UNIVERSIDAD
NACIONAL
DE COLOMBIA

Introducción a la Minería de Datos

Verónica Guarín Escudero
Escuela de Estadística

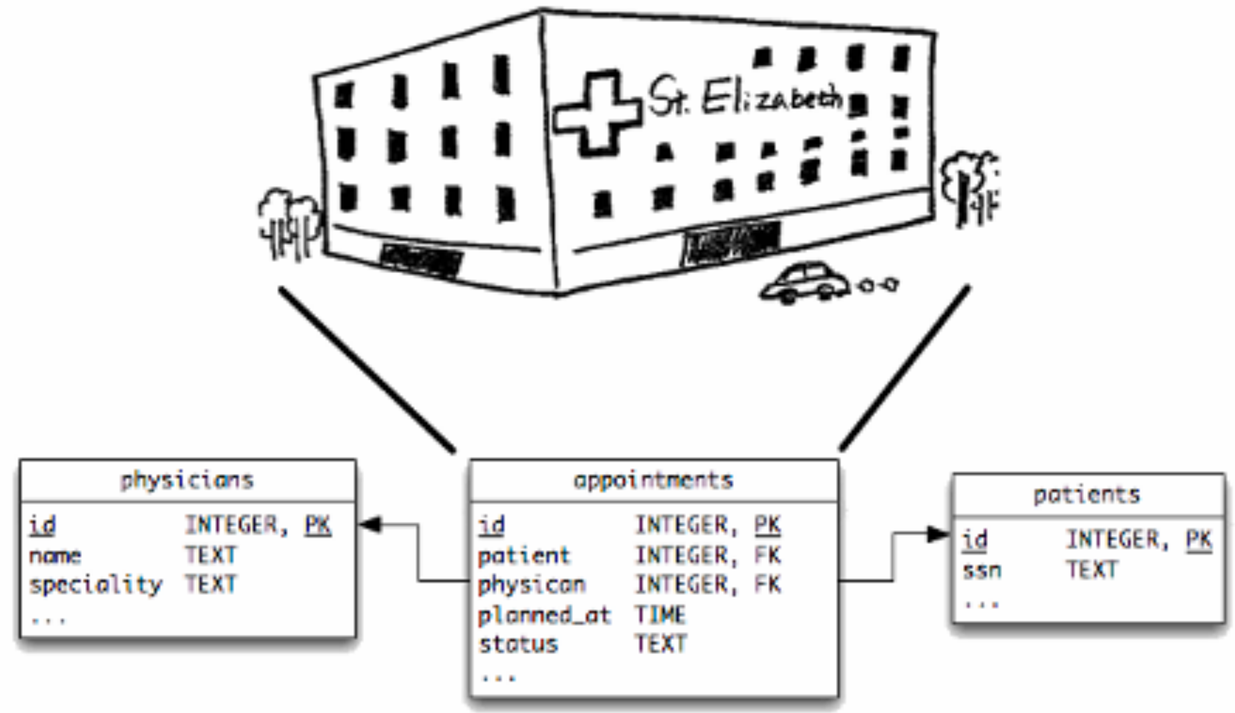
Correo: jvguarine@unal.edu.co

Datos no estructurados.
De bases de datos textuales a características.

Datos Estructurados

Elementos bien definidos, relaciones entre elementos

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



Datos no Estructurados

Francisco de Goya

152 idiomas

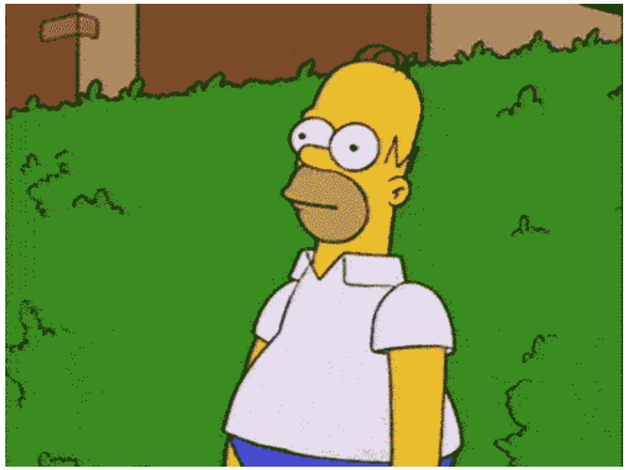
Artículo Discusión Leer Editar Ver historial Herramientas

«Goya» redirige aquí. Para otras acepciones, véase Goya (desambiguación).

