1
relación	1
al	1
cada	1
pensamiento	1
sensual	1
Refiriéndose	1
baile	1
baila	1
máximos	1
relación	1
sus	1
esa	1
Discépolo	1
persona	1
abrazada	1
triste	1
propone	1
pareja	1
cuerpo	1

Modelo de Bolsa de Palabras

El orden de las palabras no importa

¿Cuál es la probabilidad de sacar la palabra "tango" de la bolsa?

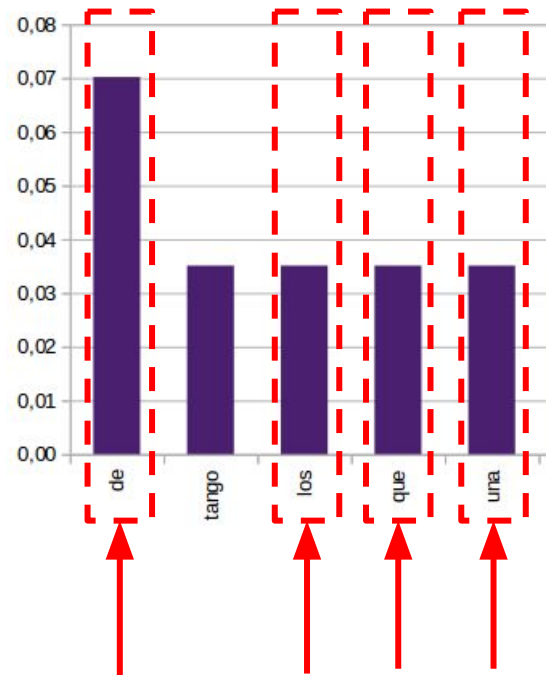
En el contexto de un problema de aprendizaje las palabras se tratan como **dimensiones** (o features) y sus valores corresponden a las frecuencias de ocurrencia.



Palabras que no ayudan

- ❑ Modelos de bolsa de palabras: muchas palabras frecuentes no ayudan.
- ❑ Podemos remover esas palabras de la bolsa.
- ❑ Esos términos son llamados: **stopwords o palabras vacías**.
- ❑ Podemos utilizar listas de **stopwords** ya curadas para cada idioma.

un una unas unos uno sobre todo también tras otro
algún alguno alguna algunos algunas ser es soy
eres somos sois estoy esta estamos estais estan
como en para atras porque por qué estado estaba
ante antes siendo ambos pero por poder puede
puedo podemos podeis pueden fui fue fuimos
fueron hacer hago hace hacemos haceis hacen
cada....



¿Qué tan útiles son estas palabras para entender la semántica?

Colección/Corpus

Un conjunto de datos corresponde a una colección de documentos, que también se conoce como **corpus**.

Esos documentos tienen que estar en un formato de texto plano.



Matriz de Término/Documento

Una matriz de Término-Documento es una forma de representar las palabras en el texto como una tabla (o matriz) de números.

Las **filas de la matriz representan** las unidades de estudio (**los documentos**) de texto que se analizarán, y las **columnas de la matriz representan los términos** del texto que se utilizarán en el análisis.

	El	popular	01	placentero	baile	Gardel	Escuchar	fue	tango	sueños	liquido	un	del	Tango	gran	valor	el	vendido	los	es	Con
El tango es un baile popular	1	1			1				1			1	1							1	
Escuchar tango es placentero				1			1		1											1	
Gardel es un gran valor del tango						1			1			1			1	1				1	
Con el Tango liquido los sueños										1	1			1			1		1		1
El Tango 01 fue vendido	1		1					1						1				1			

Características:

- ❑ La mayoría de los valores de las dimensiones son **cero**, y solo unas pocas dimensiones adquieren **valores positivos**.
- ❑ La matriz de TD es una representación de **alta dimensión, dispersa y no negativa**.

Stemming

Stemming es el proceso de consolidar palabras relacionadas con la misma raíz

Stemming se refiere al proceso de extracción de la raíz morfológica de una palabra, y varias heurísticas crudas se utilizan para lograr este objetivo

Form	Suffix	Stem
stud ies	-es	studi
stud ying	-ing	study
niñ as	-as	niñ
niñ ez	-ez	niñ

- ❑ Semi-automatic lookup tables
- ❑ Suffix stripping

efectúa	efectu
efectuaba	efectu
efectuada	efectu
efectuadas	efectu
efectuado	efectu
efectúan	efectu
efectuar	efectu
efectuará	efectu
efectuarán	efectu
efectuaría	efectu
efectuaron	efectu
efectuarse	efectu
efectúen	efectu
efectuo	efectu
efectúo	efectu
efectuó	efectu

Lematización

- ❑ La lematización es un enfoque más sofisticado porque usa la parte específica del habla para determinar la raíz de una palabra.
- ❑ Las reglas de normalización dependen de la parte del discurso y, por lo tanto, son altamente específicas del idioma.

