

Clase 20: Procesamiento del lenguaje natural (NPL)

Laboratorio de datos, FCEyN, 4/6/2021



Motivación

Imaginemos que...



Motivación

Imaginemos que...

vimos estas dos grandes clases



Motivación

Imaginemos que...



vimos estas dos grandes clases

T. Cicchini y S. Pinto

Clase N° 18

Interacción con páginas webs
mediante APIs

Laboratorio de Datos

28 de Mayo 2021

T. Cicchini y S. Pinto

Clase N° 19

Web scraping

Laboratorio de Datos

31 de Junio 2021

Motivación

Imaginemos que...



vimos estas dos grandes clases

T. Cicchini y S. Pinto

Clase N° 18

— Interacción con páginas webs mediante APIs —

Laboratorio de Datos 28 de Mayo 2021

T. Cicchini y S. Pinto

Clase N° 19

— Web scraping —

Laboratorio de Datos 31 de Junio 2021

ya somos **ases** en la minería de datos

Motivación



Motivación

qué hacemos con este texto





Comunicar y Compartir información



Comunicar y Compartir información

LENGUAJE



Comunicar y Compartir información



LENGUAJE



ALPHABETS

字母 वर्णमाला ALFABETOS
الأبجدية எழுத்துக்கள்

Comunicar y Compartir información



LENGUAJE



ALPHABETS

字母 वर्णमाला ALFABETOS
الأبجدية எழுத்துக்கள்



PALABRAS

Comunicar y Compartir información



LENGUAJE



ALPHABETS

字母 वर्णमाला ALFABETOS
الأبجدية எழுத்துக்கள்



PALABRAS formen ORACIONES

Comunicar y Compartir información



Reglas conocidas

LENGUAJE



ALPHABETS

字母 वर्णमाला ALFABETOS
الأبجدية எழுத்துக்கள்



PALABRAS forman ORACIONES

Comunicar y Compartir información



LENGUAJE



ALPHABETS

字母 वर्णमाला ALFABETOS
الأبجدية எழுத்துக்கள்



PALABRAS formen ORACIONES

Reglas conocidas



GRAMÁTICA

Motivación

qué hacemos con este texto



son datos **no-estructurados**:
-no tienen un modelo predefinido

ej: no hay una forma natural de ordenarlos

Motivación

qué hacemos con este texto



son datos **no-estructurados**:
-no tienen un modelo predefinido

ej: no hay una forma natural de ordenarlos

Text mining/Text analytics: es el proceso de generar información útil de textos de lenguaje natural

Motivación

Text mining/Text analytics: es el proceso de generar información útil de textos de lenguaje natural

Motivación

Estructurar los datos



Text mining/Text analytics: es el proceso de generar información útil de textos de lenguaje natural

Motivación

Encontrar patrones



Estructurar los datos



Text mining/Text analytics: es el proceso de generar información útil de textos de lenguaje natural

Motivación

Encontrar patrones



Estructurar los datos



Text mining/Text analytics: es el proceso de generar información útil de textos de lenguaje natural

Motivación



Texto es el tipo de dato que todos (casi todos) generamos más rápidamente y en grandes cantidades

Motivación



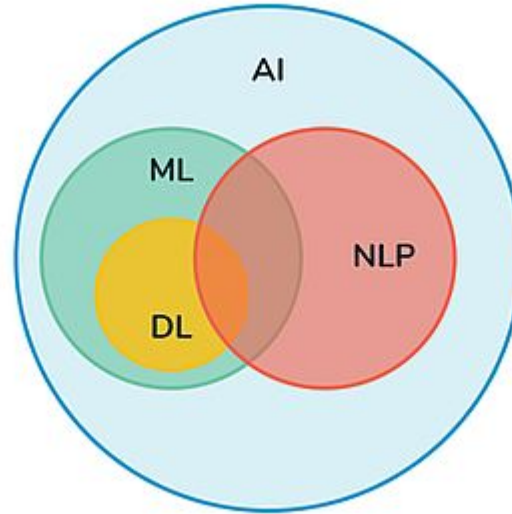
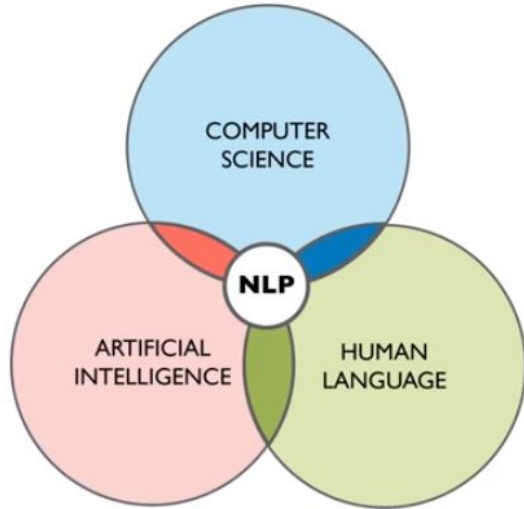
Texto es el tipo de dato que todos (casi todos) generamos más rápidamente y en grandes cantidades