Francisco José de Goya y Lucientes[1] (Fuendetodos, 30 de marzo de 1746[2] -Burdeos, 16 de abril de 1828[a]) fue un pintor y grabador español. Su obra abarca la pintura de caballete y mural, el grabado y el dibujo. Su estilo evolucionó desde el rococó, pasando por el neoclasicismo, hasta el prerromanticismo, siempre interpretados de una forma personal y original,[3] y siempre con un rasgo subyacente de naturalismo, del reflejo de la realidad sin una visión idealista que la edulcore ni desvirtúe, donde es igualmente importante el mensaje ético. Para Goya la pintura es un vehículo de instrucción moral, no un simple objeto estético.[4] Sus referentes más contemporáneos fueron: Giambattista Tiepolo y Anton Raphael Mengs, aunque también recibió la influencia de Diego Velázquez y Rembrandt.[5] El arte goyesco supone uno de los puntos de inflexión que entre los siglos XVIII y XIX anuncian la pintura contemporánea y es precursor de algunas de las vanguardias pictóricas del siglo XX, especialmente el expresionismo.[5] [6] Por todo ello, se lo considera uno de los artistas españoles más relevantes y uno de los grandes maestros de la historia del arte mundial.

Retrato del pintor Francisco de Goya (1826), por Vicente López, Museo del Prado, Madrid.

Además, su obra refleja el convulso periodo histórico en que vive, particularmente la guerra de la Independencia, de la que la serie de estampas de *Los desastres de la guerra* es casi un reportaje moderno de las atrocidades cometidas[7] y compone una visión exenta de heroísmo donde las víctimas son siempre los individuos de cualquier clase y condición. Elogiado por Gustave Doré y E.T.A. Hoffmann, Charles Baudelaire describió su *Capricho 43*, *El sueño de la razón produce monstruos*, como "*cauchemar plein de choses inconnues*" (una pesadilla llena de cosas desconocidas).[8]



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

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.

Datos no Estructurados

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 documentos de texto

Representación en Texto Libre de un sitio Web

Tango

17 idiomas

Artículo Discusión | Leer Editar Ver historial Herramientas

Para otros usos de este término, véase Tango (desambiguación).

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, pero 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 afrorrioplatenses, con la cultura gauchesca, española, 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 conocidos.[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]

Tango

Patrimonio de la Humanidad de la Unesco

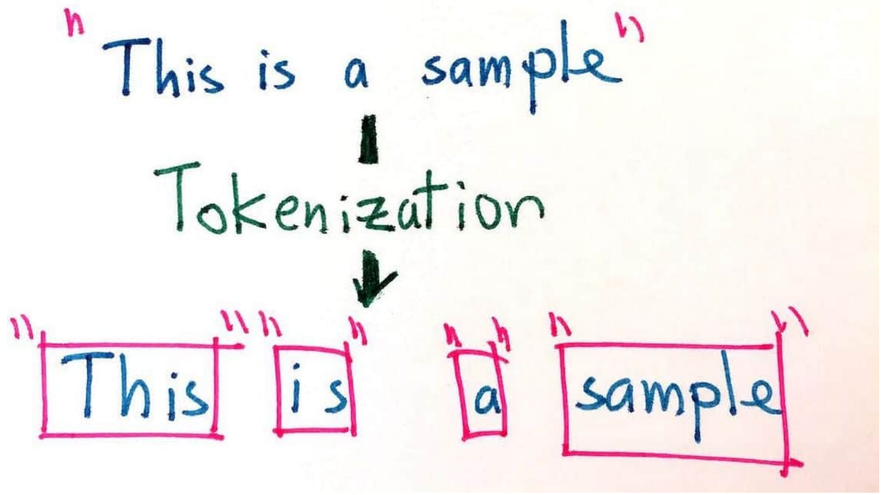
Pareja bailando tango en San Telmo, Buenos Aires.

