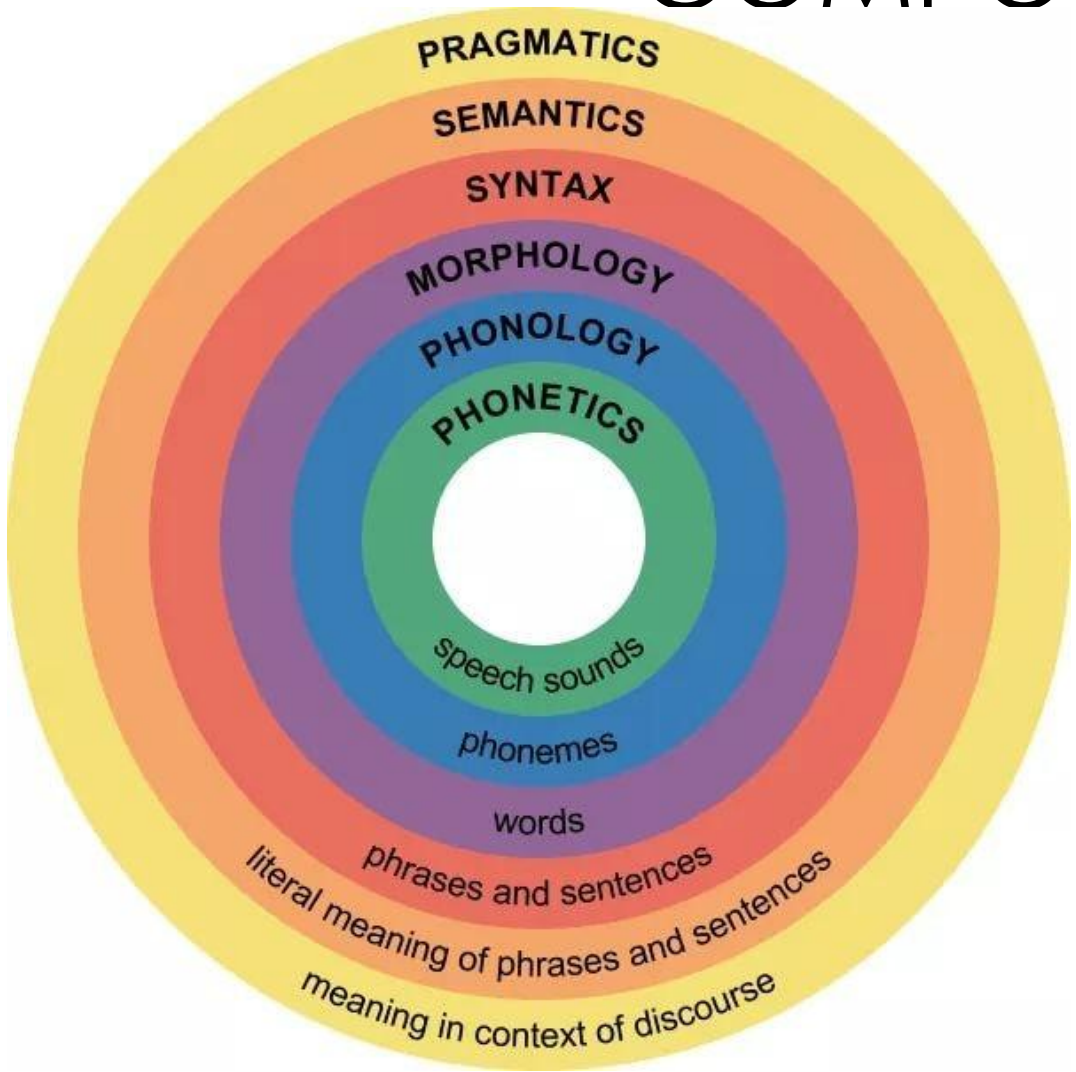
# INTRODUCCIÓN AL PROCESAMIENTO DE LENGUAJE NATURAL

Dra. Cecilia Reyes Peña

# LENGUAJE

- El **lenguaje** es un sistema complejo integrado por **componentes** los cuales se pueden agrupar en gramaticales (sintaxis, morfología y fonología), de contenido (semántico) y de uso (pragmático).

# COMPONENTES DEL LENGUAJE



# FONOLOGÍA

- El componente fonológico se ocupa del **aspecto sonoro** del lenguaje, tanto las reglas de su estructura, como la secuencia de sonidos.

- La partícula más sencilla que estudia la fonología es el **fonema**, que puede ser un solo sonido.

# MORFOLOGÍA

- Estudia los morfemas y sus combinaciones para formar palabras a partir de un **lexema** (raíz).
- **Morfema**: unidad más pequeña de la lengua que tiene significado léxico o gramatical y no puede dividirse en unidades significativas menores.

- Ejemplo:
- Camin□**lexema**
- Caminante□**morfema**
- Caminar□**morfema**

# LEMATIZACIÓN

- Es el proceso de identificación de los lexemas a partir de morfemas.

Escribir
Escritura
Escribiste
Escribamos
Escrituración

| Lematización: | Stemming: |
| --- | --- |
| Escribir o Escrito | Escri |

Lexema Escribir->Morfemas Escribiste, Escritura, Escribí, Escribiré, Escribamos

# SINTÁCTICA

- **Léxico:** conjunto de palabras que constituyen un lenguaje.
- **Vocabulario**: conjunto de palabras que pertenecen a un lenguaje.
- La **sintaxis** se define como el conjunto de **reglas gramaticales** para formar frases de un lenguaje, determinando el orden correcto de las palabras.
- Busca dar coherencia a las expresiones.

# TOKENIZACIÓN

- La tokenización es el proceso de **delimitar** secciones de una cadena de caracteres de entrada.

- Ejemplo: El perro, el gato y el ratón no son amigos.

- Tokens: [‘El’, ’perro’, ‘,’,‘el’, ‘gato’, ‘y’, ‘el’, ‘ratón’, ‘no’, ‘son’, ‘amigos’,’.’]

# N-GRAMAS

- Es el conjunto de n-tokens consecutivos, independientemente de su componente semántico.

- Ejemplo: El perro, el gato y el ratón no son amigos.

Unigramas: El perro , el gato y el ratón no son amigos .