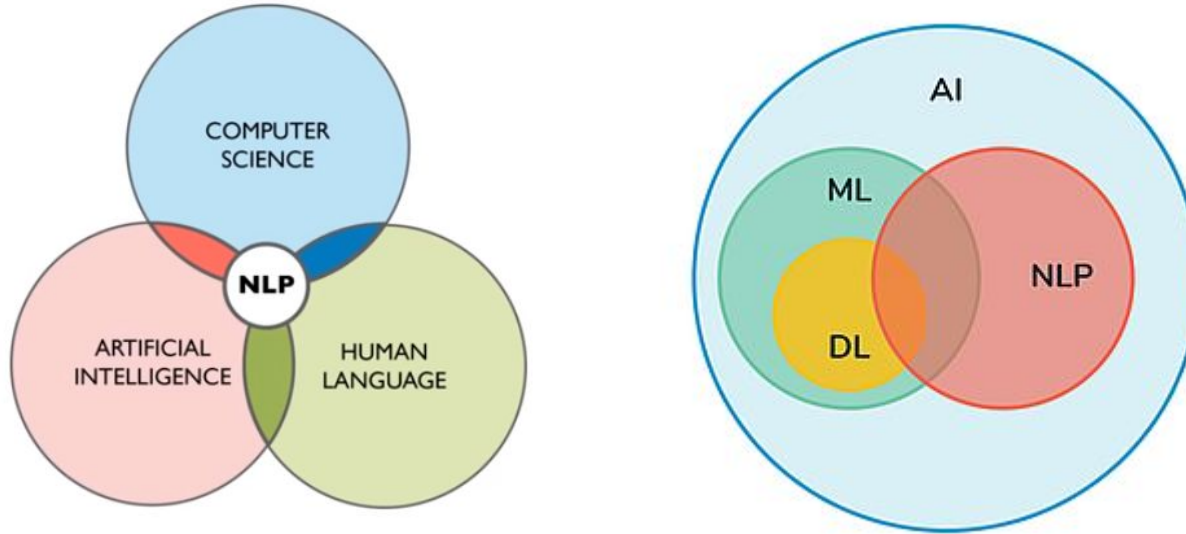
Text mining busca *transformar* **textos** en **datos** **para analizar** mediante la aplicación del Procesamiento del Lenguaje Natural (PLN/NLP)

Procesamiento del lenguaje natrual (PLN)

Procesamiento del lenguaje natrual (PLN)



Procesamiento del lenguaje natrual (PLN)



PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Procesamiento del lenguaje natrual (PLN)

PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Procesamiento del lenguaje natrual (PLN)



“La materia es muy interesante”

“Odio las aceitunas”

Análisis de
sentimiento

PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Procesamiento del lenguaje natrual (PLN)



“La materia es muy interesante”
“Odio las aceitunas”

Análisis de
sentimiento



Reconocimiento
de voz

PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Procesamiento del lenguaje natrual (PLN)



“La materia es muy interesante”
“Odio las aceitunas”

Análisis de
sentimiento

Chatbot



Reconocimiento
de voz

PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Procesamiento del lenguaje natrual (PLN)



“La materia es muy **interesante**”
“**O**dio las aceitunas”

