



Text Mining

Luis Gascó Sánchez, Ph.D.



Sobre mi:

Formación

- Doctor en Ingeniería (Universidad Politécnica de Madrid)
- Certificado de doctorado europeo en innovación (EIT Digital Doctoral School)
- Master en Big Data e Inteligencia de Negocio (Escuela de Organización Industrial)
- Master Universitario en Ingeniería Acústica (Universidad Politécnica de Madrid)
- Ingeniero de Imagen y Sonido (Telecomunicaciones) (Universidad Politécnica de Madrid)

Puesto actual

Text Mining Research Engineer
@ Text Mining Unit (BSC)



Experiencia



NOKIA Bell Labs

Atos



Cronograma

Día 1:

1. Introducción
 1. Contexto histórico
 2. ¿Qué es el Text Mining?
 3. Librerías de programación para Text Mining
2. Técnicas y conceptos básicos de NLP
3. Representación numérica de documentos textuales



Cronograma

Día 2:

1. Técnicas de Text Mining:
 1. Flujo de los datos
 2. Clasificación
 3. Topic Modeling
2. Caso de estudio: Análisis de sentimiento



Evaluación

Clases teórico prácticas

+

Ejercicio(s) prácticos





cuál es la capital de zimbabue



Todo

Noticias

Maps

Imágenes

Vídeos

Más

Configuración

Herramientas

Aproximadamente 3.170.000 resultados (1,21 segundos)

Zimbabue / Capital



Harare

Harare (denominada Salisbury hasta 1982) es la ciudad más poblada y capital de Zimbabue. Tiene una población estimada de 1.600.000 habitantes, con unas 2.800.111 personas en su área metropolitana (2006). Es el centro administrativo, comercial, y de comunicaciones de Zimbabue.

<https://es.wikipedia.org/wiki/Harare>

[Harare - Wikipedia, la enciclopedia libre](#)

Sugerencias

Otras preguntas de los usuarios

¿Dónde está Zimbabue?



¿Cómo se llama la capital de Zambia?



¿Qué país es ahora Rodesia?



Por qué



- por qué
- por qué **o porqué**
- por qué **ronronean los gatos**
- por qué **el cielo es azul**
- por qué **no te callas**
- por qué **la sal derrite el hielo**
- por qué **se inventaron los cereales**
- por qué **la nieve es blanca**
- por qué **a los madrileños se les llama gatos**
- por qué **matan las mujeres**

Buscar con Google

Voy a tener suerte

[Denunciar predicciones inadecuadas](#)
[Más información](#)

Buscar en Twitter


< ti COVID-19 Tendencias Noticias Deportes Entretenimiento

Tendencias de España

1 · Fútbol · Tendencia

Chelsea

UEFA Champions League
Atletico Madrid vs Chelsea



156 mil Tweets

2 · Fútbol · Tendencia

Lemar

4.620 Tweets

3 · Tendencias

#pasapalabra200

1.343 Tweets

4 · Tendencias

#FirstDates23F

Tendencias sobre #CiudadNova23Feb, #MatrimonioDivinity23F

5 · Fútbol · Tendencia

Bayern

UEFA Champions League
Lazio vs Bayern Munich





Alyssa Leader @alittleleader · Mar 30

Fastest service ever— thanks @zapier for offering our project free support for our COVID related work! We love y'all big time.



Alyssa Leader @alittleleader · Mar 30

Hey @zapier! I'm an organizer that put together a team of over 3,000 law students to provide pro bono support to attorneys working on COVID-19 related matters! We are using Zapier to automate some of our work flows and loving it! But we quickly used up our free automations! 1/

Show this thread

3

2

52



Zapier

@zapier

Replying to @alittleleader

We're happy to help! 