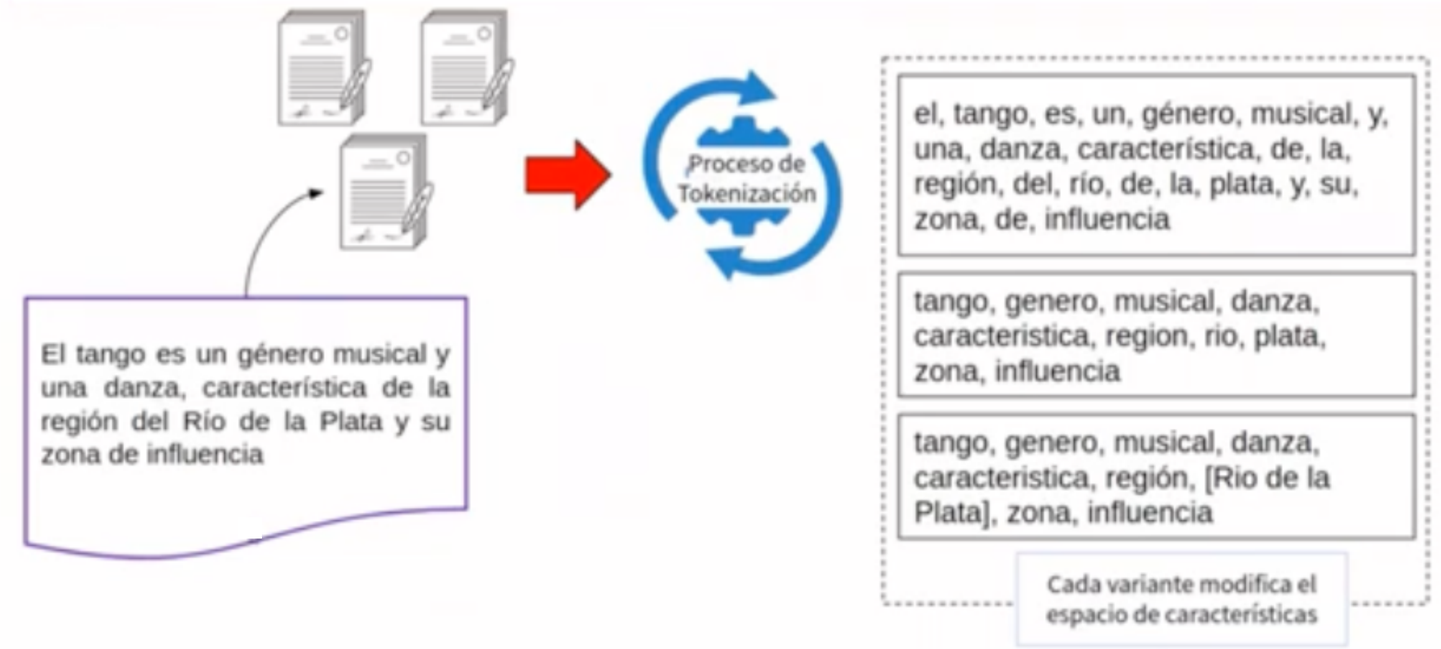
Localización

UNIVERSIDAD NACIONAL DE COLOMBIA

Representación documentos de texto

Tokenización

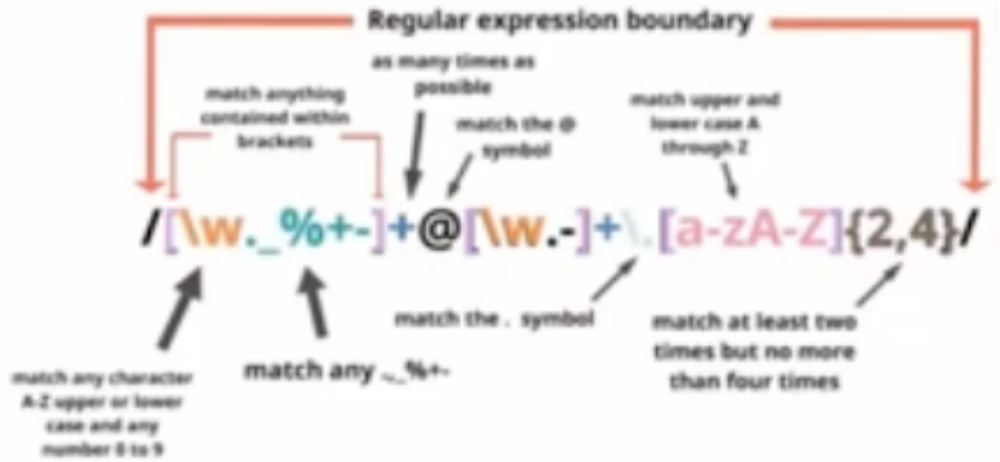




Representación documentos de texto

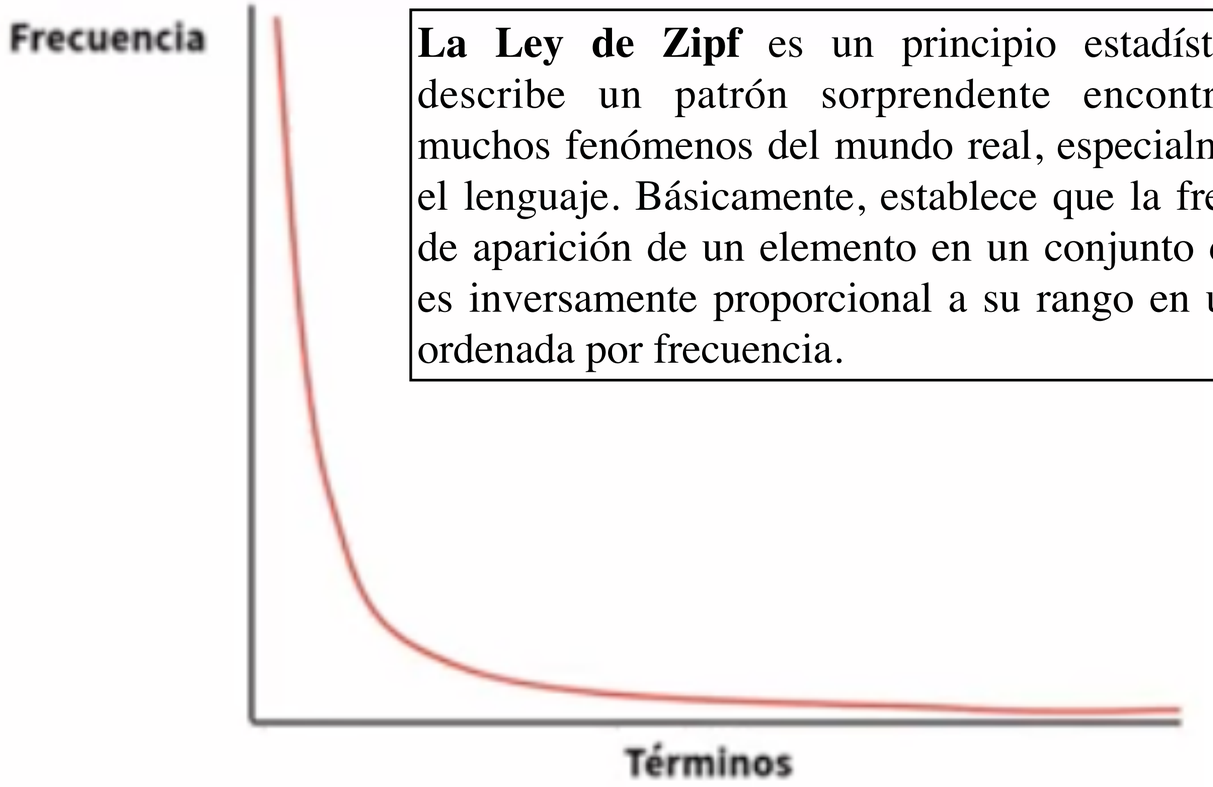
Técnicas

- Tokenización y Normalización
  - Expresiones Regulares
  - Extraer fechas, URLs, direcciones de mail, etc.
- Detección de Entidades
  - Por ejemplo, nombres propios de interés.
- Eliminar Palabras Vacías
- Stemming
  - computer, computing, compute, computation → **comput**



Representación documentos de texto

Ley de Zipf



La Ley de Zipf es un principio estadístico que describe un patrón sorprendente encontrado en muchos fenómenos del mundo real, especialmente en el lenguaje. Básicamente, establece que la frecuencia de aparición de un elemento en un conjunto de datos es inversamente proporcional a su rango en una lista ordenada por frecuencia.

