{ Minería de textos }

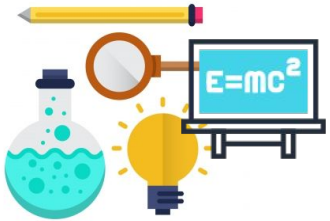
PROCESO DE ANÁLISIS

{PREGUNTAS}

{Datos}

{ EDA }

{ Modelo }



{ Visualización}

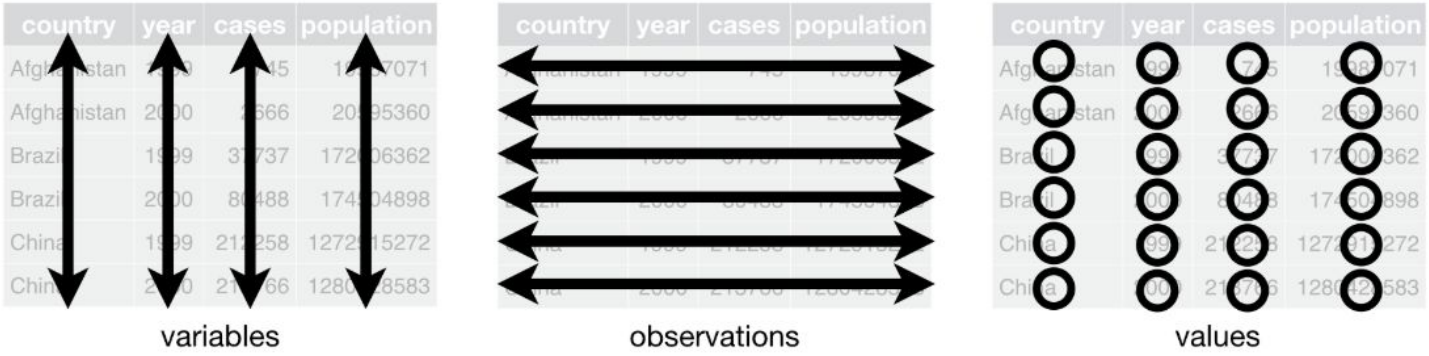
{ Toma de decisiones }

{ Conocimiento }

{ TIDY DATA }

Cada conjunto ordenado tiene que cumplir tres reglas:

- Cada variable debe tener una columna.
- Cada observación debe tener una fila.
- Cada valor debe tener su propia celda.



{ TEXT - MINING }

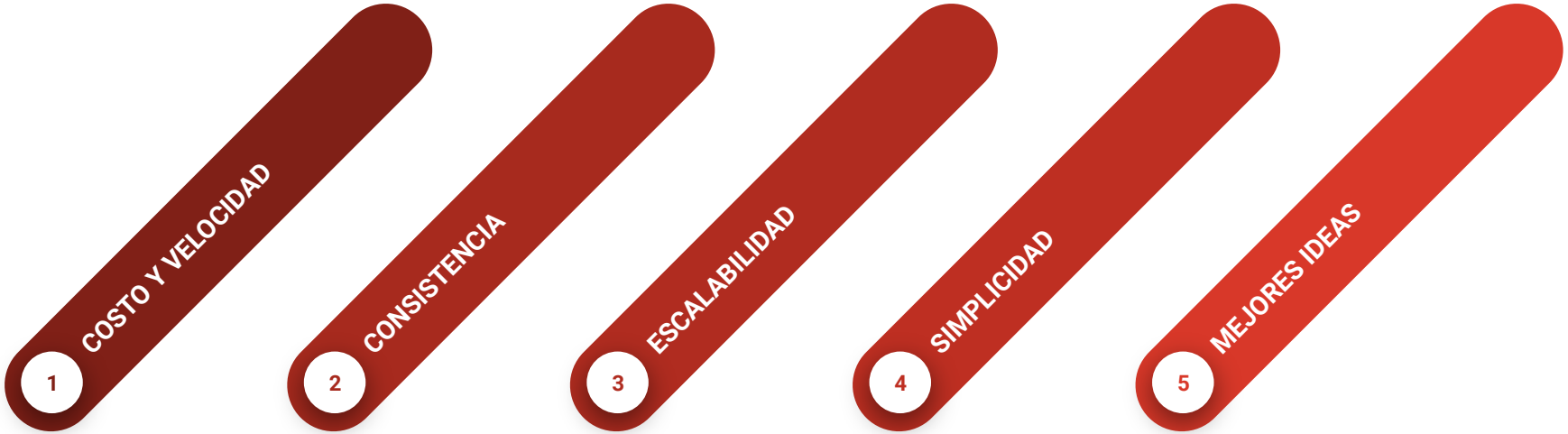
Text-Mining **es el proceso de analizar colecciones de texto con el objetivo de capturar conceptos y temas clave**.
Esta técnica permite detectar patrones que mediante una exploración tradicional de los datos muchas veces no se podría realizar debido a que las relaciones son demasiado complejas o por el volumen de datos que se maneja.