9:38 AM · Mar 31, 2020 · Twitter Web App

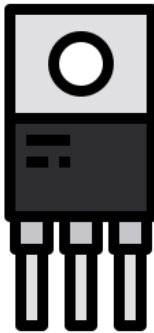
1. Introducción

Contexto histórico

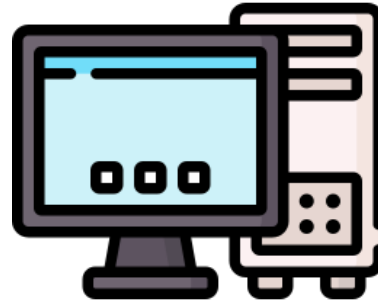


Revolución Digital: La Tercera Revolución Industrial

Transistor



Ordenador personal



ARPANET



1. Introducción

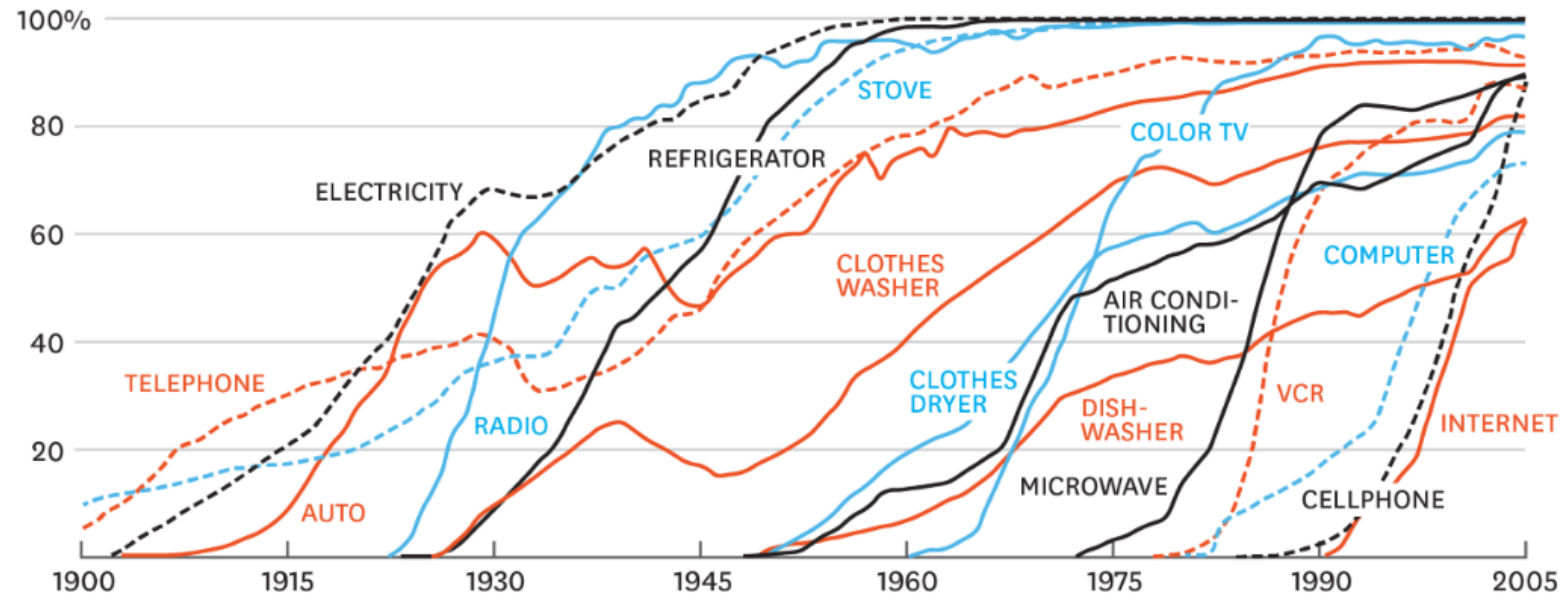
Contexto histórico



Revolución Digital: *La Tercera Revolución Industrial*

CONSUMPTION SPREADS FASTER TODAY

PERCENT OF U.S. HOUSEHOLDS



SOURCE NICHOLAS FELTON, THE NEW YORK TIMES

HBR.ORG

Reducción del tiempo de **adopción tecnológica**, gracias a:

- Mejora de las líneas de producción
 - + producción
 - - coste
- Incremento exponencial de la potencia. Ley de Moore.
- Bonanza económica

1. Introducción

Contexto histórico



Revolución digital



Era de la información

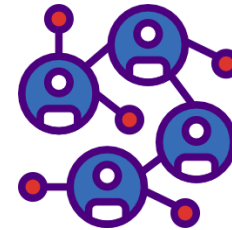


Importancia de las
TICs

Tecnología como motor de la evolución social



Web 2.0.



Redes sociales



Digitalización y
sensorización



DATOS

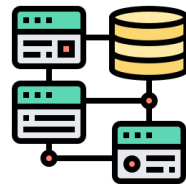
1. Introducción

Contexto histórico



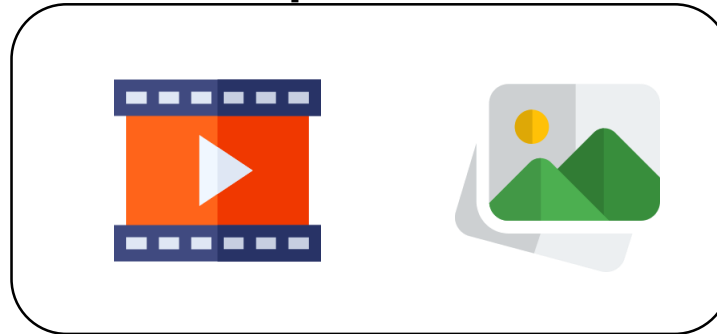
DATOS

Estructurados



No estructurados

Visión computacional



Text
Mining

1. Introducción

Text Mining



Definición

El proceso de transformación de datos textuales no estructurados a un formato tabular que permita analizarlos y extraer conocimiento

1. Introducción

Text Mining



Búsqueda y recuperación de información

Agrupación de documentos

Clasificación de documentos

Extracción de información

Extracción de conceptos

Procesado de Lenguaje Natural

1. Introducción

Librerías de Text Mining

NLTK

*University of
Pennsylvania*

Año 2001

**Orientada a
investigación**

Principalmente para
preprocesado

Spacy

*Ines Montani y
Matthew Honnibal*

Año 2015

**Orientada a
producción**

Preprocesado,
construcción de
modelos...

Gensim

Radim Rehurek

Año 2009

**Orientada a
producción**

Análisis preliminar

TextBlob

CoreNLP

Polyglot

1. Introducción

NLTK 3.5 documentation

[NEXT](#) | [MODULES](#) | [INDEX](#)

spaCy

🚀 Out now: spaCy v3.0

[USAGE](#)

[MODELS](#)

[API](#)

[UNIVERSE](#)



19,659

Industrial-Strength Natural Language Processing

IN PYTHON

Get things done

spaCy is designed to help you do real work — to build real products, or gather real insights. The library respects your time, and

Blazing fast

spaCy excels at large-scale information extraction tasks. It's written from the ground up in carefully memory-managed Cython. If

Awesome ecosystem

In the five years since its release, spaCy has become an industry standard with a huge ecosystem. Choose from a variety of plugins,

```
>>> sentence = """At eight o'clock on Thursday morning  
... Arthur didn't feel very good."""
```


