Minería de datos

Datos complejos
Web mining

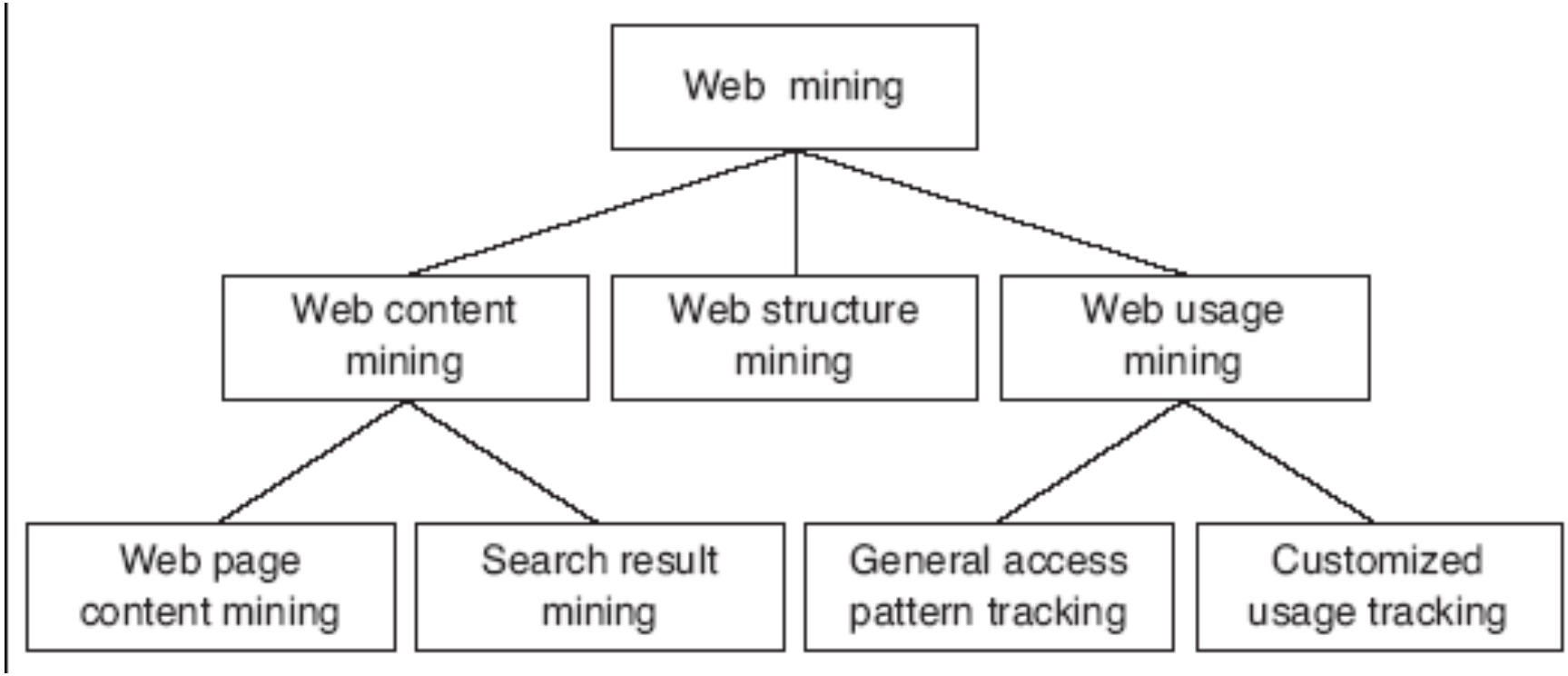
Prof. Lorena Zúñiga S.

Marzo,2020

Web Mining

- El descubrimiento y análisis de información útil e interesante de la web y sobre la web

Tipos de Web Mining



Usos

- Diseñar estrategias de marketing
- Evaluar campañas promocionales
- Enfocar anuncios y cupones electrónicos hacia ciertos grupos de usuarios
- Predecir el comportamiento del usuario
- Presentar información dinámica a los usuarios

Usos

- En comercio electrónico:
  - Mejorar la transformación de un visitante en un cliente
    - Identificar factores que llevan a una compra
  - Identificar anuncios efectivos (ad clicks)
  - *Branding* (Incrementar el reconocimiento y mejorar la imagen de marca)
  - Llevar al cliente a páginas específicas *Target Pages*