{ APLICACIONES }

Aplicaciones que puede tener:

- La extracción de información
- El análisis de sentimientos o minería de opiniones
- La clasificación documental
- La elaboración de resúmenes
- La extracción de conocimiento.

{ BENEFICIOS }


1
COSTO Y VELOCIDAD
2
CONSISTENCIA
3
ESCALABILIDAD
4
SIMPLICIDAD
5
MEJORES IDEAS

{ PROCESO }

El proceso para poder analizar un texto cuenta con 6 pasos...

1 — 2 — 3 — 4 — 5 — 6

Obtener los datos — **Limpieza** — **Ordenar** — **Procesar** — **Visualizar** — **Analizar**

{ Limpieza de datos }

1) Se pasan las palabras a minúsculas
2) Se eliminan signos de puntuación.
3) Se modifican las vocales por tildes por vocales sin tildes.
4) Se elimina la última letra s.

{ TOKEN }

Un token es una unidad significativa de texto, es una palabra, que estamos interesados en usar para el análisis, y la tokenización es el proceso de dividir el texto en tokens.
Para la minería de texto ordenado, el **token** que se almacena en cada fila es a menudo una sola palabra, pero también puede ser un n-gramo, oración o párrafo.

{ STOP WORDS }

Las stop word o palabras vacías como se denominan en español son todas las palabras que carecen de un significado concreto o por si solas.
Ejemplos de palabras vacías pueden ser:

- Articulos
- Preposiciones
- Conjunciones
- Pronombres
- Etc

Por lo cual en la minería de texto estas palabras deben ser filtradas ya que no tienen un significado dentro del análisis, son solo conectores.