2. Técnicas y terminología básica de NLP



2. Técnicas y terminología básica de NLP

Procesado de Lenguaje Natural, Natural Language Processing (NLP)

“Un lenguaje natural es una forma de lenguaje humano con fines comunicativos que tiene asociadas una serie de reglas sintácticas, conocidas como sintaxis”



Lenguajes formales

vs



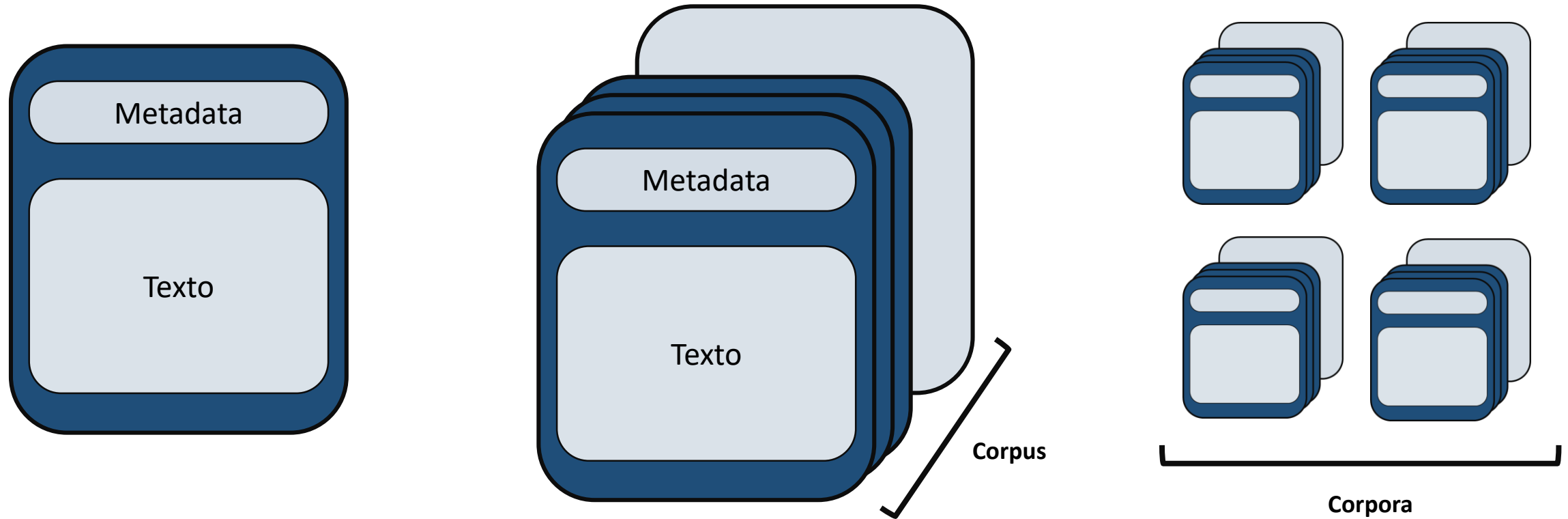
Lenguajes naturales

- Ej. Lenguajes de programación
- Artificiales
- Utilizados en aplicaciones específicas
- Sintaxis

- Ej. Español, Inglés, ...
- Naturales, evolución de otros
- Comunicación entre humanos
- Sintaxis y semántica

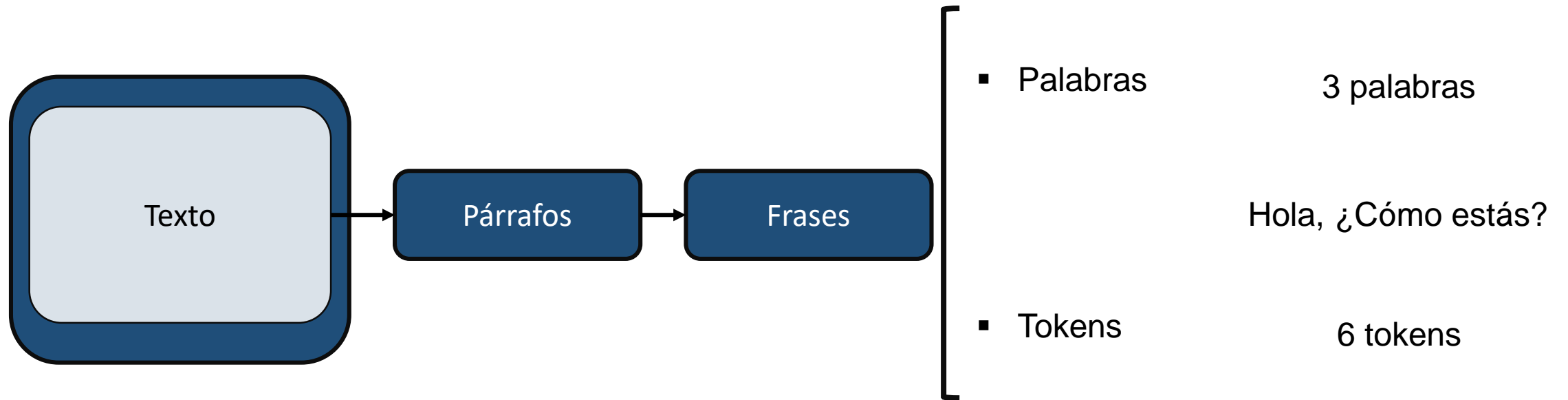
2. Técnicas y terminología básica de NLP