¿Por qué es importante la Ley de Zipf?

- **Lingüística:** Ayuda a entender cómo se estructura el lenguaje y cómo las personas usan las palabras.
- **Informática:** Se utiliza en compresión de datos, búsqueda de información y análisis de texto.
- **Otras disciplinas:** Se aplica en economía, biología, física y muchas otras áreas.

Aplicaciones de la Ley de Zipf:

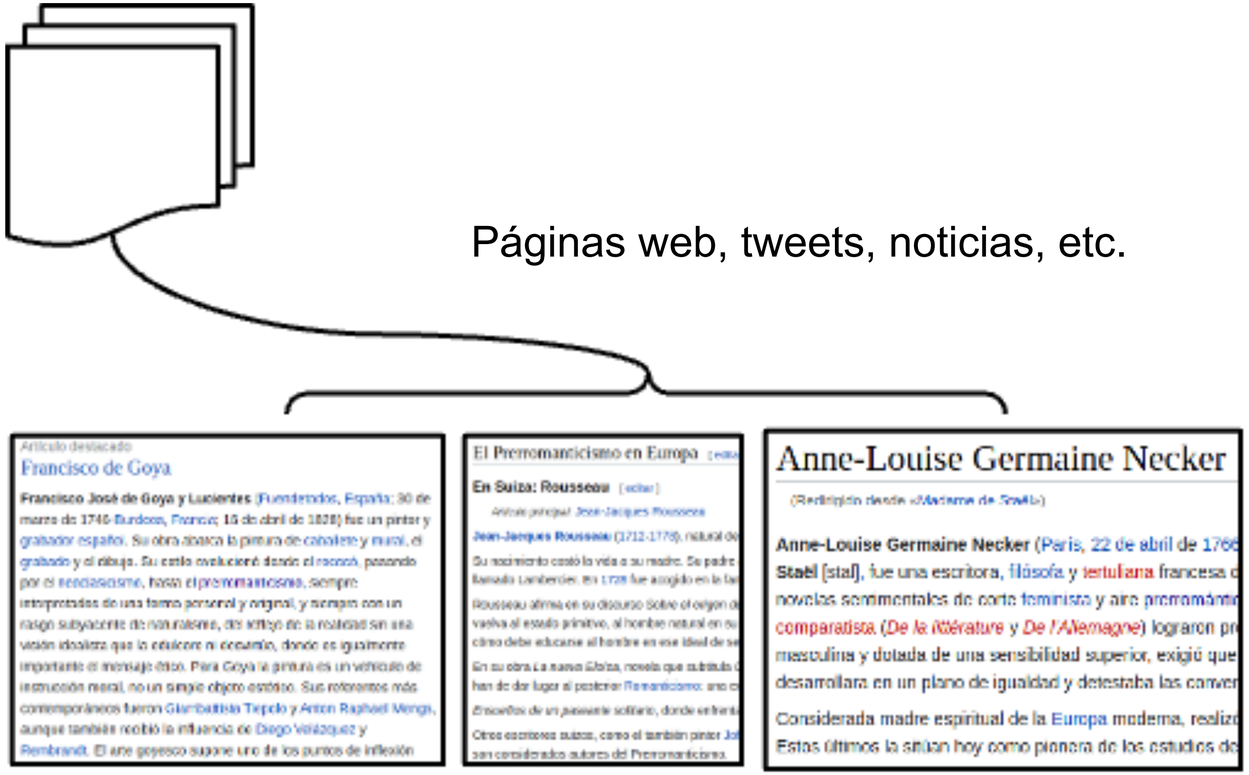
- **Optimización de motores de búsqueda:** Al entender qué palabras son más comunes, los motores de búsqueda pueden mejorar los resultados de las búsquedas.
- **Análisis de redes sociales:** Se utiliza para estudiar la frecuencia con la que se mencionan ciertos temas o palabras clave en las redes sociales.
- **Compresión de datos:** La Ley de Zipf se puede aprovechar para desarrollar algoritmos de compresión de datos más eficientes.