Bigramas: (El,perro),(perro,,),(,,el),(el,gato),(gato,y)……

Trigramas: (El,perro,,),(perro,,,el),(,,el,gato),(el,gato,y)…..

# ETIQUETADO MORFOSINTÁCTICO

- Asignación de etiquetas de acuerdo a las características de cada token según el discurso ***Part-Of-Speech*** (POS-Tags).

- Cada modelo utiliza sus etiquetas propias.

- Spacy utiliza las etiquetas de **Dependencias Universales** (UD), el cual es un framework para la anotación coherente de la gramática (partes del discurso, características morfológicas y dependencias sintácticas) en diferentes idiomas humanos.

# ETIQUETADO MORFOSINTACTICO

| Etiqueta | Descripción | Ejemplo |
|---|---|---|
| ADJ | adjective | *big, old, green, incomprehensible, first* |
| ADP | adposition | *in, to, during* |
| ADV | adverb | *very, tomorrow, down, where, there* |
| AUX | auxiliary | *is, has (done), will (do), should (do)* |
| CONJ | conjunction | *and, or, but* |
| CCONJ | coordinating conjunction | *and, or, but* |
| DET | determiner | *a, an, the* |
| INTJ | interjection | *psst, ouch, bravo, hello* |
| NOUN | noun | *girl, cat, tree, air, beauty* |

# ETIQUETADO MORFOSINTACTICO

| Etiqueta | Descripción | Ejemplo |
|---|---|---|
| NUM | numeral | *1, 2017, one, seventy-seven, IV, MMXIV* |
| PART | particle | *'s, not,* |
| PRON | pronoun | *I, you, he, she, myself, themselves, somebody* |
| PROPN | proper noun | *Mary, John, London, NATO, HBO* |
| PUNCT | punctuation | *., (, ), ?* |
| SCONJ | subordinating conjunction | *if, while, that* |
| SYM | symbol | *$, %, §, ©, +, −, ×, ÷, =, :), * |
| VERB | verb | *run, runs, running, eat, ate, eating* |
| X | other | *sfpksdpsxmsa* |
| SPACE | space | |

# SEMÁNTICA

- Estudia el **significado** de las palabras, frases u oraciones.
- Su análisis debe ser restringido de acuerdo al **dominio**.
- Utiliza un modelo de lenguaje

# STOP WORDS

- Son las palabras más utilizadas en el lenguaje que carecen de significado por si solas como artículos, pronombres, preposiciones.

# ENTIDADES NOMBRADAS

- Son objetos pertenecientes al mundo real, el cual es identificado por medio de etiquetas como: locación, persona, organización, evento, entre otros.

- https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/named-entity-recognition/concepts/named-entity-categories

# PRAGMÁTICA

- La pragmática toma en consideración los **factores extralingüísticos** que condicionan el uso del lenguaje, esto es, todos aquellos factores a los que no se hace referencia en un estudio puramente formal.

- Son las estrategias para usar el lenguaje apropiadamente en varios **contextos**.

- Significado intencional

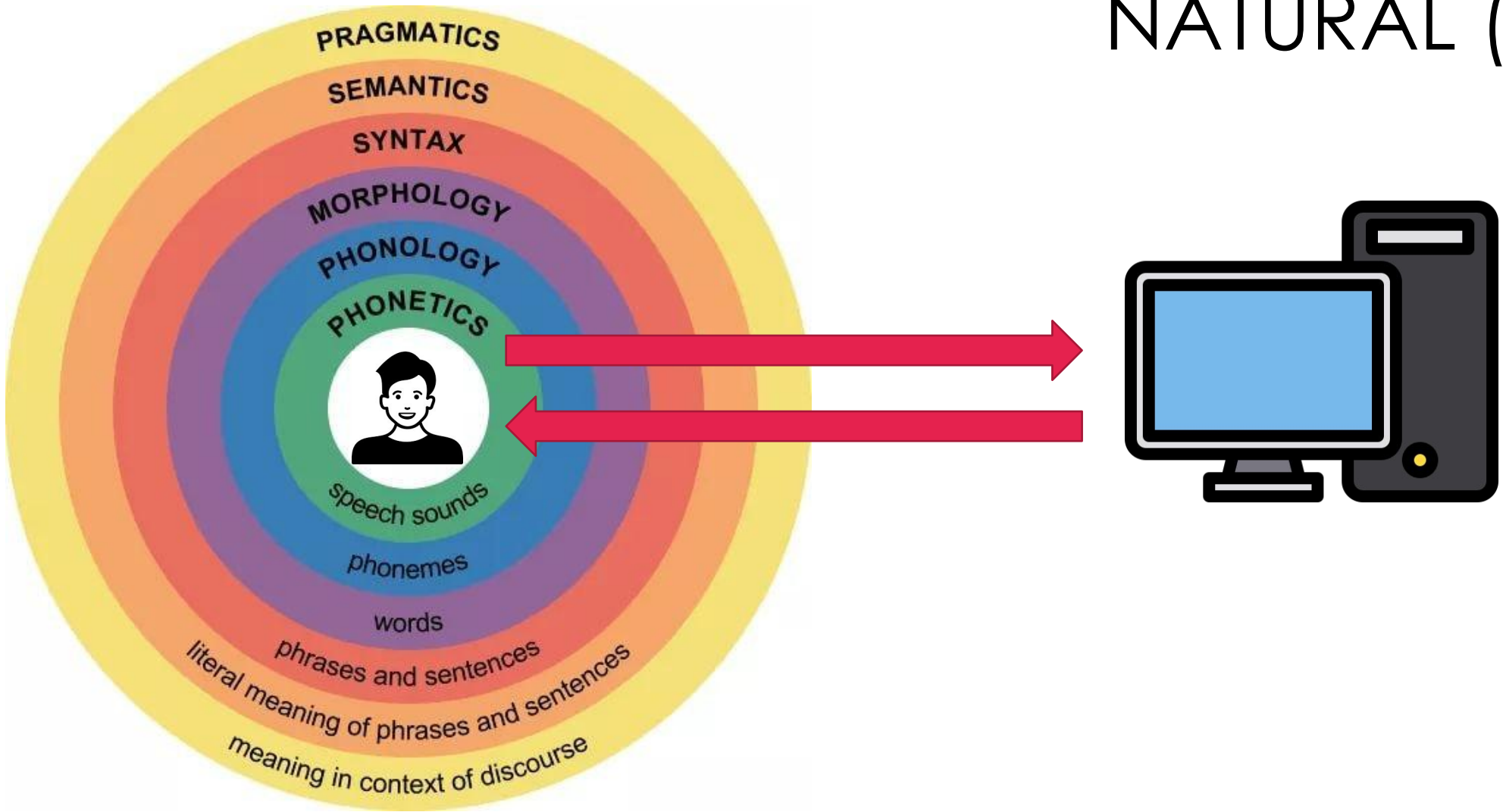
- Va más allá de la frase como la referencia a los pronombres

# LENGUAJE NATURAL VS LENGUAJE ARTIFICIAL

- Lenguaje natural: utilizado para la comunicación entre humanos
- Lenguaje artificial: lenguajes formales, notaciones matemáticas, lenguajes de programación, entre otros. Estos lenguajes pueden ser entendidos por computadoras.