Corpora



2. Técnicas y terminología básica de NLP

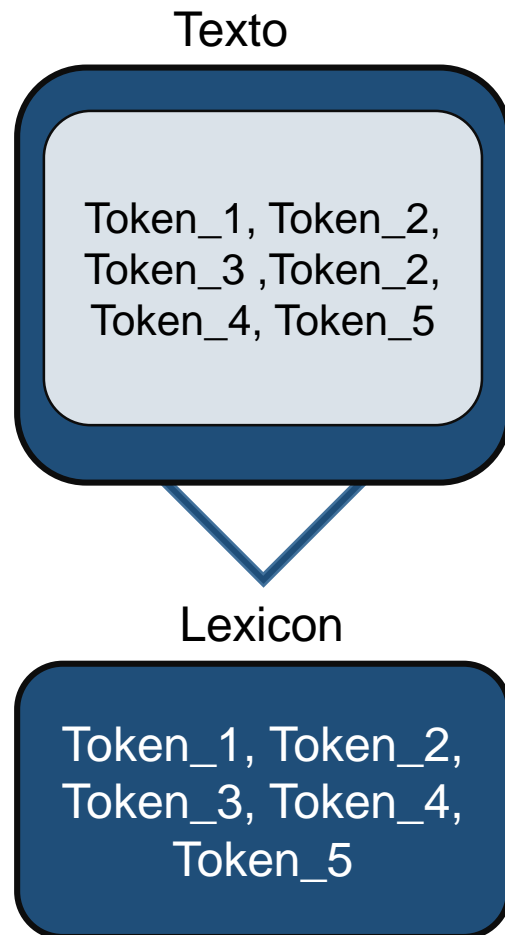
Tokens



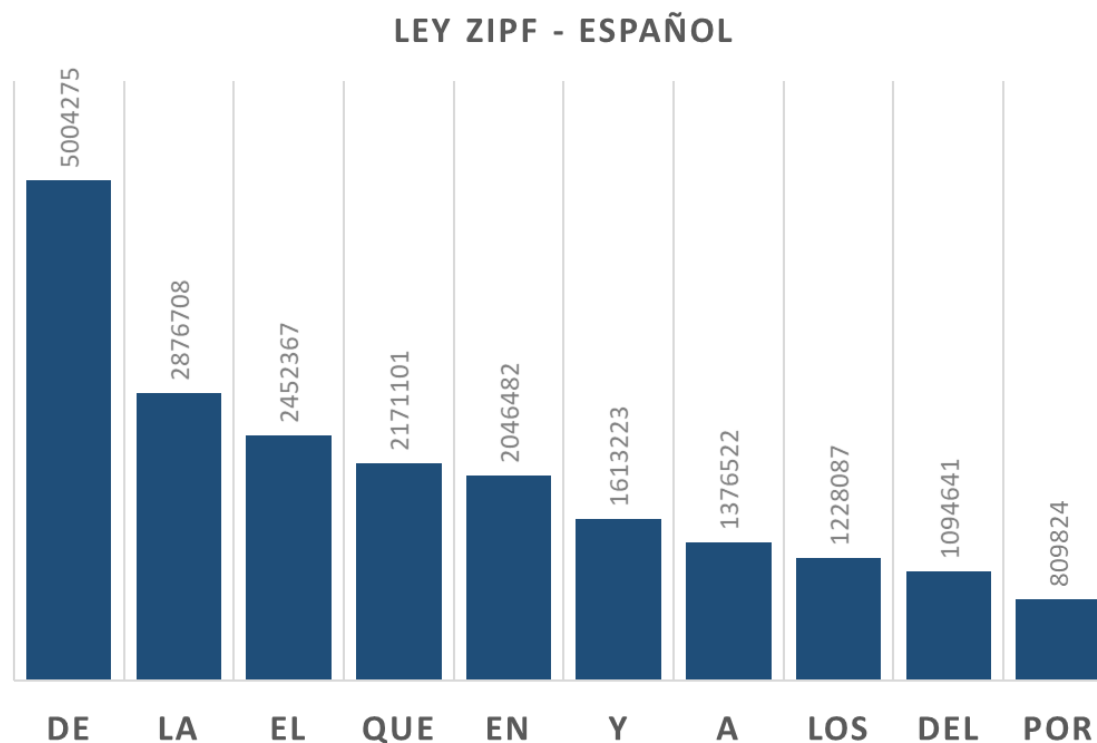
2. Técnicas y terminología básica de NLP