{ STOP WORDS }

TECNO

Instagram eliminará más rápidamente los perfiles que infrinjan sus normas de uso

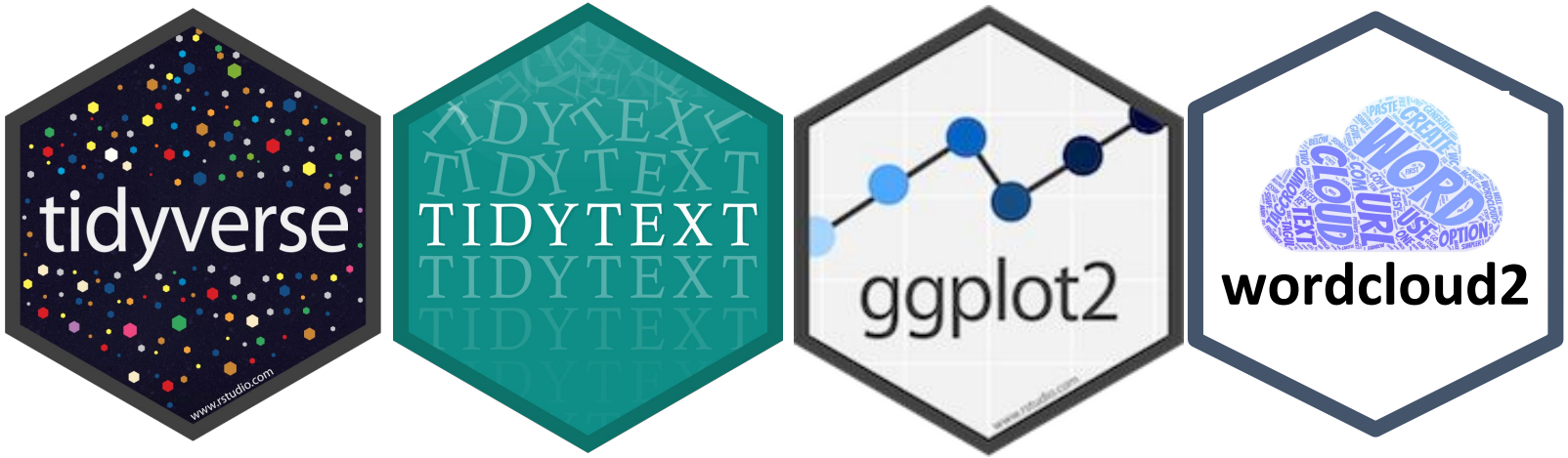
La red social anunció cambios vinculadas a la inhabilitación de cuentas. La compañía dice que busca implementar sus políticas “de un modo más consistente” y que de este modo esperan que “las personas asuman responsabilidad por lo que comparten” en la plataforma

{ TIDY TEXT }

Definimos el formato de texto ordenado como **una tabla con un token por fila.**

Un token es una unidad significativa de texto, como una palabra, que estamos interesados en usar para el análisis, y la tokenización es el proceso de dividir el texto en tokens

{ PAQUETES }



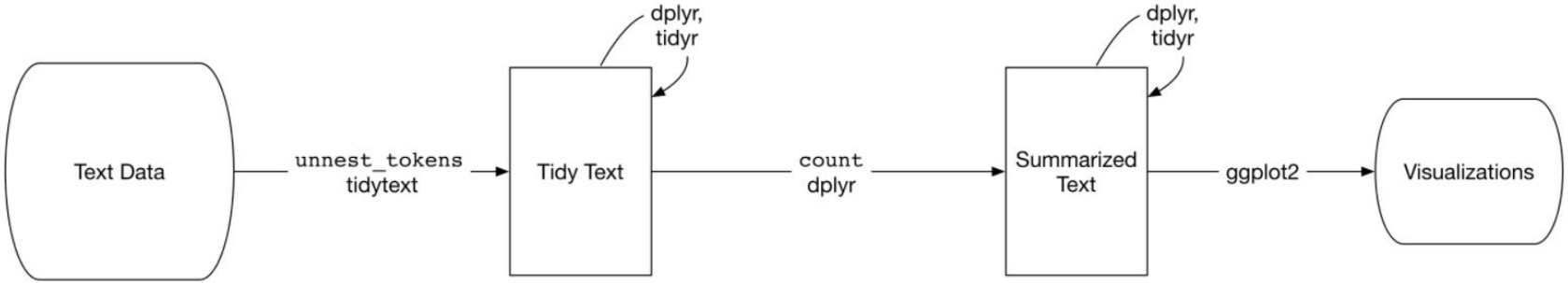
{ N-grams }

Es muchas veces interesante poder ver las relaciones entre las palabras y ver cómo se vinculan las mismas.

Los n-grams son secuencias de palabras consecutivas que se repiten en un texto .

Si se establece un n=2, lo que estamos buscando son dos palabras consecutivas, por lo cual a esta estructura se la denominada bigramas.

{ PROCESO }



https://www.tidytextmining.com/

{ Generar tokens }

```
Df = df_original %>%
        select(id,category,description) %>%
        unnest_tokens(palabra,description)
```

Nueva columna con tokens

Columna que contiene el texto

{ Generar tokens }

```
Df = Df %>%
        anti_join(sw,by="palabra")
```

Dataset con las stop word

Key para cruzar los dos dataframes

{ bigrams }

```
Df = df_original %>%
        select(id,category,description) %>%
          unnest_tokens(    Palabra,
                            description,
                            token = "ngrams",
                            n = 2)
```

n-gram

{ Ejercicio I }

1) Creamos una variable de texto y cargamos el titular de infobae:
“Instagram eliminará más rápidamente los perfiles que infrinjan sus normas de uso.”