- Elementos considerados dentro del lenguaje natural:
  - Texto
  - Señales
  - Imagen**

# PROCESAMIENTO DEL LENGUAJE NATURAL (PLN)



# TAREAS DEL PLN

- Síntesis del discurso (text-to-speech)
- Reconocimiento del habla (speech recognition)
- Síntesis de voz (speech-to-text)
- Traducción automática (Machine traslation)
- Respuesta a preguntas (Question answering)
- Recuperación de la información (Information retrieval)
- Extracción de la información (Information Extraction)
- Análisis de sentimientos (sentiment analysis)
- Generación automática de resúmenes (Summarization)
- Identificación de entidades nombradas (Named-Entity Recognition)
- Etiquetado sintáctico (Parts-of-Speech tagging (POS))

# PROCESAMIENTO DE VOZ

- Reconocimiento de voz, que trata el análisis del contenido lingüístico de una señal de voz.
- Reconocimiento de locutores, que tiene como objetivo identificar al hablante.
- Mejora de la señal de voz, por ejemplo reducción de ruido.
- Codificación de voz para compresión de datos y transmisión de la voz.
- Análisis de voz con propósitos médicos, para el análisis de disfunciones vocales.
- Síntesis de voz: la síntesis artificial del habla, lo que habitualmente significa habla generada por computador.

# PROCESAMIENTO DE IMAGEN

- Compresión de imágenes
- Detección de objetos
- Mejora de imágenes
- Segmentación
- Restauración
- Reconstrucción

# PROCESAMIENTO DE TEXTO

# RECUPERACIÓN DE LA INFORMACIÓN

- La recuperación de información es el conjunto de actividades orientadas a facilitar la localización de determinados datos u objetos, y las interrelaciones que estos tienen a su vez con otros.

# BÚSQUEDA DE DOCUMENTOS POR PALABRA

- Se tiene un conjunto de *n* documentos. Se quiere recuperar los que contengan la palabra *w*.

| Doc | Palabra |
|---|---|
| 1 | Pido perdón a los niños por haber dedicado este libro a una persona mayor. Tengo |
| 2 | una seña excusa: esta persona mayor es el mejor amigo que tengo en el mundo. |
| 3 | Pero tengo otra excusa: esta persona mayor es capaz de comprenderlo todo, |
| 4 | incluso los libros para niños. Tengo una tercera excusa todavía: esta persona |
| … | … |
| n | mayor vive en Francia, donde pasa hambre y frío. Tiene, por consiguiente, una |

# SOLUCIÓN

| word | No. De documento |
|---|---|
| carro | 1,2,4,5,…. |
| casa | 2,4,6,…. |
| | |
| | |
| | |
| | |
| | |
| wn | |

# SOLUCIÓN

- Carro y casa

| | Lista de documentos | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| carro | 1 | 2 | 4 | 6 | 10 | 12 | 15 | 16 | … |
| | | | ↑ | | | | | | |
| | | | | | | | | | |
| casa | 1 | 3 | 7 | 8 | 9 | 10 | 16 | 17 | … |
| | | | ↑ | | | | | | |
| | | | | | | | | | |
| salida | 1 | | | | | | | | |

# SOLUCIÓN

- Carro o casa

| | Lista de documentos | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| carro | 1 | 2 | 4 | 6 | 10 | 12 | 15 | 16 | … |
| | | | | | | | | | |
| | | | | | | | | | |
| casa | 1 | 3 | 7 | 8 | 9 | 10 | 16 | 17 | … |
| | | | | | | | | | |
| | | | | | | | | | |
| salida | 1 | 2 | 3 | 4 | 6 | 7 | … | | |

# TAREA

- Negación de existencia de palabras
- (Perro y gato) not raton
- Not carro y (casa or coche)

# SIMILITUD LÉXICA

# SIMILITUD LÉXICA

- La similitud léxica es el grado de semejanza de dos o más palabras entre si mismas.
- La similitud léxica puede utilizarse para comparar lenguajes.
- En PLN se puede utilizar para la detección de errores de escritura.

# ONE-HOT ENCODING

- Representación vectorial con 0's y 1's, donde el 1 indica la existencia de una palabra en una oración y 0 la ausencia de esta.
- T1: Me gustó el carro rojo de María Luisa
- T2: Me encantó el carro rojo de Luisa
- T3: Me gusto el carro de María Luisa y de Luisa

| | Me | gustó | encantó | el | carro | rojo | de | María | Luisa | y |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| T3 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |

# BOLSA DE PALABRAS

- Representación vectorial donde un entero indica la frecuencia de una palabra en una oración y 0 la ausencia de esta.
- T1: Me gustó el carro rojo de María Luisa
- T2: Me encantó el carro rojo de Luisa
- T3: Me gusto el carro de María Luisa y de Luisa

| | Me | gustó | encantó | el | carro | rojo | de | María | Luisa | y |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| T3 | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 2 | 1 |

Coeficiente de .

| | Me | gustó | encantó | el | carro | rojo | de | María | Luisa | y |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| T3 | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 2 | 1 |

$$sim_J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Jac(t1,t2)= 6/9

## Distancia de Jaccard

$$J_\delta(A,B) = 1 - sim_J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

# COEFICIENTE DE DICE

- Es una variación del coeficiente de Jaccard

$$sim_D(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Dice(t1,t2)=(2*6)/(8+7)
Dice(t2,t3)=(2*6)/(8+7)
Dice(t1,t3)=(2*6)/(10+8)

| | Me | gustó | encantó | el | carro | rojo | de | María | Luisa | y |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| T3 | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 2 | 1 |

