

# Tarea 01

Kevin Hernández

**Exercise 1.2.1:** Using Information from Section same 1.2.3, what would be the number of suspected pairs if following changes were made to the data (and all other numbers remained as they were in that section)?

- (a) The number of days of observation was raised to 2000.
- (b) The number of people observed was raised to 2 billion (and there were therefore 200,000 hotels).
- (c) We only reported a pair as suspected if they were at the same hotel at the same time on 3 different days.

## Solución 1.2.1

- a. *Respuesta.*  $((2000)^2)/2 = 2 * 10^6$
- b. *Respuesta.*  $((2 * 10^9)^2)/2 = 2 * 10^{18}$
- c. *Respuesta.* El número de dos personas que se conozca en 3 diferentes días es:  $13 * 10^9$

**Exercise 1.3.2:** Suppose there is a repository of ten million documents, and word  $w$  appears in 320 of them. In a particular document  $d$ , the maximum number of occurrences of a word is 15. Approximately what is the TF.IDF score for  $w$  if that word appears (a) once (b) five times?

## Solución 1.3.2

- a. once:
  - $N = 10^7$  documents.
  - $n_i$  = appears in 320 of them.
  - $IDF_w = \log_2(N/n_i) = \log_2(10^7/320)$
  - appears of word = 1
  - maximum number of occurrences = 15
  - $TF_{wk} = 1/15$ , Suppose that in document  $k$ , word  $w$  appears once, while the maximum number of occurrences of a word is 15.
  - *Respuesta.*  $TF.IDF = TF_{wk} \times IDF_w = 1/15 \times \log_2(10^7/320) = \mathbf{0.9954379046216117}$
- b. five times:

- appears of word = 5
- $TF_{wk} = 5/15$
- *Respuesta.*  $TF.IDF = TF_{wk} \times IDF_w = 5/15 \times \log_2(10^7/320) =$   
**4.9771895231080587**

**Exercise 1.3.4:** In terms of e, give approximations to:

- (a)  $(1.01)^{500}$
- (b)  $(1.05)^{1000}$
- (c)  $(0.9)^{40}$

**Solución 1.3.4**

- a.  $(1.01)^{500} = (1 + 0.01)^{500} \approx e^{ab} \approx e^{0.01 \cdot 500} \approx$  **148.4131591025766034**
- b.  $(1.05)^{1000} = (1 + 0.05)^{1000} \approx e^{ab} \approx e^{0.05 \cdot 1000} \approx$  **5184705528587072464100**
- c.  $(0.9)^{40} = (1 - 0.1)^{40} \approx e^{-ab} \approx e^{-0.1 \cdot 40} \approx$  **0.0183156388887342**

**Exercise 3.3.2:** Using the data from Fig. 3.4, add to the signatures of the columns the values of the following hash functions:

- (a)  $h3(x) = 2x + 4 \bmod 5$ .
- (b)  $h4(x) = 3x - 1 \bmod 5$ .

**Solución 3.3.2**

Row	$S_1$	$S_2$	$S_3$	$S_4$	$x + 1 \bmod 5$	$3x + 1 \bmod 5$	$2x + 4 \bmod 5$	$3x - 1 \bmod 5$
0	1	0	0	1	1	1	4	-1
1	0	0	1	0	2	4	1	2
2	0	1	0	1	3	2	3	0
3	1	0	1	1	4	0	0	3
4	0	0	1	0	0	3	2	1