2) Debemos eliminar las tildes (utilice la funcion gsub)

3) Colocamos los términos en minúsculas (utilice la función tolower)

4) Elimine la stop words utilizando el archivo “sw.cvs”

https://raw.githubusercontent.com/r0mymendez/FuncionesR/master/Text/sw.csv

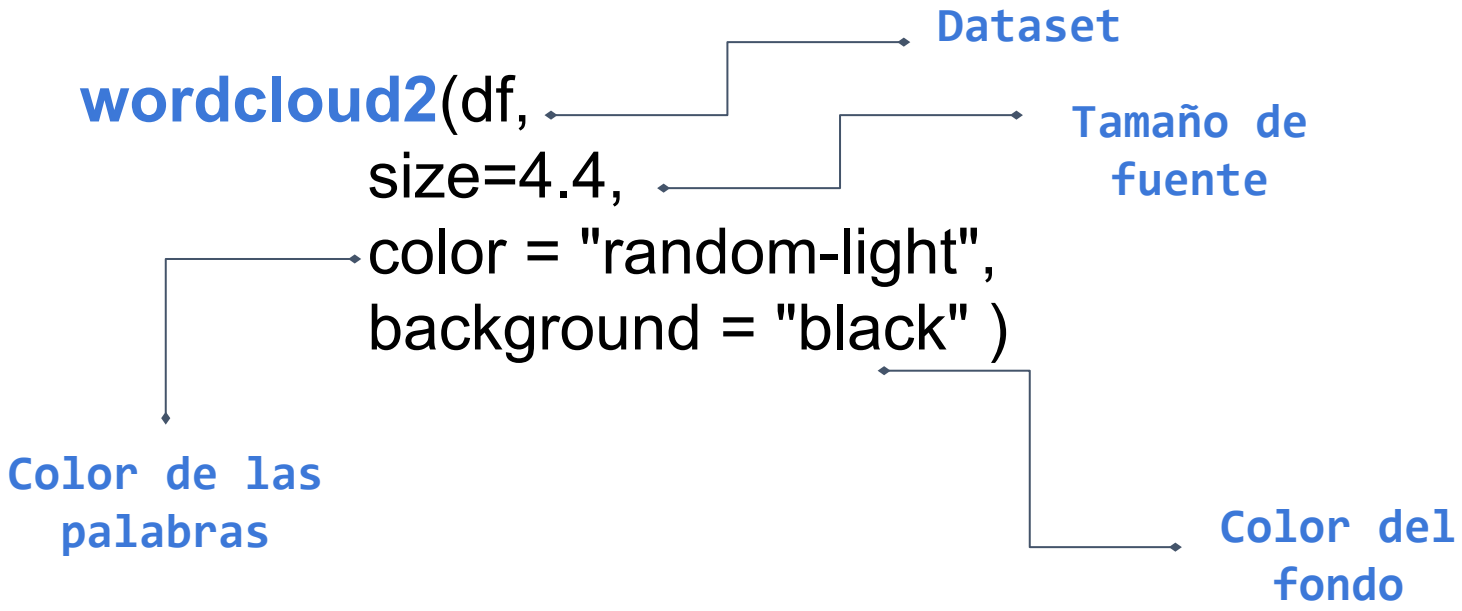
5) Genere una tabla con los tokens.

library(wordcloud2)

Una nube de palabras es una representación visual de datos de tipo carácter.
Las palabras se visualizan con diferente tamaño dependiendo de la frecuencia con la que aparecen en un texto.

Es un recurso que facilita la identificación de términos clave o mas significativos.

{ wordcloud2 }


wordcloud2(df,
size=4.4,
color = "random-light",
background = "black")
Dataset
Tamaño de fuente
Color de las palabras
Color del fondo