# DISTANCIA LEVENSHTEIN

- Es el número mínimo de operaciones que se necesitan para transformar una cadena en otra.
- Las operaciones son:
  - Inserción de un carácter
  - Eliminación de un carácter
  - Sustitución de un carácter
  - Trasposición de dos caracteres

P1: Casa
P2: Cazador

D=1+3=4

P3: Traza
P4: Transporte

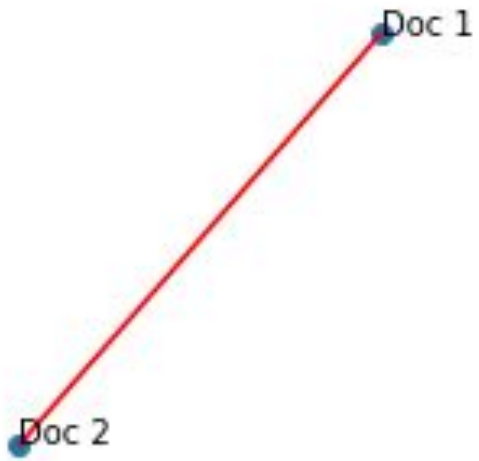
D(p3,p4): 1+1+5=7

# CLIDIANA

- La

| | Me | gustó | encantó | el | carro | rojo | de | María | Luisa | y |
|---|---|---|---|---|---|---|---|---|---|---|
| T1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| T2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| T3 | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 2 | 1 |

$$\text{Euclidiana}(t1,t2)=\sqrt{\sum_{0}^{n-1}(t1_n - t2_n)}$$



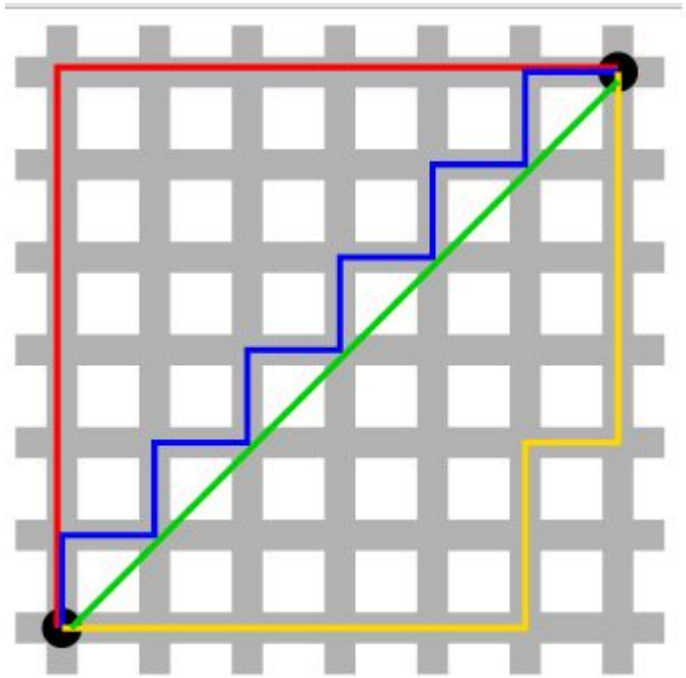

Euc(t1,t2)=raíz(0+-1+1+0+0+0+0+-1+0)^2
raíz(1)

Euc(t1,t3)=raíz(0+-1+1+0+0+-1+1+0+1+1)^2
Raiz(4)

# DISTANCIA MANHATTAN

- Es la suma de las diferencias de sus componentes.

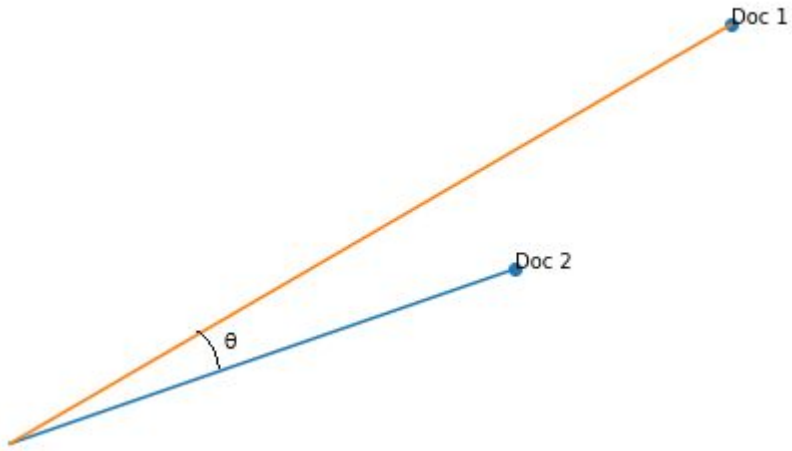
$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$



# SIMILITUD POR COSENO

- Es una medida de la similitud existente en un **rango [-1,1]** entre dos vectores en un espacio que posee un producto interior con el que se evalúa el valor del **coseno del ángulo** comprendido entre ellos.
- Esta función trigonométrica proporciona un valor igual a **1** si el ángulo comprendido es **cero**, es decir si ambos vectores apuntan a un mismo lugar.
- Cualquier ángulo existente entre los vectores, el coseno arrojaría un valor inferior a uno.
- Si los vectores fuesen **ortogonales** (90 grados) el coseno se **anularía**, y si apuntasen en **sentido contrario** su valor sería **-1**.

$$\text{cosine}(\mathbf{v},\mathbf{w}) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