Lexicon y stopwords

Lexicon o vocabulario



Stopwords



2. Técnicas y terminología básica de NLP

n-gramas

Francisco ha comido demasiadas patatas

Unigrama/token

1	Francisco
2	ha
3	comido
4	demasiadas
5	patatas

Bigrama

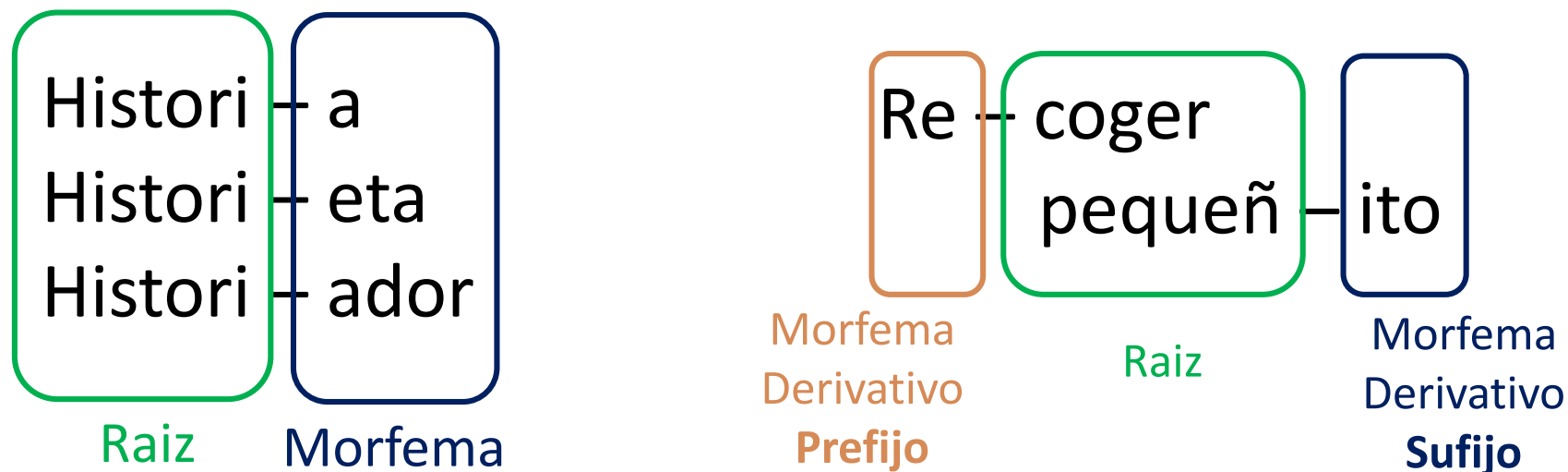
1	Francisco ha
2	ha comido
3	comido demasiadas
4	demasiadas patatas

Trigrama

1	Francisco ha comido
2	ha comido demasiadas
3	comido demasiadas patatas

2. Técnicas y terminología básica de NLP

Lexemas/Lemas y morfemas



Los morfemas aportan más significado a los lexemas, a costa de incrementar la dimensionalidad de los datos y en ocasiones que los modelos aprendan peor.

¿Solución? Eliminar los morfemas

2. Técnicas y terminología básica de NLP

Stemming y Lemmatization

Stemming

Uso de reglas sintácticas para eliminar stems

Ej: Porter Stemming o Snowball Stemming

Speaking
Speaks
Speaker
Bus
Buses

Original form

Speak
Speak
Speak
Bus
Buse

Stemmed form

Lematización

Búsqueda de lemas a partir de diccionarios jerárquicos o similar

Ej: WordNetLemmatizer()

Computes
Computing
Computed

Original form

Compute
Compute
Compute

Lemmatized form

2. Técnicas y terminología básica de NLP

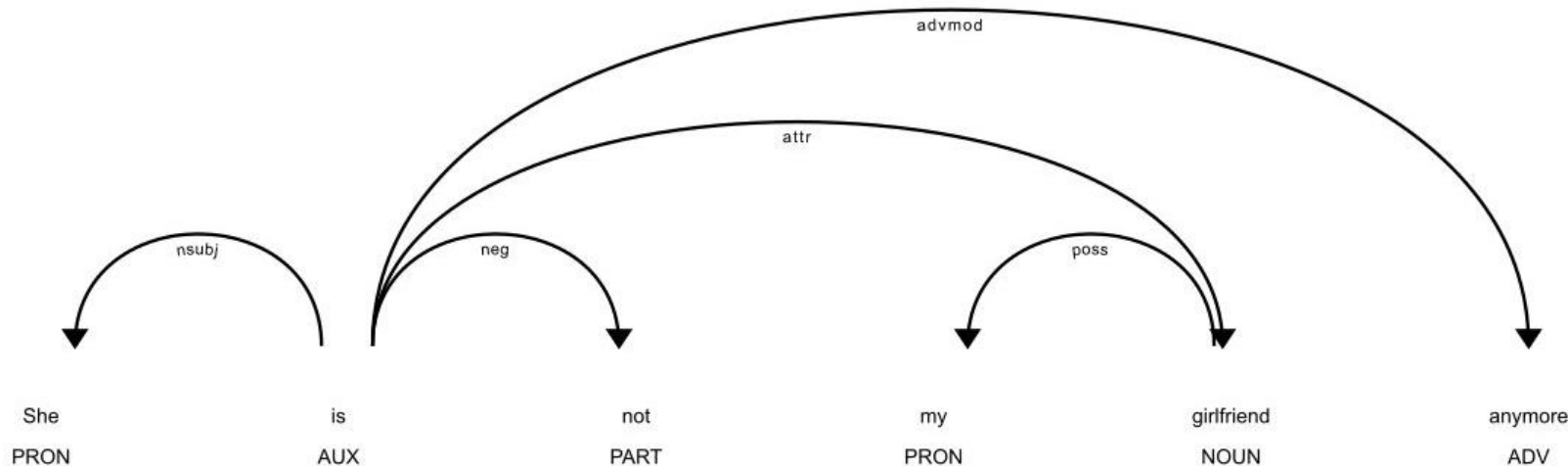
PoS y Parsing

PoS = Part-of-Speech (Categoría gramatical)

Parsing = Análisis sintáctico

She is not my girlfriend anymore

text	lemma_	pos_	tag_
She	she	PRON	PRP
is	be	AUX	VBZ
not	not	PART	RB
my	my	PRON	PRP\$
girlfriend	girlfriend	NOUN	NN
anymore	anymore	ADV	RB



2. Técnicas y terminología básica de NLP

PoS y Parsing

Las categorías gramaticales de las palabras son de utilidad para varias tareas en el Text Mining:

Information
Retrieval

Information
Extraction

Clasificación

Topic Modeling

2. Técnicas y terminología básica de NLP

NER

La trabajadora María Martínez fue trasladada en el año 2019 a las oficinas de San Francisco (California) por su gran rendimiento en el año 2019. Sin embargo, le ofrecieron ser jefa de tecnología de la empresa Youtube y decidió aceptar.