Representación documentos de texto

Corpus/colección

Un conjunto de documentos es denominado **corpus** o **colección**

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



Representación documentos de texto

Lexicon

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

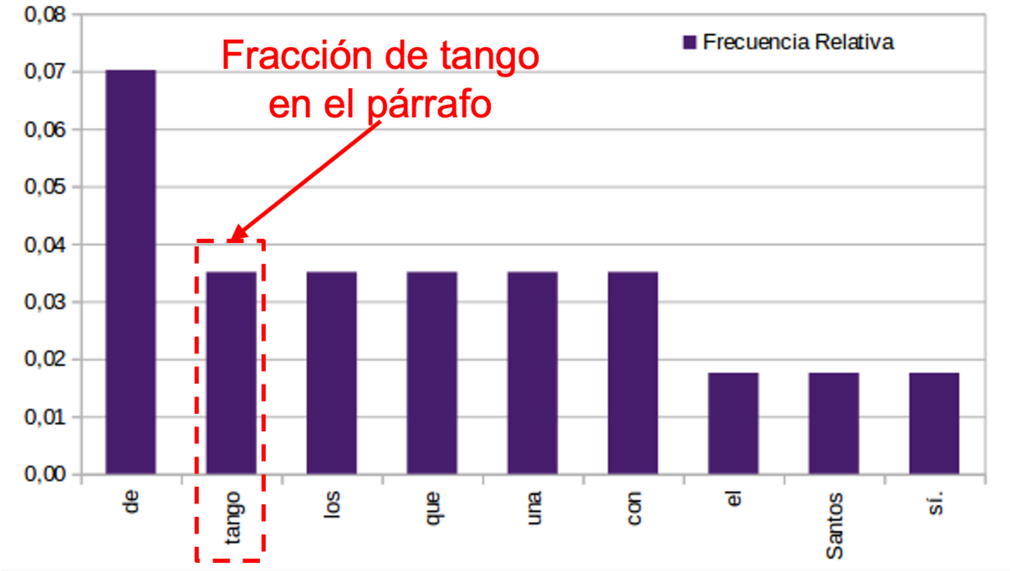
el Santos sí danza relación al cada pensamiento El sensual Refiriéndose baile baila tango máximos los relación sus esa Discépolo persona abrazada de triste propone pareja cuerpo y que revolucionó como emocional poetas bailarines a introduciendo entre propio su definió una un cuerpos popular Enrique uno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

Representación documentos de texto

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

Representación documentos de texto

Modelo de Bolsa de palabras

El modelo de bolsa de palabras es una técnica fundamental en el procesamiento del lenguaje natural (PLN) que se utiliza para representar documentos textuales de una manera que las máquinas puedan entender.

El orden de las palabras no importa

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

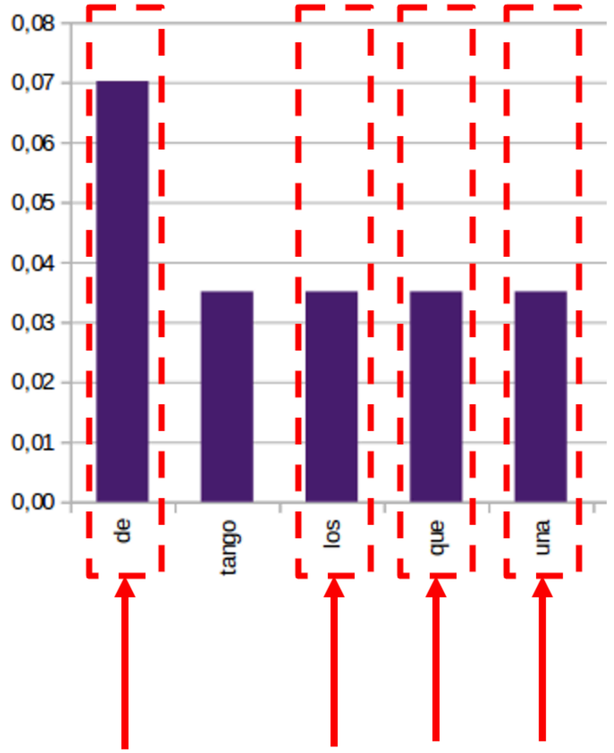
En el contexto de un problema de aprendizaje las palabras se tratan como **dimensiones** (o features) y sus valores corresponden a las frecuencias de ocurrencia.