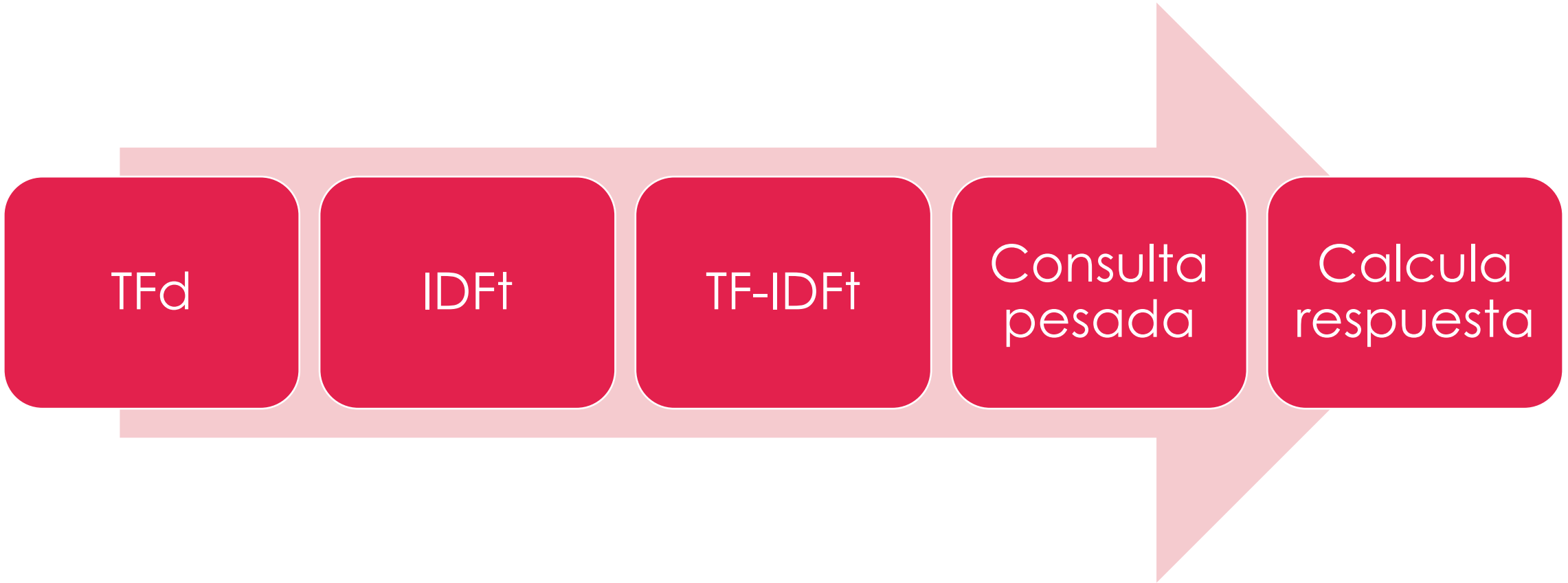


# RANKING DE PALABRAS

# RANKING

- El *Ranking* se puede definir como un proceso que permite asignar un valor de **relevancia** a cada término dentro un documento perteneciente a un corpus, con la finalidad de que puedan ser ordenados para satisfacer una tarea de pregunta-respuesta.
- **Ranking simple** consiste en aplicar algoritmos y crear listas de relevancia basándose en las características propias de los documentos
- **Ranking de agregación** retoma otras listas de Ranking de los documentos ya establecidas para formar una nueva.

# RANKING SIMPLE



Palabra [doc1, doc2, ……]
[#doc1, #doc2,….]

Tamano de la coleccion

# RANKING SIMPLE

- Este algoritmo de Ranking simple trabaja con valores asociados a:
- La **frecuencia del término** en el documento ($TF_{t,d}$, Term Frequency), que el número de veces que aparece el término en un documento.
- El **total de documentos** de la colección (N)
- El **número de documentos** de la colección en el que aparece un término ($DF_t$ Document Frequency)

$$IDF_t = \log_{10} \frac{N}{DF_t}$$

- Donde el $IDF_t$ **frecuencia inversa del término en el documento**

Palabra [doc1, doc2, ……]
[#doc1, #doc2,….]

Tamano de la coleccion

# RANKING

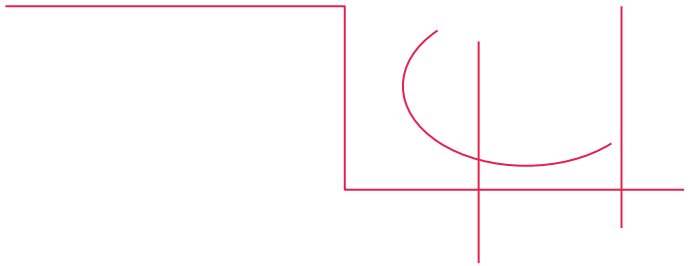
- Posteriormente, se multiplica la frecuencia del término en el documento ($TF_{t,d}$) por el IDFt, para obtener el puntaje de los términos que será utilizado para realizar las consultas.

$$TF - IDF_{t,a} = TF_{t,a} \times IDF_t$$

# RANKING SIMPLE

- Un peso alto en **TF-IDF** se alcanza con una **alta frecuencia** de la palabra en el **documento** dado y una **baja frecuencia** de la palabra en la **colección completa de documentos**.
- Como el cociente dentro de la función logaritmo del idf es siempre mayor o igual que 1, el valor del idf (y del tf-idf) es mayor o igual que 0.

- Cuando un término aparece en muchos documentos, el cociente dentro del logaritmo se acerca a 1, ofreciendo un valor de idf y de tf-idf cercano a 0



# RANKING SIMPLE

- Para la consulta, lo primero es **asignar un peso** a cada uno de los términos de la **consulta** ($w_{t,q}$) dentro de un vector, para lo cual se obtiene el log10 de la frecuencia del término dentro de la consulta ($TF_{t,q}$) y se multiplica por el $IDF_t$ que ya se tiene almacenado.

$$w_{t,q} = \log_{10}(1 + TF_{t,q} \times IDF_t)$$

[casa, perro, gato, casa]

# RANKING SIMPLE

- Para la recuperación, se hace el producto punto entre el vector de pesos en la consulta con el vector compuesto de TF-IDFt,d de cada término para obtener el puntaje final que representa la relevancia de los documentos respecto a la consulta. Los K vectores de mayor puntaje son los que encabezan el Ranking.

