Minería de Datos para Seguridad

Lectura 2: Recuperación de la información

Aldo Hernández, MEng

IPN SEPI-ESIME Culhuacan

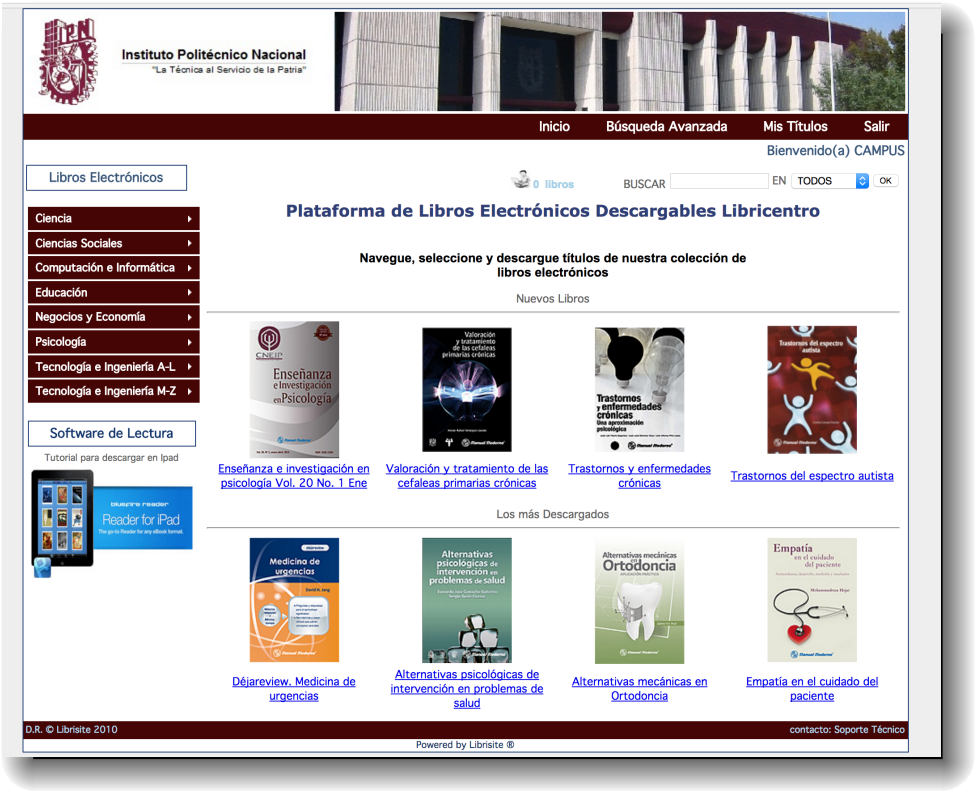
Agosto, 2017

"Todo lo que nos informe acerca de algo útil, que realmente no conozcamos, es una señal en potencia. Si importa y merece una respuesta , su potencial se actualiza."

— Stephen Few, Signal: Understanding What Matters in a World of Noise

¿Cómo recuperamos la información?

- ¿ Acaso vamos a la biblioteca ? ó,
- ¿ Buscamos en su catálogo en linea?



Recuperación y presentación de la información

- Dado un conjunto de documentos, es nuestro objetivo encontrar aquellos k más similares dada una consulta de texto
- Debe de existir una forma de representar dichos documentos
  - Qué sea fácil de visualizar a partir de datos en crudo
  - Que resalten los aspectos más importantes de los documentos y supriman aquellos que no aporten información valiosa

Representación mediante Bolsas de Palabras (1/2)

- La representación por bolsas de palabras es aquella que lista aquella por el número de apariciones dentro del conjunto de documentos

  "El cálculo vectorial o análisis vectorial es un campo de las matemáticas referidas al análisis real multivariable de vectores en 2 o más dimensiones".
  $El = 1; cálculo = 1; vectorial = 2; o = 2; análisis = 2; es = 1; un = 1; campo = 1; de = 1; las = 1; matemáticas = 1; referidas = 1; al = 1; real = 1; multivariable = 1; vectores = 1; en = 1; 2 = 1; más = 1; dimensiones = 1$

Representación mediante Bolsas de Palabras (2/2)

- La palabra dimensión pude tener muchos significados, pero a partir de ella se pueden aprender otros conceptos
  - *Análisis* y *consideración* hacen referencia a dimensión como un punto de vista
  - *Masa*, *longitud* y *tiempo* hacen referencia a dimensión como producto de unidades físicas

Conteo de Palabras (1/3)

- Realizar una lista de todas las palabras presentes en los documentos y una consulta
- Indexar las palabras $w = 1, .., W$ y los documentos $d = 1, .., D$
- Para cada documendo d contar cuantas veces la palabra w aparece y representarla por X_{dw}. El vector $X_d = (X_{d1}, ..., X_{dW})$ indica el conteo de palabras para el $i - ésimo$ documento
- Realizar los mismos pasos para la consulta ; sea Y_w el número de veces que la $i - ésima$ palabra ocurre ; lo cual resulta en el vector $Y = (Y_1, ..., Y_W)$