Análisis de
sentimiento

Chatbot



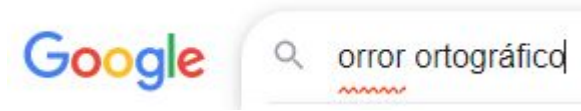
Reconocimiento
de voz

Traductor de
máquina



PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

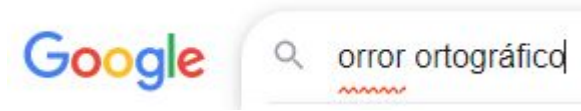
Procesamiento del lenguaje natrual (PLN)



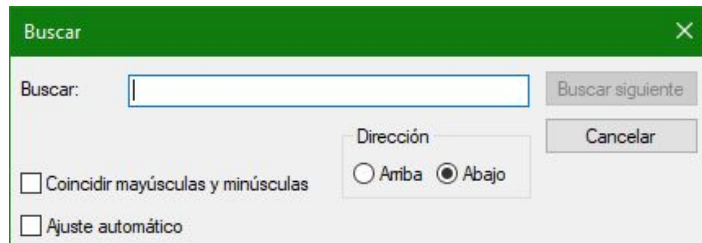
Corrector
ortográfico

PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Procesamiento del lenguaje natrual (PLN)



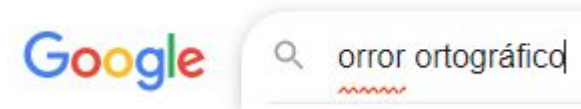
Corrector
ortográfico



Buscador de
palabras

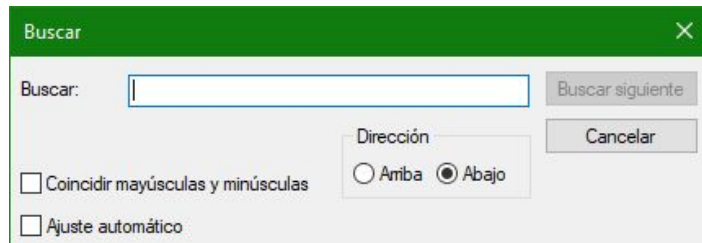
PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Procesamiento del lenguaje natrual (PLN)



Corrector
ortográfico

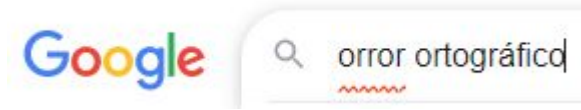
Predictor de
palabras



Buscador de
palabras

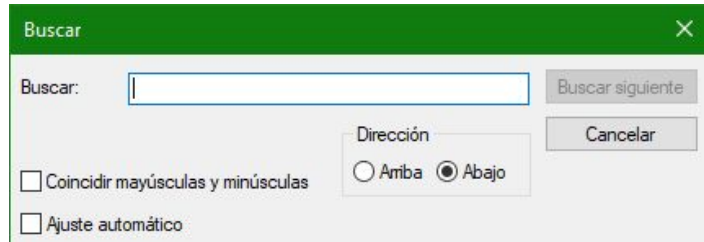
PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Procesamiento del lenguaje natrual (PLN)



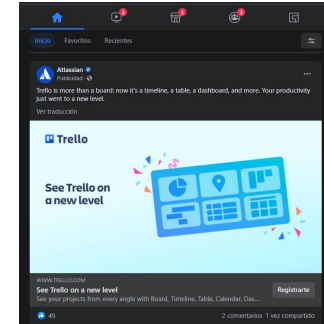
Corrector
ortográfico

Predictor de
palabras



Buscador de
palabras

Publicidad
dirigida



PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

Hay sesgo en las aplicaciones de PLN?

Virtualmente, cualquier lengua humana puede ser tratada por los ordenadores. Lógicamente, limitaciones de *interés económico o práctico* hace que solo las **lenguas más habladas o utilizadas en el mundo digital tengan aplicaciones en uso**.