Representación documentos de texto

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 pueden fui fue fuimos fueron hacer hago hace hacemos hacen cada….



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

Representación documentos de texto

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	sueldos	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 sueldos										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**.

Representación documentos de texto

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
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

¿Por qué es útil el stemming?

- **Reducción de la dimensionalidad:** Al reducir las palabras a su raíz, disminuimos el número de términos únicos en un corpus, lo que simplifica los modelos de análisis.
- **Mejora de la precisión:** Palabras con la misma raíz, pero diferentes flexiones, son consideradas como la misma palabra. Esto es especialmente útil en tareas como la búsqueda de información y la clasificación de textos.
- **Aumento del recall:** Al considerar las palabras en su forma raíz, es más probable encontrar documentos relevantes en una búsqueda.

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

Consideremos las palabras "corriendo", "correr" y "corrida". El stemming reduciría estas palabras a su raíz común: "corr".

Representación documentos de texto

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

Consideremos la palabra "corriendo".

- **Stemming:** Podría reducirla a "corri", que no es una palabra válida.
- **Lematización:** Identificaría "correr" como el lema, que es la forma base y válida de la palabra.

Representación documentos de texto

Frases como un solo término

El reconocimiento de entidades es una técnica fundamental dentro del procesamiento del lenguaje natural (PLN) que se encarga de identificar y clasificar elementos específicos dentro de un texto. Estos elementos, llamados **entidades**, pueden ser personas, organizaciones, ubicaciones, fechas, cantidades, etc.

¿Para qué sirve?

Imagina que tienes un montón de reseñas de productos. El reconocimiento de entidades te permitiría identificar automáticamente los nombres de los productos, las marcas y las opiniones de los usuarios. Esto sería muy útil para realizar análisis de sentimiento, estudios de mercado o para crear sistemas de recomendación.

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»

Estas expresiones se denominan ENTIDADES

Otros ejemplos: Buenos Aires, Santa Cruz, etc.
Necesitamos reconocer Entidades (NER Named Entity Recognition)

Es un subtarea 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**

Representación documentos de texto

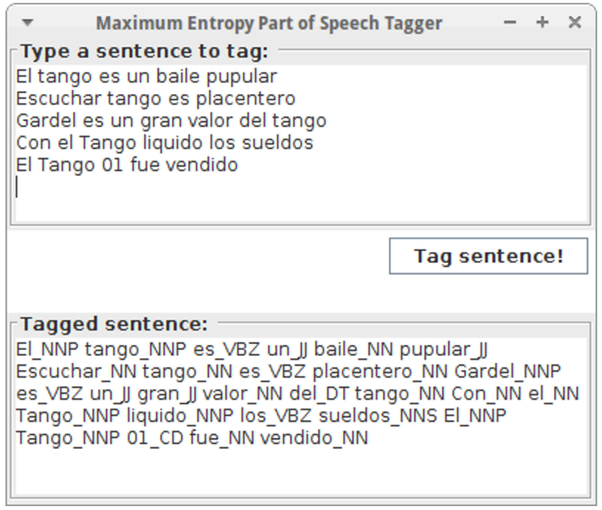
Otras tareas del Procesamiento del Lenguaje Natural

Part-of-speech tagging (Etiquetado morfosintáctico): averiguar qué son sustantivos, verbos, adjetivos, etc.

El etiquetado morfosintáctico, también conocido como *Part-of-Speech tagging* (POS tagging), es una técnica fundamental en el procesamiento del lenguaje natural (PLN) que consiste en asignar a cada palabra de una oración una etiqueta que indica su categoría gramatical o parte del discurso. Estas categorías pueden ser sustantivos, verbos, adjetivos, adverbios, preposiciones, conjunciones, etc.

¿Para qué sirve?

- **Análisis sintáctico:** Es el primer paso para comprender la estructura gramatical de una oración y las relaciones entre las palabras.
- **Análisis semántico:** Ayuda a determinar el significado de las palabras en el contexto de una oración.
- **Otras tareas de PLN:** Es una tarea previa necesaria para muchas otras aplicaciones de PLN, como la traducción automática, el resumen de textos, la generación de lenguaje natural y el análisis de sentimientos.



El_NNP tango_NNP es_VBZ un_JJ baile_NN pupular_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.

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Gracias