



UNIVERSIDAD TECNOLÓGICA NACIONAL  
FACULTAD REGIONAL CÓRDOBA

# Diplomatura en Data Science Aplicada



UNIVERSIDAD TECNOLÓGICA NACIONAL  
FACULTAD REGIONAL CÓRDOBA

# Procesamiento de Texto

# Agenda

- Expresiones regulares (RegEx)
- Procesamiento de texto
  - tokenización
  - normalización
  - segmentación



# Expresiones regulares (aka RegEx)

[^]\*?@[^]\*?\. [^]\*

*“Una expresión regular es una secuencia especial de caracteres que nos ayudan a encontrar una cadena o conjunto de cadenas de texto”*

Todos los lenguajes de programación traen librerías para el manejo de expresiones regulares.

Incluso UNIX, provee el comando tr para manipulación de texto a través de opciones basadas en regex.

Algunos usos más comunes para RegEx:

- Reconocimiento de patrones.
- Web scrapping.
- Extracción de datos.
- Datascience, etc.

# Expresiones regulares (aka RegEx)

`[^]*?@[^]*?\. [^]*`

El uso de RegEx no está libre de errores, por lo que es algo que tenemos que evaluar.

- Disminuimos los Falsos Positivos (FP)?
- Aumentamos los Falsos Negativos (FN)?

Esto va impactar en la métrica que busquemos mejorar.

Podemos usar RegEx en Python gracias a la librería **re**.

Funciones más interesantes:

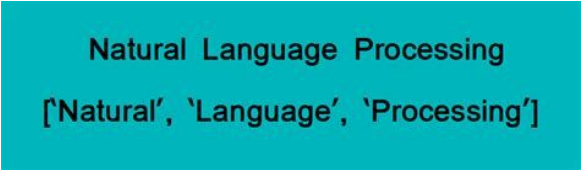
- `split`
- `findall`
- `sub`

# Procesamiento de Texto

## Tokenización

# Tokenización

*“Es un proceso el cual demarca y posibilita la clasificación en secciones de una cadena de **strings**. El resultado es un conjunto de **tokens** que son pasados posteriormente para ser sometido a otros procesos o transformaciones.”*



Natural Language Processing  
['Natural', 'Language', 'Processing']

**Dentro del proceso de PLN vamos a encontrar los siguientes pasos:**

- *Segmentación o tokenización de nuestro corpus.*
- Normalizar palabras o tokens.
- Segmentar las oraciones del corpus.

# Tokenización - Cantidad de palabras

¿Cuántas palabras hay?

*“En un lugar de la Mancha, de cuyo nombre no quiero acordarme”*

**Types (V)**: elemento unívoco dentro de un vocabulario.

**Token (N)**: instancia de un tipo dentro de un texto.

*“Now cracks a noble heart. Good-night, sweet prince; and flights of angels sing thee to thy rest.”*



# Tokenización - Cantidad de palabras

	Tokens	Types
El Quijote de la mancha	~2.000.000	~380.000
La tragedia de Hamlet	~190.000	~30.000

Los idiomas o lenguajes van evolucionando, no son estáticos.



## Problemas comunes:

- token o palabras fuera del vocabulario (OOV).
- Contracciones de palabras propias del lenguaje.
- Nombres compuestos. Buenos Aires (es 1 o 2).
- Manejar vocabularios muy grandes, trae aparejado problemas de espacio.

¿Cuan grande será el vocabulario, de Google Assistance o Siri?

# Procesamiento de Texto

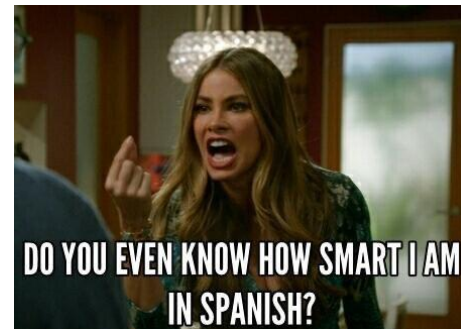
## Normalización

# Normalización

## ¿Porque normalizar?

La evolución en cómo nos comunicamos y expresamos cambió respecto al siglo pasado.

- En el año 2011 en US, se enviaron ~196.9 billion de sms (12.5 billion in 2006).
- Brandwatch mostró que en solo en el mes de Mayo de 2016 fueron enviados ~500 millones de tweets cada día (aprox. 6.000 tweets por seg).
- Las palabras pueden significar/estar escritas/habladas acuerdo a su contexto y situación.



*“Buscamos agrupar las distintas formas o inflexiones de una misma palabra en un mismo lemma o forma del diccionario.”*

- Equivalencia de clases: U.T.N. F.R.C → UTN FRC
- Expansión asimétrica: jarra → jarras, jarrón
- Llevar a minúscula por facilidad, pero existen algunas excepciones de acuerdo el caso.

# Normalización - Lemmatization

*Buscamos reducir las variaciones de las palabras hacia su forma “raíz”. Basándose en:*

- Análisis morfológico de la palabra
- Su importancia en el diccionario

auto, autos, automovil → auto

flores, floripondio, flor → flor

**Entonces:**

Los autos corren carreras y pisan flores → los auto corren carrera y pisan flor

# Normalización - Stemming

*Es una forma simplificada de lematización, la cual reduce palabras a su “raíz” cortando las mismas basándose en prefijos y sufijos.*

Son algoritmos que van recorriendo distintos pasos hasta llevar a la palabra raíz.

monastery → monasteri

walking → walk

sing → sign

automation, automatic → automat

# Normalización - Casi final

- Pasar todo el corpus a minúscula
- Filtrar el corpus a través de stopwords
- Hacer un análisis sobre aquellos “términos”, que nos pueden o no servir.

Finalizado este proceso, tendremos un Bag of Words (BoW)