Pensemos en cuántas lenguas hablan Siri (**20**) o Google Assistant (**8**). El inglés, español, alemán, francés, portugués, chino, árabe y japonés (no necesariamente en este orden) son las que cuentan con más aplicaciones que las entienden. Google Translate es la que más lenguas trata, superando el centenar... pero hay entre **5000** y **7000** lenguas en el mundo.

Aplicación básica de PLN

Aplicación básica de PLN



Tokenization



Stemming



Lemmatization

Lemmatization



POS Tags



Named Entity Recognition



Chunking

Aplicación básica de PLN



Tokenization

Aplicación básica de PLN



Tokenization

Preprocesamiento de texto:

Aplicación básica de PLN



- Limpieza del texto (texto característico de los formatos)
- Expandir contracciones típicas del inglés (ej. simple: “can’t” -> “can not”, hay algunas ambigüas que dependen de la oración)
- Cambiar mayúsculas por minúsculas
- Eliminar signos de puntuación (#\$%&?!.,)

Aplicación básica de PLN



Tokenization

Preprocesamiento de texto:

- Limpieza del texto (texto característico de los formatos)
- Expandir contracciones típicas del inglés (ej. simple: “can’t” -> “can not”, hay algunas ambigüas que dependen de la oración)
- Cambiar mayúsculas por minúsculas
- Eliminar signos de puntuación (#\$%&?!.,)

Dividir el texto en unidades más pequeñas. De párrafo a oraciones. De oración a palabras.

Aplicación básica de PLN



Tokenization

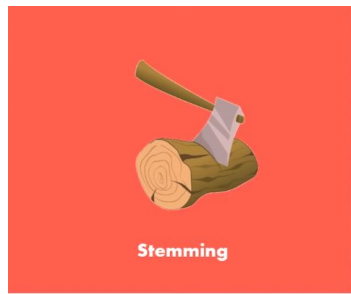
Preprocesamiento de texto:

- Limpieza del texto (texto característico de los formatos)
- Expandir contracciones típicas del inglés (ej. simple: “can’t” -> “can not”, hay algunas ambigüas que dependen de la oración)
- Cambiar mayúsculas por minúsculas
- Eliminar signos de puntuación (#\$%&?!.,)

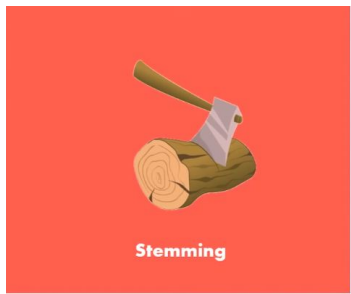
Dividir el texto en unidades más pequeñas. De párrafo a oraciones. De oración a palabras.

[“dividir el texto en unidades más pequeñas”, “de párrafo a oraciones”, ...]
[“dividir”, “el”, “texto”, “en”, ...]

Aplicación básica de PLN



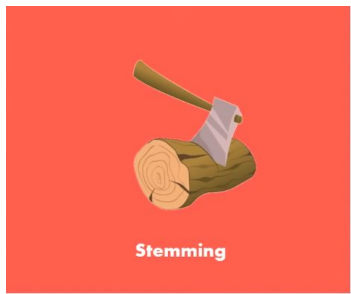
Aplicación básica de PLN



Hay palabras que provienen de una misma ***raíz***.

Quiero cortar las palabras para quedarme únicamente con la raíz.

Aplicación básica de PLN

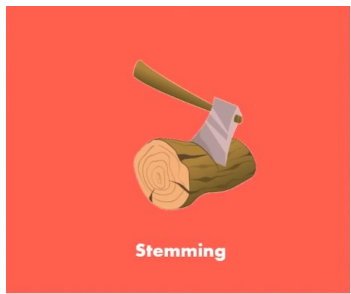


Hay palabras que provienen de una misma ***raíz***.

Quiero cortar las palabras para quedarme únicamente con la raíz.

“Affectation” “Affects” “Affections” “Affected” “Affection” “Affecting”

Aplicación básica de PLN

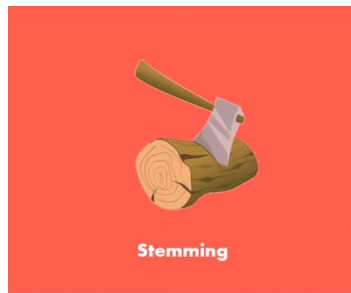


Hay palabras que provienen de una misma **raíz**.