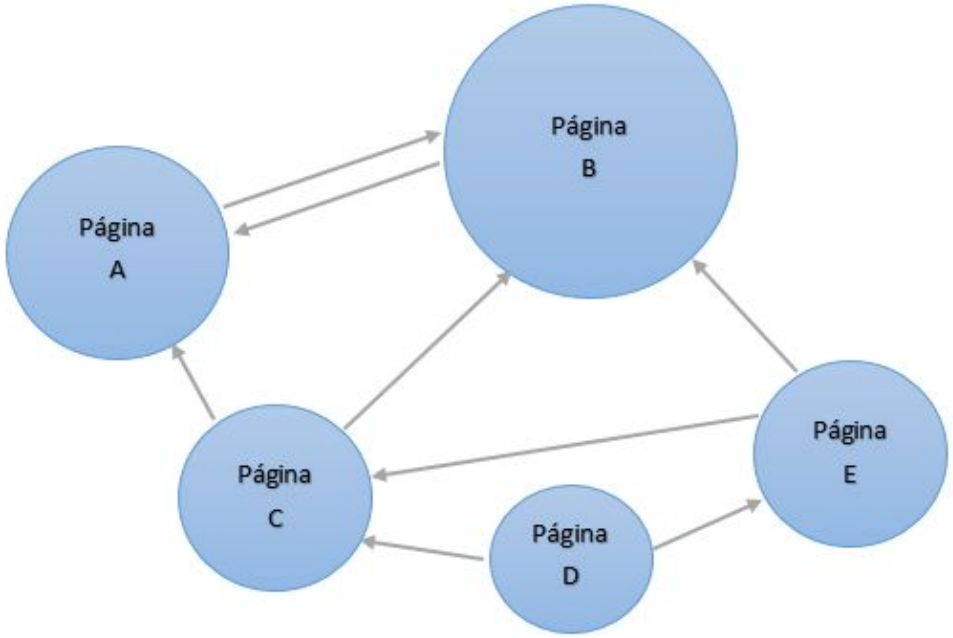
$$Ranking = \overrightarrow{w_{t,q}} \cdot \overrightarrow{TF-IDF_{t,d}}$$

# PAGERANK

# PAGERANK

- Es un método desarrollado Google para asignar **relevancia** a una página web.

- Maneja a la web como un **grafo** donde cada página web es un **nodo** que tiene aristas entrantes y salientes las cuales representan a los **hipervínculos** que apuntan a la página desde otros sitios.



# PAGERANK

- Analiza los **enlaces** hacia una página, al igual que la página que los emite.
- Los enlaces emitidos por las páginas consideradas **importantes**, es decir con un PageRank elevado, valen más, y ayudan a hacer a otras páginas importantes.
- Si no hay enlaces a una página web, no hay apoyo a esa página específica.
- El PageRank de la barra de Google va de 0 a 10, donde:
  - **10** es el máximo PageRank posible (pocos sitios)
  - **1** es la calificación mínima que recibe un sitio normal
  - **0** significa que el sitio ha sido penalizado o aún no ha recibido una calificación de PageRank.

# PAGERANK

- El cálculo es de forma recursiva, donde:
- ***PR(A)*: *PageRank* de la página *A***
- ***d***: factor de amortiguamiento
- ***PR(i)*: *PageRank* de cada una de las páginas *i* que enlazan a *A***
- ***C(i)*:** es el número total de enlaces salientes de la página ***i***.

- 
- $PR(A) = (1 - d) + d \sum_{i=1}^{n} \frac{PR(i)}{C(i)}$

# PAGERANK

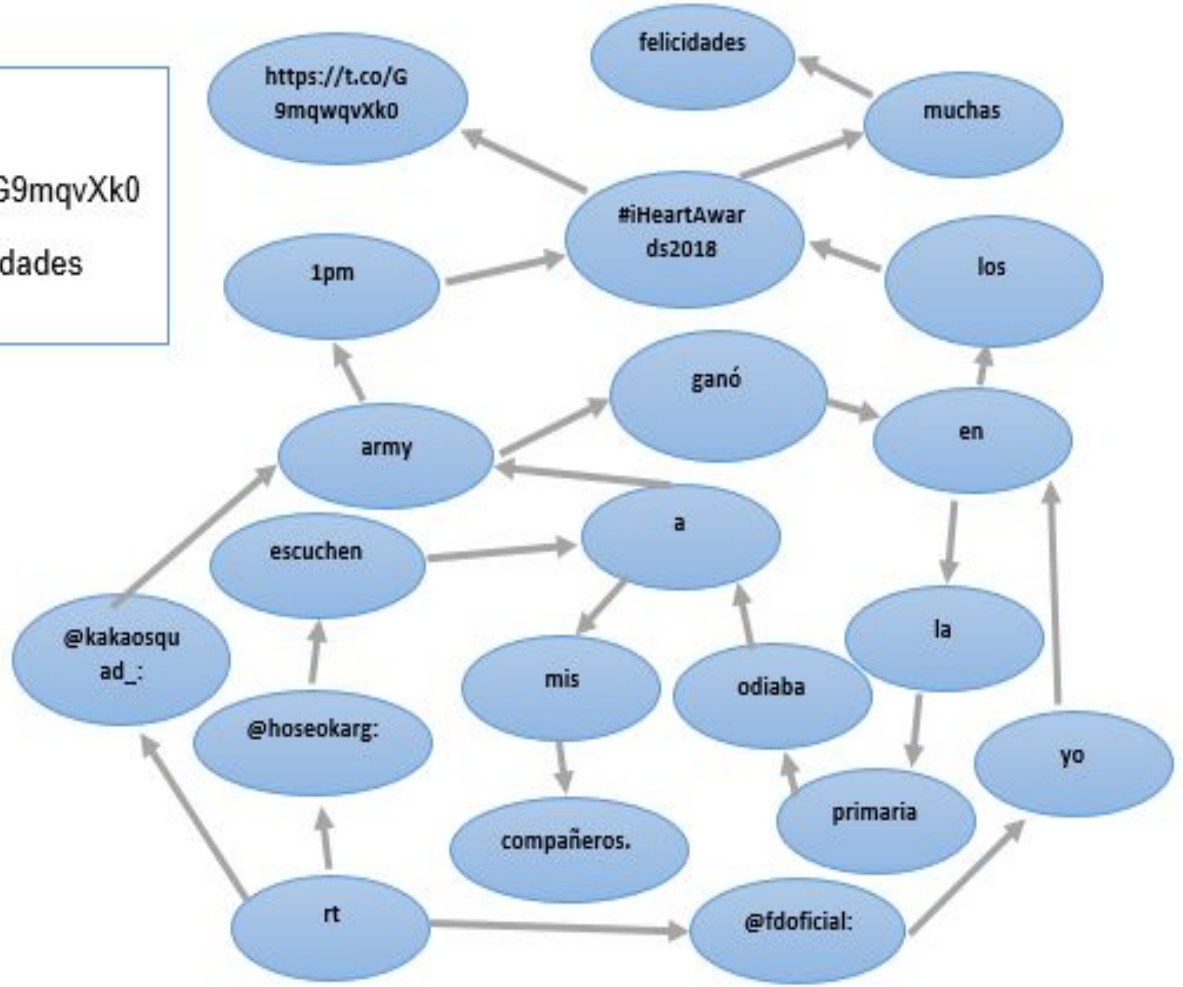
- Los enlaces de una página a sí misma se ignoran.
- Los enlaces salientes múltiples de una página a otra página se tratan como un solo enlace.
- PageRank se inicializa con el mismo valor para todas las páginas (1/cantidad total de paginas).
- PageRank sostiene que un navegante que hace clic aleatoriamente en los enlaces eventualmente dejará de hacerlo. La probabilidad, en cualquier paso, de que la persona continúe es un factor de amortiguamiento d (alrededor de 0.85).