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

Frases como un solo término

Existen términos que por sí solos no gravitan en el vocabulario y pierden valor.

Nombres compuestos, es un caso:

Refiriéndose a esa relación, **Enrique Santos Discépolo**, uno de sus máximos poetas, definió al tango como «un pensamiento triste que se baila»

Otros ejemplos: Buenos Aires, Santa Cruz, etc.

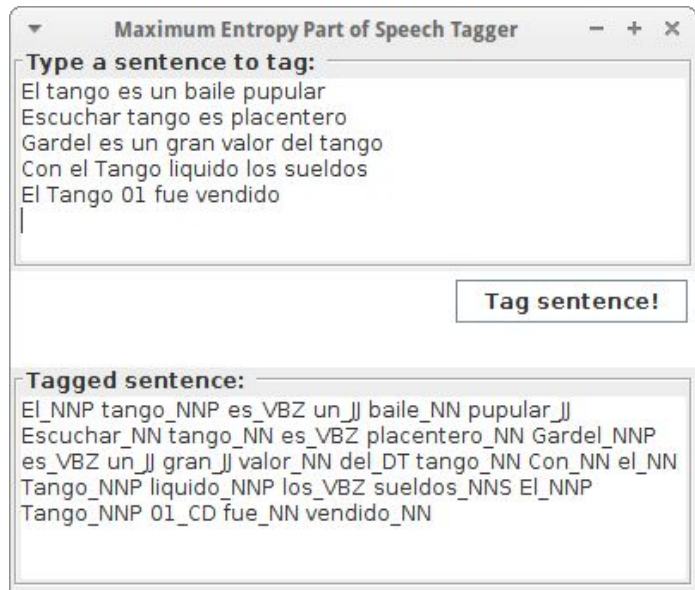
Necesitamos reconocer Entidades (NER Named Entity Recognition)

Es un subtask de **Extracción de Información (IE)** cuyo objetivo es identificar y clasificar expresiones de un texto que hacen referencia a **personas, organizaciones, lugares, marcas comerciales e incluso fechas, horas y medidas**

Estas expresiones se denominan ENTIDADES

Otras tareas del Procesamiento del Lenguaje Natural

Part-of-speech tagging: averiguar qué son sustantivos, verbos, adjetivos, etc.



The screenshot shows a web-based interface for a Maximum Entropy Part of Speech Tagger. It has a title bar with a dropdown arrow, the text "Maximum Entropy Part of Speech Tagger", and standard window controls. The main area is divided into two sections. The top section, titled "Type a sentence to tag:", contains a text input field with the following text: "El tango es un baile popular", "Escuchar tango es placentero", "Gardel es un gran valor del tango", "Con el Tango liquido los sueldos", and "El Tango 01 fue vendido". Below this input field is a button labeled "Tag sentence!". The bottom section, titled "Tagged sentence:", displays the output of the tagging process, where each word is followed by its assigned part-of-speech tag in a smaller font.

Maximum Entropy Part of Speech Tagger

Type a sentence to tag:

El tango es un baile popular
Escuchar tango es placentero
Gardel es un gran valor del tango
Con el Tango liquido los sueldos
El Tango 01 fue vendido

Tag sentence!

Tagged sentence:

El_NNP tango_NNP es_VBZ un_JJ baile_NN popular_JJ
Escuchar_NN tango_NN es_VBZ placentero_NN Gardel_NNP
es_VBZ un_JJ gran_JJ valor_NN del_DT tango_NN Con_NN el_NN
Tango_NNP liquido_NNP los_VBZ sueldos_NNS El_NNP
Tango_NNP 01_CD fue_NN vendido_NN

El_NNP tango_NNP es_VBZ un_JJ baile_NN
popular_JJ Escuchar_NN tango_NN es_VBZ
placentero_NN Gardel_NNP es_VBZ un_JJ
gran_JJ valor_NN del_DT tango_NN Con_NN
el_NN Tango_NNP liquido_NNP los_VBZ
sueldos_NNS El_NNP Tango_NNP 01_CD
fue_NN vendido_NN

Bibliografía

- ❑ 95-865: Unstructured Data Analytics (Spring 2018 Mini 4) Clase 1 [[Slides](#)]
- ❑ Aggarwal, C. C. (2018). Machine Learning for Text. Springer, Cham.