Quiero cortar las palabras para quedarme únicamente con la raíz.

“**Affect**ation” “**Affect**s” “**Affect**ions” “**Affect**ed” “**Affect**ion” “**Affect**ing”

Aplicación básica de PLN



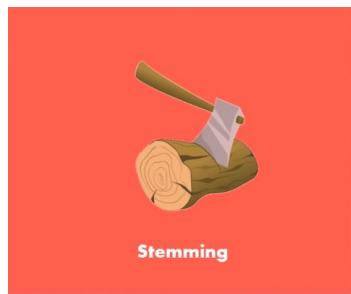
Hay palabras que provienen de una misma **raíz**.

Quiero cortar las palabras para quedarme únicamente con la raíz.

“Affectation” “Affects” “Affections” “Affected” “Affection” “Affecting”

“Affect” “Affect” “Affect” “Affect” “Affect” “Affect”

Aplicación básica de PLN



Hay palabras que provienen de una misma **raíz**.

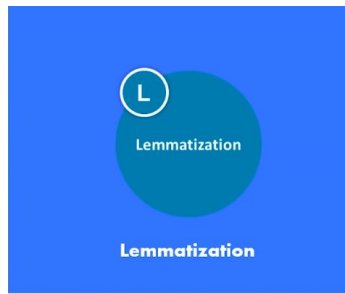
Quiero cortar las palabras para quedarme únicamente con la raíz.

“Affectation” “Affects” “Affections” “Affected” “Affection” “Affecting”

“Affect” “Affect” “Affect” “Affect” “Affect” “Affect”

Inconveniente: No funciona siempre. Hay palabras que su raíz **depende del contexto** de la oración. Se requiere un análisis morfológico.

Aplicación básica de PLN



Aplicación básica de PLN



En vez de buscar la raíz de la palabra se busca su **lema**

Aplicación básica de PLN



En vez de buscar la raíz de la palabra se busca su **lema**

el lema es la palabra que nos encontraríamos en el diccionario tradicional:

Aplicación básica de PLN



En vez de buscar la raíz de la palabra se busca su **lema**

el lema es la palabra que nos encontraríamos en el diccionario tradicional:

- singular para sustantivos (“Mesa” -> “Mesas”)
- masculino singular para adjetivos (“guapas” -> “guapo”)
- infinitivo para verbos (“dije”, “diré”, “dijéramos” -> “decir”)

Aplicación básica de PLN



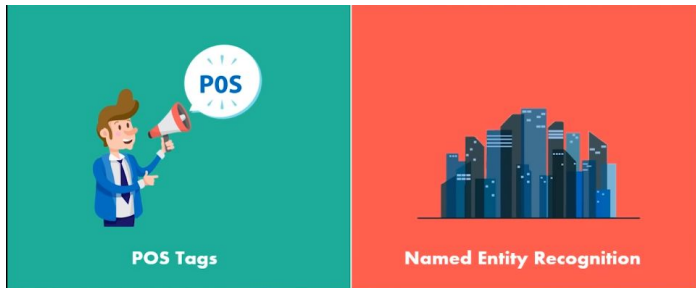
En vez de buscar la raíz de la palabra se busca su **lema**

el lema es la palabra que nos encontraríamos en el diccionario tradicional:

- singular para sustantivos (“Mesa” -> “Mesas”)
- masculino singular para adjetivos (“guapas” -> “guapo”)
- infinitivo para verbos (“dije”, “diré”, “dijéramos” -> “decir”)

Es *similar* a **stemming** ya que mapea muchas palabras a una sola pero el resultado de **lemmatization** es una palabra mientras que en stemming puede no serlo