# DETECCIÓN DE TRENDING TOPICS

RT @FDoficial: Yo en la primaria odiaba a mis compañeros.

RT @HoseokArg: ESCUCHEN A ARMY 1PM #iHeartAwards2018 https://t.co/G9mqvXk0

RT @kakaosquad_: ARMY ganó en los #iHeartAwards2018 Muchas Felicidades



# CLASIFICACIÓN ESTADÍSTICA

# ADIVINA LA PALABRA SIGUIENTE

# ADIVINA LA PALABRA SIGUIENTE

- De

- De la

- De la sierra

- De la sierra, morena

- De la sierra, morena cielito

# ADIVINASTE!!!

- De la sierra, morena cielito lindo vienen bajando. Un par de ojitos negros cielito lindo de contrabando.

# PROBABILIDAD DE OCURRENCIA

- La probabilidad de que una palabra aparezca después de otra de forma independiente es:

$$S = P(w_1) \times P(w_2) \times \ldots \times P(w_n)$$

- Esto solo es posible considerando unigramas de un texto y cada uno tiene una probabilidad de 1 entre la longitud del vocabulario de un lenguaje.

# PROBABILIDAD

- Sin embargo, si se tiene un indicio de la palabra anterior la probabilidad está condicionada, ya que dependería de la probabilidad de la ocurrencia de la palabra anterior.

$$S = P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_n|w_{n-1})$$

- Esto es gracias a los N-gramas. Cabe resaltar que cada grado de N-gramas contiene más información que el grado N-1.
- Propiedad de Markov indica que la probabilida

# REGLA DE LA CADENA

- Es una generalización de las palabras ocurridas anteriormente

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1..w_2)...P(w_n|w_1...w_{n-1})$$

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_1^{k-1})$$

# TEOREMA DE BAYES

- Teorema de Bayes (Thomas Bayes, 1763), el cual específica que:
- Sea A={A1, A2,…, An} un conjunto de sucesos mutuamente excluyentes y exhaustivos, donde la probabilidad de cada uno de ellos es distinta de cero.
- Sea B un suceso cualquiera del que se conocen las probabilidades condicionales P(B | Ai) y Ai un suceso perteneciente a A.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

# TEOREMA DE BAYES

- Donde P(Ai) es la probabilidad a priori del suceso Ai,
- P(B | Ai) es la probabilidad verosímil del suceso B supuesto que hubiera ocurrido en el suceso Ai y P(Ai | B) es la probabilidad a posteriori del suceso Ai.

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)}$$

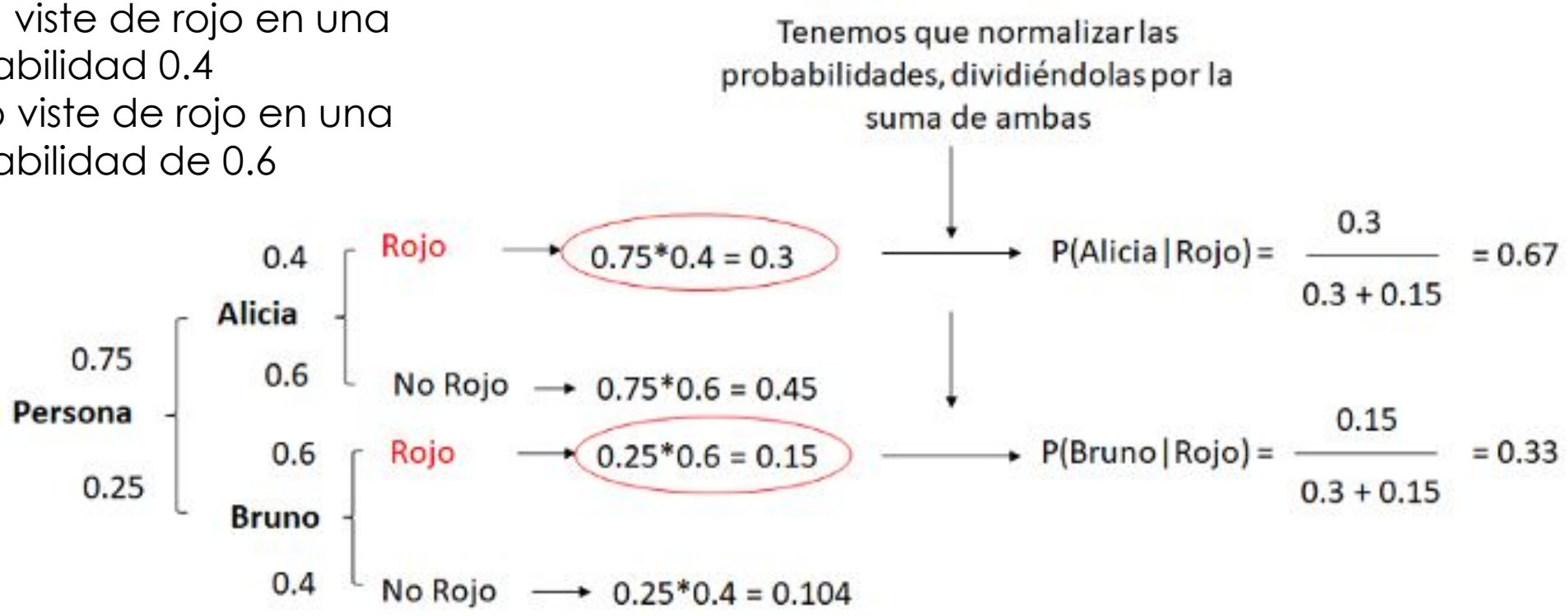
# NAÏVE BAYES

- Está basado en la simplificación del Teorema de Bayes y asume que cada una de las características de una clase, contribuyen de forma independiente a la probabilidad de que un objeto pertenezca a dicha clase.
- Su entorno de aprendizaje es supervisado y no necesita una gran cantidad de elementos para el entrenamiento.