La trabajadora María Martínez PER fue trasladada en el año 2019 a las oficinas de San Francisco LOC (California LOC) por su gran rendimiento en el año 2019. Sin embargo, le ofrecieron ser jefa de tecnología de la empresa Youtube ORG y decidió aceptarr

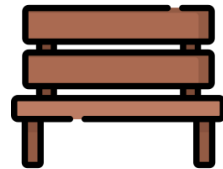
2. Técnicas y terminología básica de NLP

Semántica - WordNet

Semántica

Rama que estudia el significado de las palabras, así como las diversas relaciones de sentido que se establecen entre ellas.

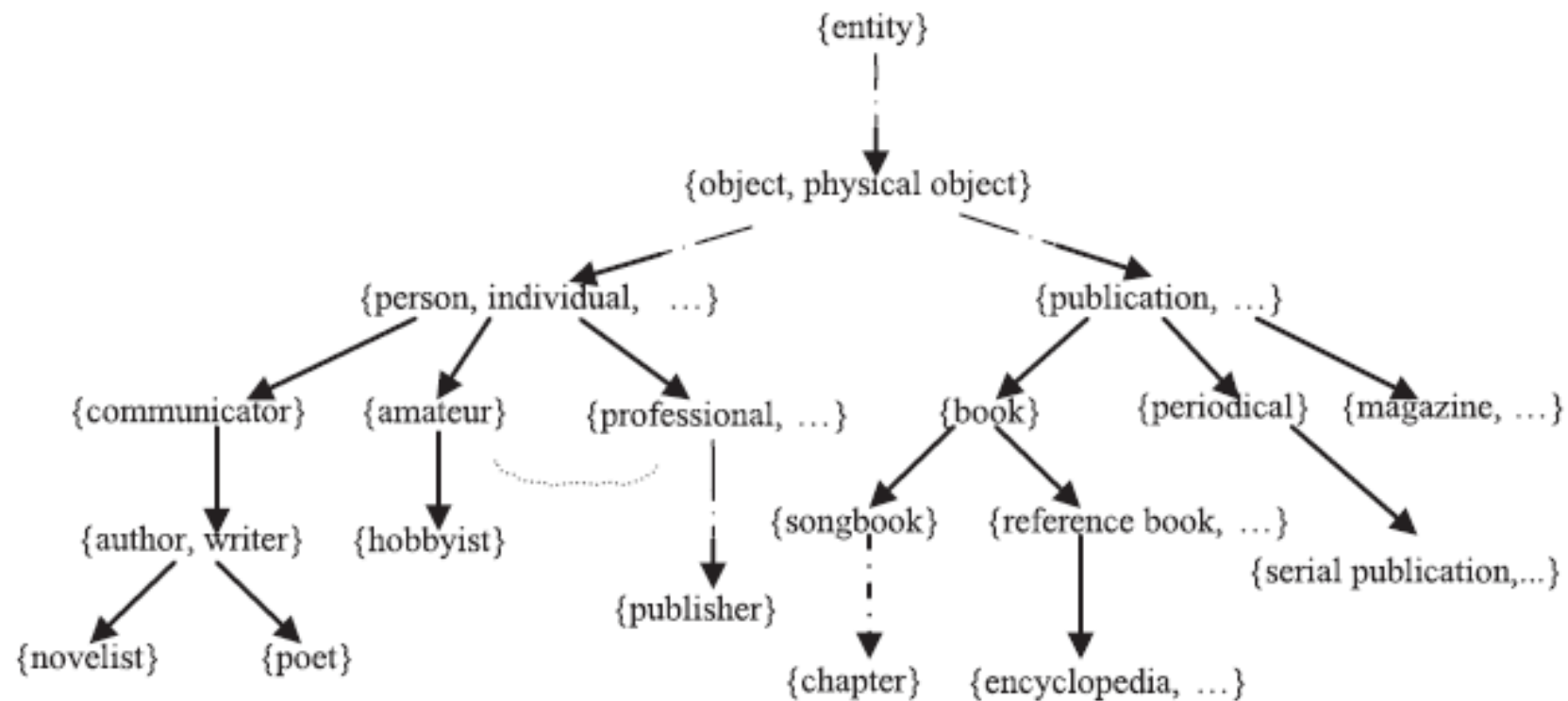
Vamos al banco, estoy cansado



WordNet 3.0

2. Técnicas y terminología básica de NLP

Semántica - WordNet



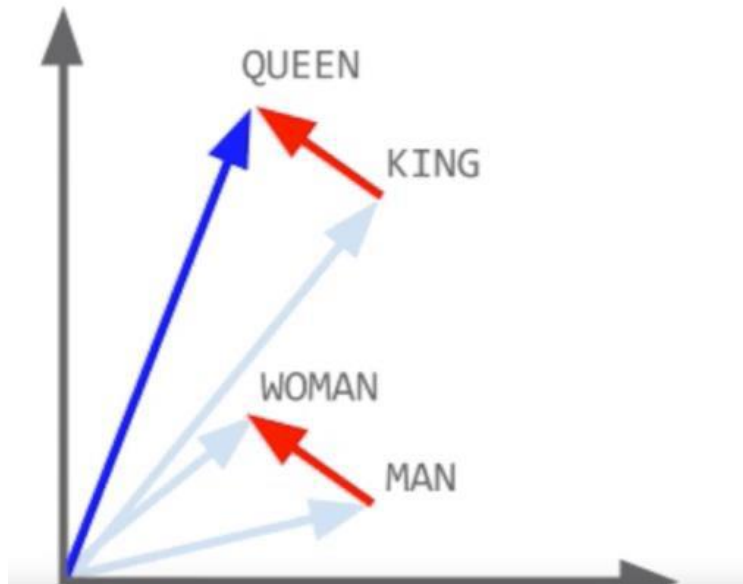
Hiperónimos (Hypenyms)

Hipónimos (Hyponyms)

2. Técnicas y terminología básica de NLP

Semántica – Word Embeddings

El texto es unidimensional. ¿Y si proyectamos las palabras a espacios multidimensionales?