Aplicación básica de PLN



Aplicación básica de PLN



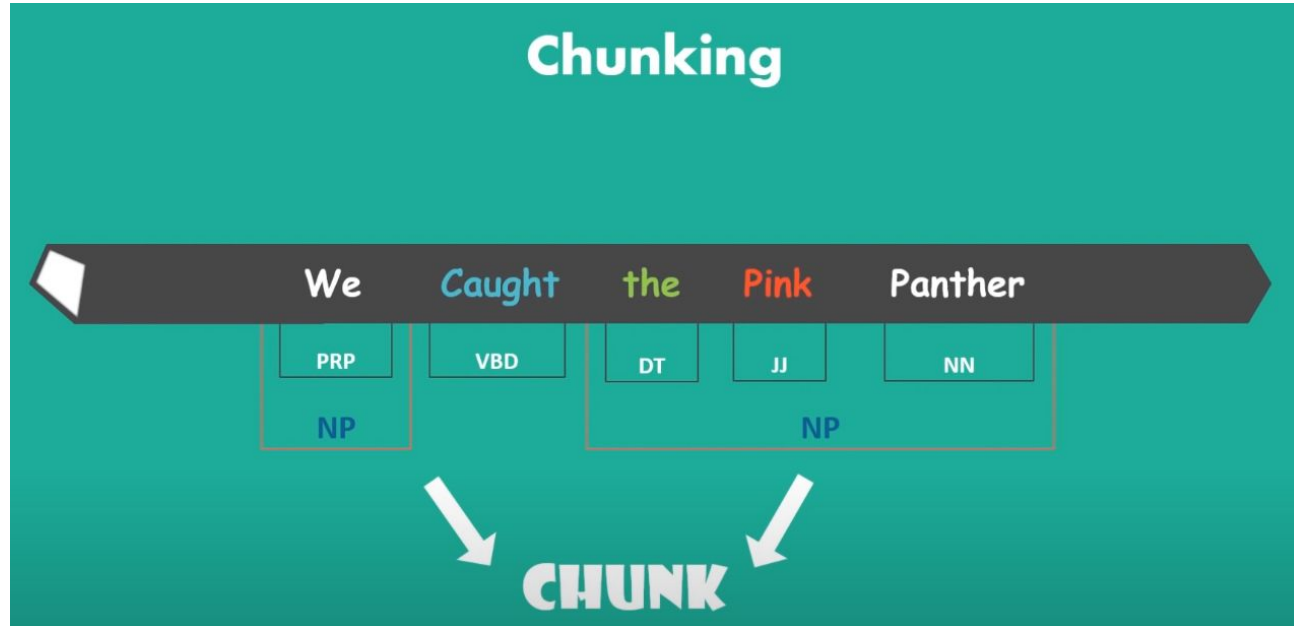
Part of Speech taggin:

Identificamos el rol gramatical de cada palabra (si es un verbo, adjetivo, sustantivo, etc.)

Named Entity Recognition:

Reconocer entidades como (empresas, personas, películas, lugares)

Aplicación básica de PLN



Aplicación básica de PLN



Tokenization



Stemming



Lemmatization



POS Tags



Named Entity Recognition



Chunking

Matriz de frecuencia/Bolsa de palabras (BoW)

Matriz de frecuencia/Bolsa de palabras (BoW)

Oración 1: “Tengo suculentas en mi escritorio.”

Oración 2: “Mi escritorio es de madera y tiene suculentas encima.”

Matriz de frecuencia/Bolsa de palabras (BoW)

Oración 1: “Tengo suculentas en mi escritorio.”

Oración 2: “Mi escritorio es de madera y tiene suculentas encima.”

tengo | suculentas | en | mi | escritorio | es | de | madera | y | tiene | encima

Matriz TF-IDF (Term Frequency - Inverse Document frequency)

Medida de la **importancia de cada término (palabra) en nuestro corpus**

Matriz TF-IDF (Term Frequency - Inverse Document frequency)

Medida de la **importancia de cada término (palabra) en nuestro corpus**

$$w_d = f_{w,d} \cdot \log \left(\frac{|D|}{f_{w,D}} \right)$$

Matriz TF-IDF (Term Frequency - Inverse Document frequency)

Medida de la importancia de cada término (palabra) en nuestro corpus

para una palabra w en el documento d

$$w_d = f_{w,d} \cdot \log \left(\frac{|D|}{f_{w,D}} \right)$$

frecuencia de w en d

tamaño de colección de documentos
(cantidad de documentos)

cantidad de documentos donde aparece w

A tener en cuenta

Cuando trabajamos con texto de chats o posteos nos solemos encontrar con:

- Errores de tipeo
- Palabras con letras repetidas “Siiiiiii”
- Abreviaturas
- Usar números para reemplazar palabras “d1”
- emoticones
- palabras en distintos idiomas

Depende del análisis que queremos hacer se puede atacar esto de distintas formas

Notebook

Notebook



casciari

Enviar mensaje

...