Parsing

- *Web scraping*
  1. Inspeccionar contenido en la red y determinar si es valioso para un análisis
  2. Importar las páginas HTML (en R) para extraer la información de interés
- Parsing → comprender la gramática del documento
- Descartar nodos

Expresiones regulares

- Son patrones de texto generalizables
- Permiten buscar y manipular datos textuales
- Ejemplo:

```
> library(stringr)
> strEjemplo<-'1. Una primera oracion. 2. La segunda oracion.'
> str_extract(strEjemplo,'oracion')
[1] "oracion"
> str_extract(strEjemplo,'larga')
[1] NA
```

Expresiones regulares

```
> str_extract_all(strEjemplo,'oracion')
[[1]]
[1] "oracion" "oracion"

> str_extract(strEjemplo,'ORAcion')
[1] NA
> str_extract(strEjemplo,regex('ORAcion',ignore_case = T))
[1] "oracion"
```

- Sin embargo, no es un patrón generalizable

Expresiones regulares

Expresión	Significado
[:digit:]	Dígitos
[:lower:]	Letras en minúscula a..z
[:upper:]	Letras mayúsculas A..Z
[:alpha:]	Caracteres alfabéticos a..z A..Z
[:alnum:]	Caracteres alfanuméricos
[:punct:]	Signos de puntuación
[:blank:]	Espacio en blanco, tabulación
[:space:]	Espacio en blanco, tab, nl

Expresiones regulares

```
> strEjemplo
[1] "1. Una primera oracion. 2. La segunda oracion."
> str_extract_all(strEjemplo,'[:punct:]')
[[1]]
[1] "." "." "." "."

> str_extract_all(strEjemplo,'[:digit:]')
[[1]]
[1] "1" "2"
```

Expresiones regulares

Símbolo	Significado
\w	Caracter alfanumérico
\W	Carácter no alfanumérico
\s	Espacio en blanco [:blank:]
\S	No espacio en blanco
\d	Dígitos
\D	No dígitos
[a..z]	Letras en minúscula
[A..Z]	Letras en mayúscula
[0-9]	Dígitos

Expresiones regulares

```
> str_extract_all(strEjemplo,'\\d')
[[1]]
[1] "1" "2"

> str_extract_all(strEjemplo,'\\D')
[[1]]
 [1] "." " " "U" "n" "a" " " "p" "r" "i" "m" "e" "r" "a" " " "o" "r" "a" "c"
[19] "i" "o" "n" "." " " "." " " "L" "a" " " "s" "e" "g" "u" "n" "d" "a" " "
[37] "o" "r" "a" "c" "i" "o" "n" "."
```

Cuantificadores

Expresión	Significado
?	Opcional, aparece 0 o 1 vez
*	Aparece 0 o más veces
+	1 o más veces
{n}	Exactamente n veces
{n,}	N o más veces
{n,m}	Entre n y m veces

Otros símbolos

Expresión	Significado
.	Cualquier caracter menos \n
\\.	El punto
\\+	Signo de suma
\\$	Signo de dólar
\\[\\] \\(\\)	Paréntesis izquierdo, derecho

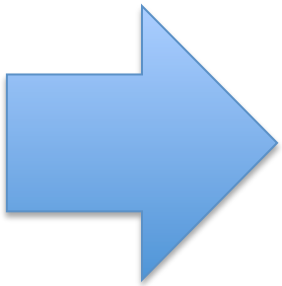
Ejemplos uso expresiones regulares

A partir del siguiente texto:

En total, el país exportó 11494 millones de dólares en el 2018. De esto, más de 1000 millones eran algún tipo de dispositivo médico (agujas, jeringas, catéteres, cánulas o similares), 621 millones de prótesis médicas y 833 millones aparatos médicos y 925 millones piñas frescas.

Ejemplos uso expresiones regulares

Crear la siguiente tabla:



	rubros	montos
1	dispositivo médico	1000
2	prótesis médicas	621
3	aparatos médicos	833
4	piñas frescas	925

En R : Paquete stringr

- string_extract
- string_extract_all
- str_locate
- str_locate_all
- str_split
- str_detect
- str_count
- str_replace
- str_replace_all
- str_remove
- Str_remove_all

Proceso genérico

1. Localizar los datos
2. Obtener los datos
3. Limpiar y/o transformar datos obtenidos
4. Aplicar algún algoritmo o técnica: clustering, PCA, correlación, etc.
5. Visualizar resultados

Ejemplos en R

¿Preguntas?