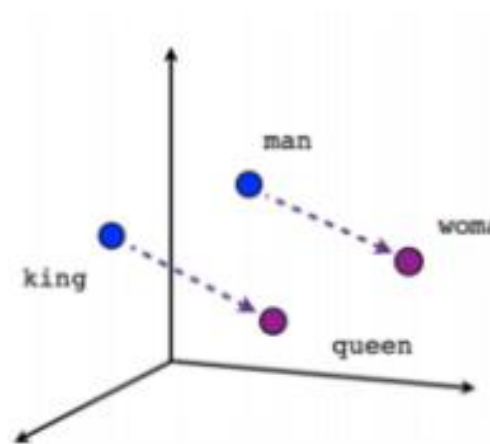
$$\text{King} - \text{man} + \text{woman} = \text{Queen}$$

- Se entrena con modelos de Deep Learning
- Aparecen relaciones semánticas en el espacio multidimensional
- Word2Vec, Glove...

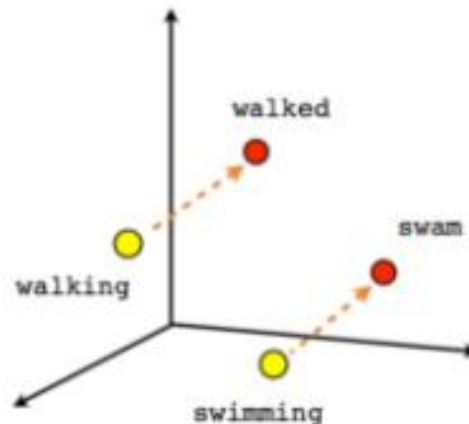
2. Técnicas y terminología básica de NLP

Semántica – Word Embeddings

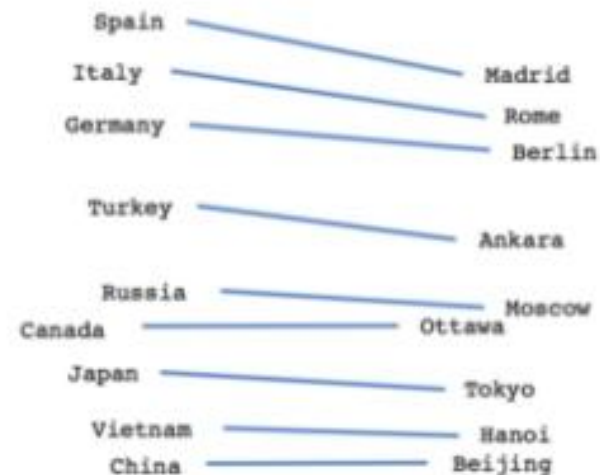
El texto es unidimensional. ¿Y si proyectamos las palabras a espacios multidimensionales?