276 publicaciones

163k seguidores

138 seguidos

Hernán Casciari

casuari@gmail.com · youtube.com/hcasciari · facebook.com/EditorialOrsai · hernancasciari.com

PUBLICACIONES

REELS

IGTV

ETIQUETADAS

Hoy nace «Streaming con Apuesta de libros». Ustedes son muy jóvenes, pero en la semifinal de la Copa América de 2015 hice una apuesta en Twitter: «Si me compran un libro, les regalo otro por cada gol de Argentina esta noche». Me compraron muchísimos libros y yo me creí un gordito marketinero, pero esa noche Argentina le ganó a Paraguay 6 a 1 (Rojo, Pastore, Agüero, Higuaín y Di María «»). Yo todavía vivía en Barcelona y en el cuarto gol supe que el chiste me iba a salir caro... Al día siguiente tuve que imprimir y mandar por DHL paquetes de seis libros a dos mil personas del otro lado del océano. La pérdida económica fue de 113.000 USD según mi contador, que renunció ese día. **Esta noche vuelvo al ruedo para tomar revancha.** En el partido de eliminatorias de las 21 hs (Argentina vs. Chile) volveré a utilizar el sistema que me hizo morder el polvo. Vos elegís mi streaming más un libro a elección y yo te mando tantos libros más por goles de diferencia de Argentina. Vengan. Desvalijéme de nuevo... si pueden.





Notebook



Hoy nace «Streaming con Apuesta de libros». Ustedes son muy jóvenes, pero en la semifinal de la Copa América de 2015 hice una apuesta en Twitter: «Si me compran un libro, les regalo otro por cada gol de Argentina esta noche». Me compraron muchísimos libros y yo me creí un gordito marketinero, pero esa noche Argentina le ganó a Paraguay 6 a 1 (Rojo, Pastore, Agüero Higuaín y Di María x2). Yo todavía vivía en Barcelona y en el cuarto gol supe que el chiste me iba a salir caro... Al día siguiente tuve que imprimir y mandar por DHL paquetes de seis libros a dos mil personas del otro lado del océano. La pérdida económica fue de 113.000 USD según mi contador, que renunció ese día. **Esta noche vuelvo al ruedo para tomar revancha.** En el partido de eliminatorias de las 21 hs (Argentina vs. Chile) volveré a utilizar el sistema que me hizo morder el polvo. Vos elegís mi streaming más un libro a elección y yo te mando tantos libros más por goles de diferencia de Argentina. **Vengan. Desvalíjenme de nuevo... si pueden.**

Hoy nace «Streaming con Apuesta de libros». Ustedes son muy jóvenes, pero en la semifinal de la Copa América de 2015 hice una apuesta en Twitter: «Si me compran un libro, les regalo otro por cada gol de Argentina esta noche». Me compraron muchísimos libros y yo me creí un gordito marketinero, pero esa noche Argentina le ganó a Paraguay 6 a 1 (Rojo, Pastore, Agüero Higuaín y Di María x2). Yo todavía vivía en Barcelona y en el cuarto gol supe que el chiste me iba a salir caro... Al día siguiente tuve que imprimir y mandar por DHL paquetes de seis libros a dos mil personas del otro lado del océano. La pérdida económica fue de 113.000 USD según mi contador, que renunció ese día. **Esta noche vuelvo al ruedo para tomar revancha.** En el partido de eliminatorias de las 21 hs (Argentina vs. Chile) volveré a utilizar el sistema que me hizo morder el polvo. Vos elegís mi streaming más un libro a elección y yo te mando tantos libros más por goles de diferencia de Argentina. **Vengan. Desvalíjenme de nuevo... si pueden.**