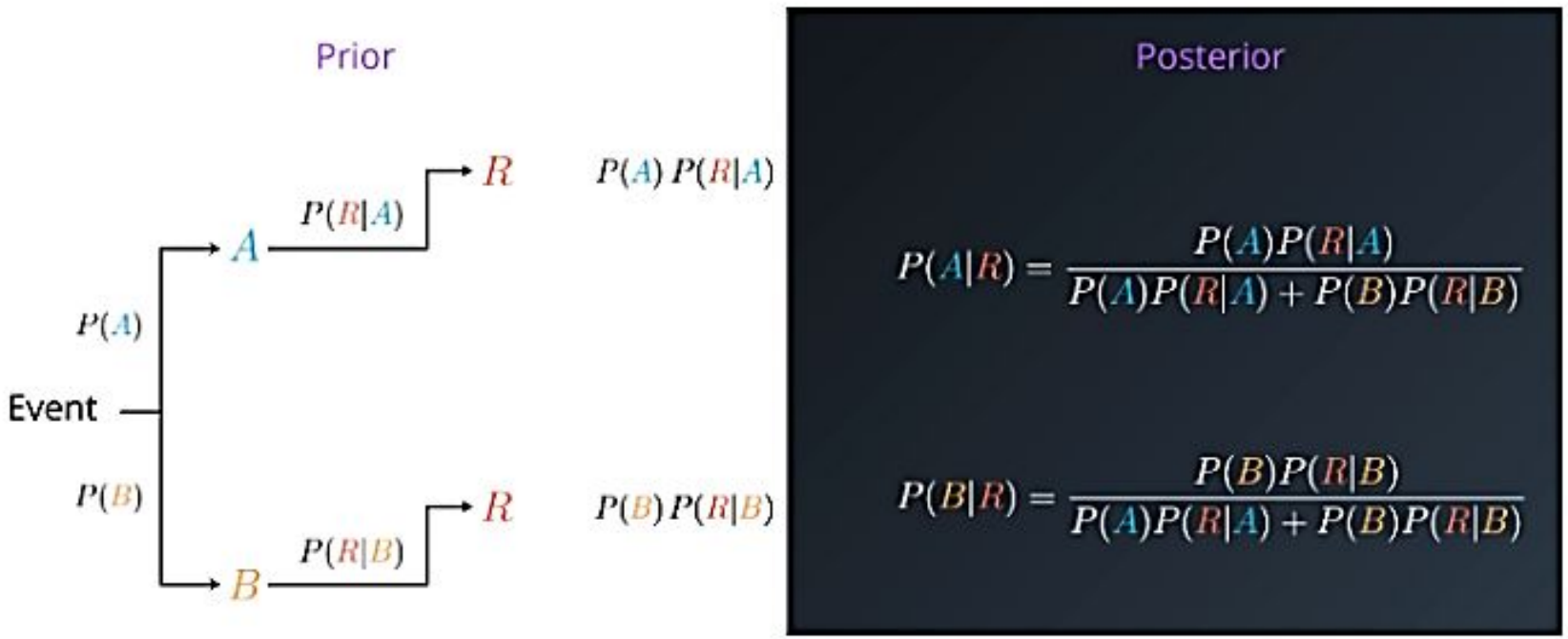
# NAÏVE BAYES

En oficina se trabajan 4 días a la semana
En esa oficina Alicia asiste 3 días a la semana y Bruno asiste solo 1
Alicia viste de rojo en una probabilidad 0.4
Bruno viste de rojo en una probabilidad de 0.6

Problema: Estando en la oficina alguien paso vistiendo de rojo y no se sabe quien es



# NAÏVE BAYES


Prior
Event
$P(A)$
$P(B)$
$A$
$B$
$P(R|A)$
$P(R|B)$
$R$
$R$
$P(A)\,P(R|A)$
$P(B)\,P(R|B)$
Posterior
$P(A|R) = \frac{P(A)P(R|A)}{P(A)P(R|A) + P(B)P(R|B)}$
$P(B|R) = \frac{P(B)P(R|B)}{P(A)P(R|A) + P(B)P(R|B)}$

# DETECCIÓN DE SPAM

- El teorema de Bayes puede utilizarse para detectar si un correo es o no spam, a partir de la probabilidad de las palabras en correos anteriores.

$$\Pr(S|W) = \frac{\Pr(W|S) \cdot \Pr(S)}{\Pr(W|S) \cdot \Pr(S) + \Pr(W|H) \cdot \Pr(H)}$$

dónde:

- $\Pr(S|W)$ es la probabilidad de que un mensaje es un correo no deseado, a sabiendas de que la palabra "réplica" está en él;
- $\Pr(S)$ es la probabilidad general de que cualquier mensaje dado es spam;
- $\Pr(W|S)$ es la probabilidad de que la palabra "réplica" aparece en los mensajes de spam;
- $\Pr(H)$ es la probabilidad general de que cualquier mensaje dado no es spam (es "legítimo");
- $\Pr(W|H)$ es la probabilidad de que la palabra "réplica" aparece en los mensajes legítimos.

$$\Pr(S) = 0.8; \Pr(H) = 0.2$$

| Bandeja entrada mensajes reales | Spam |
|---|---|
| W1, w2, wn, ….. | W2, w3, ….. |
| Oficio, documento, pago,…. | Pago, vacaciones, préstamo, ….. |
| P( r ) | P (s) |
| | |
| | |
| | |
| | |
| Mensaje nuevo: Necesitamos su pago | |

| Asistencia 10% | Tareas simples 10% | 80% | | | | |
|---|---|---|---|---|---|---|
| | La instalación de herramientas | Programa de características | | | | |
| | Adquisión de los libros | Programa recuperación | | | | |
| | | Programa tf- Idf | | | | |
| | | Naive Bayes | | | | |
| | | Examen | | | | |
| | | | | | | |