Male-Female



Verb tense

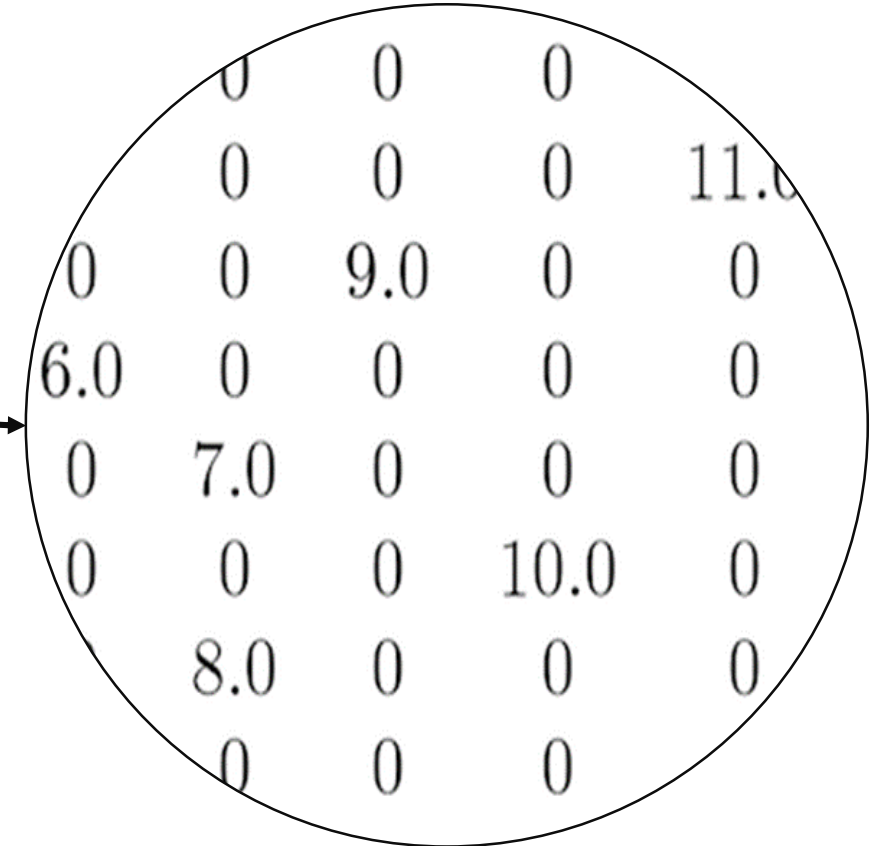
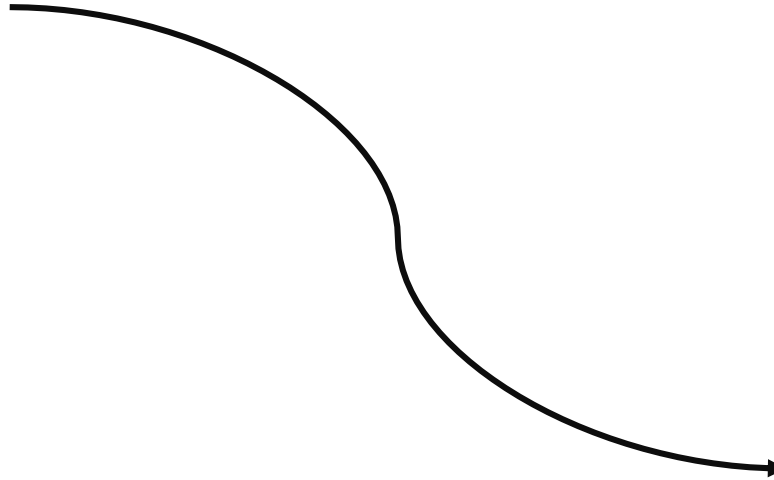
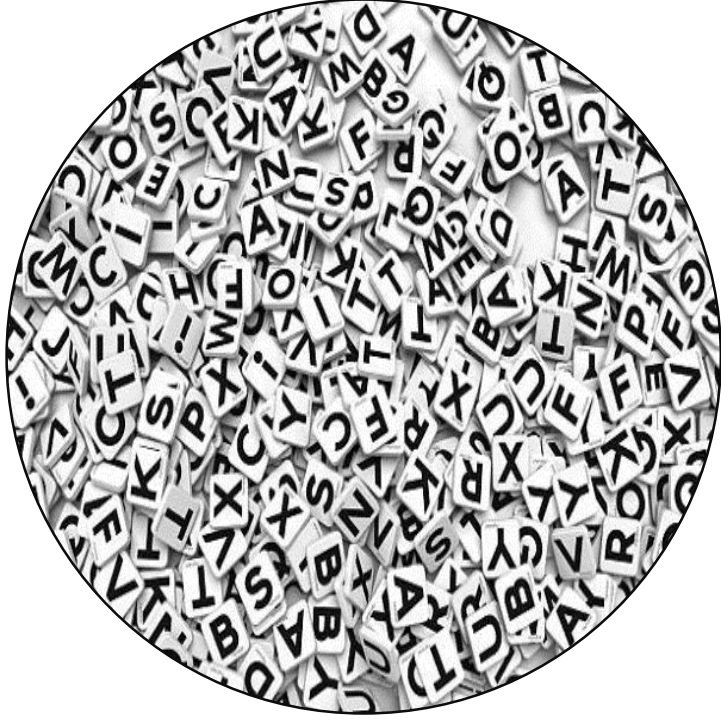


Country-Capital

3. Representación numérica de documentos textuales



3. Representación numérica de documentos textuales



3. Representación numérica de documentos textuales

Bolsa de Palabras – *Bag of Words*

Método de representación del texto basado en la aparición de palabras de un vocabulario en un documento

Se llama bolsa de palabras porque se pierde el orden de aparición de estas.

Vocabulario del
corpus

Métrica

3. Representación numérica de documentos textuales

Bolsa de Palabras – *Bag of Words*

1. Corpus

“Yo quiero agua”

“Yo quiero cocacola”

“Yo quiero agua y un agua”

“Yo no quiero vino”

“Yo quiero un entrecot”

2. Vocabulario

“agua”

“cocacola”

“entrecot”

“no”

“quiero”

“un”

“vino”

”yo”

3. Vectorización con métrica

[1 0 0 0 1 0 0 1]

[0 0 0 1 1 0 1 1]

[2 0 0 0 1 1 0 1]

[0 0 0 1 1 0 1 1]

[0 0 1 0 1 1 0 1]

3. Representación numérica de documentos textuales

Bolsa de Palabras – *Bag of Words*

Recomendaciones:

- Transformación de tokens a minúscula
- Ignorar los signos de puntuación
- Ignorar stop-words
- Normalizar palabras
- Lematizar o Stemming

3. Representación numérica de documentos textuales

Bolsa de Palabras – *Bag of Words* - Métricas

Conteo

Simple y sencillo de realizar

Mayor número es igual a mayor importancia, aunque no es así si aparece en todos los documentos

TF-IDF

Term Frequency-Inverse Document Frequency

Doble cómputo para vectorizar

Palabras que aparecen en todos los documentos son penalizadas

Se calcula IDF para token w .

$$IDF(w) = \ln \frac{N}{n_w}$$

3. Representación numérica de documentos textuales

Bolsa de Palabras – *Bag of Words* - Métricas

"Yo quiero agua"	=	[1.92	0	0	0	1	0	0	1]
"Yo quiero cocacola"	=	[0	2.61	0	0	1	0	0	1]
"Yo quiero agua y agua"	=	[3.83	0	0	0	1	1.92	0	1]
"Yo no quiero vino"	=	[0	0	0	2.61	1	0	2.61	1]
"Yo quiero un entrecot"	=	[0	0	2.61	0	1	1.92	0	1]