Conteo de Palabras (2/3)

En el siguiente ejercicio consideramos dos oraciones $D = 2$,catorce palabras $W = 14$ y una consulta Y

- Documentos
  1. X_1 =“Los alumnos aman la minería de datos”
  2. X_2 =“El profesor y los alumnos odian realmente odian la minería de datos”
- Consulta, Y = “odian minería”

Conteo de Palabras (3/3)

En las siguiente figura se muestra el conteo de palabras de los documentos D dado Y

Out[13]:

	El	Los	alumnos	aman	datos	de	la	los	minería	odian	profesor	realmente	y
X1	0	1	1	1	1	1	1	0	1	0	0	0	0
X2	1	0	1	0	1	1	1	1	1	2	1	1	1
Y	0	0	0	0	0	0	0	0	1	1	0	0	0

Conteo de ocurrencias de palabras dados X_1, X_2

En el código anexado en la carpeta **Lectura2Code deben de ir generando sus propios conjuntos de D y vectores de conteo X_d*

Distancias y Medidas de Similitud (1/4)

- Una vez obtenidos X_d , además de Y como vectores, es tiempo de medir la similitud ó de manera equivalente ¿La distancia?
- Algunas medidas de distancia entre vectores X, Y *n-dimensionales* son:
  - La distancia ℓ_2 ó Euclidiana:

$$\|X_d - Y\|_{\ell_2} = \sqrt{\sum_{i=1}^{n}(X_{d_i} - Y_i)^2}$$

  - La distancia ℓ_1 ó Manhattan:

$$\|X_d - Y\|_{\ell_1} = \sum_{i=1}^{n}\left|(X_{d_i} - Y_i)\right|$$

Distancias y Medidas de Similitud (2/4)

Existen tres documentos D que hablan de OSINT

- X_1, Inteligencia humana (espionaje) $\Rightarrow$ `https://es.wikipedia.org/wiki/Inteligencia_humana_(espionaje)`
- X_2, ¿Qué es OSINT? $\Rightarrow$ `http://h4dm.com/que-es-osint/`
- X_3, Open Source Intelligence OSINT $\Rightarrow$ `https://www.intelpage.info/open-source-intelligence-osint.html`

La consulta es $Y =$ " inteligencia OSINT información documentos "

Distancias y Medidas de Similitud (3/4)

El vector de ocurrencias X_d ó *matriz de términos* de cada documento X_1, X_2, X_3 se describe en la siguiente Figura

	100%	85%	Abiertas	Estamos	Fuentes	HUMINT	Human	Inteligencia	La	OSINT	...	sus	también	tienen	tipo	tipos	través	un	una	veces
X1	1	0	1	1	1	0	0	1	0	1	...	1	0	0	0	1	1	0	1	0
X2	0	1	0	0	0	0	0	1	0	0	...	0	1	0	1	0	0	1	0	0
X3	0	0	0	0	0	1	1	2	1	0	...	0	0	1	0	0	0	0	1	1
Y	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0

4 rows × 120 columns

Conteo de ocurrencias de palabras dados X_1, X_2, X_3

Distancias y Medidas de Similitud (4/4)

Hemos encontrado las distancias de Y para X_d

	Distancia Euclidiana
X1	14.525839
X2	13.490738
X3	9.486833
Y	0.000000

Distancia Euclidiana

	Distancia Manhattan
X1	75
X2	62
X3	52
Y	0

Distancia Manhattan

Variación de la Longitud de Documentos y Normalizar (1/3)

- La variación de la longitud calculada en las distancias suele ser un problema
  - Distancias muy grandes y variación en la cantidad de palabras en cada documento D
- Se necesitan normalizar los vectores Y y X_d
  - Hay que tomar las longitudes y estandarizar

Variación de la Longitud de Documentos y Normalizar (2/3)

- Normalizar por longitud del documento, es decir, dividir X_d por la suma de sus componentes

$$X_d \leftarrow X_d / \sum_{w=1}^{W} X_{dw}$$

- Normalizar por longitud de la norma ℓ_2

$$X_d \leftarrow X_d / \|X_d\|_{\ell_2}$$

Variación de la Longitud de Documentos y Normalizar (3/3)

De regreso a la documentos D

	NormaDoc
X1	0.503115
X2	0.523138
X3	0.496288
Y	0.000000

Normalización por longitud del documento

	Distancia Euclidiana/NormaL2
X1	1.288634
X2	1.333731
X3	1.259933
Y	0.000000

Normalización por ℓ_2

Ejercicio

Encontrar información de la MISTI en tres diferentes fuentes de internet

1. guardar el texto citado de cada una en un diferentes archivos *.txt*
2. realizar por lo menos dos consultas Y diferentes, obteniendo la distancia Euclidiana y Manhattan sin normalizar X_d
3. realizar por lo menos dos consultas Y diferentes con X_d normalizado por longitud y por ℓ_2

Próxima

MÁS MEDIDAS DE SIMILITUD :)