Análisis de palabras y extracción de información

Arturo Curiel

me@arturocuriel.com

3 de septiembre de 2019

Contenidos

Expresiones regulares

- La mayoría de los algoritmos de tokenización usan expresiones regulares.
  - ⇒ Normalmente para lenguajes separados por espacio en blanco.
- Se pueden entender como un lenguaje de especificación de lenguajes regulares.
  - ⇒ Permiten describir todas las cadenas de un lenguaje regular.

Formalidad (yay!)

Lenguaje

Un **lenguaje** A es un conjunto de cadenas.

- Una cadena es una secuencia de símbolos sobre un alfabeto Σ.
  - $\Rightarrow$ *e.g.* si $\Sigma = \{0, 1\}$ entonces 0001, 11000 y 101010 son cadenas sobre Σ.
- *e.g.* $A = \{00, 01, 10, 11\}$ es el lenguaje de todos los números binarios representables con dos bits.

Formalidad (yay!)

Lenguaje Regular

Un **lenguaje regular** R es un lenguaje reconocido por un autómata finito.

Formalidad (yay!)

Autómata finito

Un **autómata finito** is una 5-tupla $(Q, \Sigma, \delta, q_0, F)$ donde,

1. Q es un conjunto finito de **estados**,
2. Σ en un conjunto finito de símbolos llamado **alfabeto**,
3. $\delta : Q \times \Sigma \to Q$ es una **función de transición**,
4. $q_0 \in Q$ es un estado inicial, y,
5. $F \subseteq Q$ es un **conjunto de estados de aceptación** (finales).

Informalidad :-(

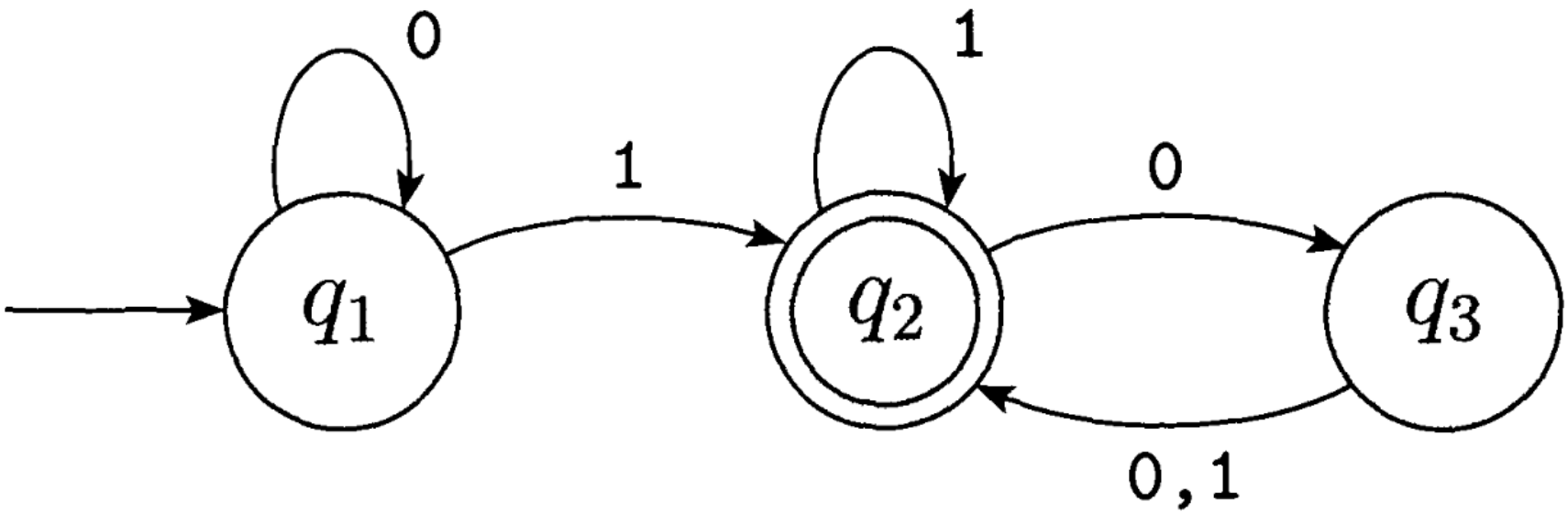

Figura: Diagrama de un autómata finito M_1

Formalidad (yay!)

$M_1 = (Q, \Sigma, \delta, q_1, F)$ tal que

- $Q = \{q_1, q_2, q_3\}$,
- $\Sigma = \{0, 1\}$,
- δ se puede describir como:

	0	1
q_1	q_1	q_2
q_2	q_3	q_2
q_3	q_2	q_2

- q_1 es el estado inicial, y
- $F = \{q_2\}$

Formalidad (yay!)

- Un autómata *M acepta* una cadena *s* si:
  - ⇒ al recibir los símbolos de la cadena, uno por uno, de izquierda a derecha, *M* termina en un estado final.

Informalidad :-(

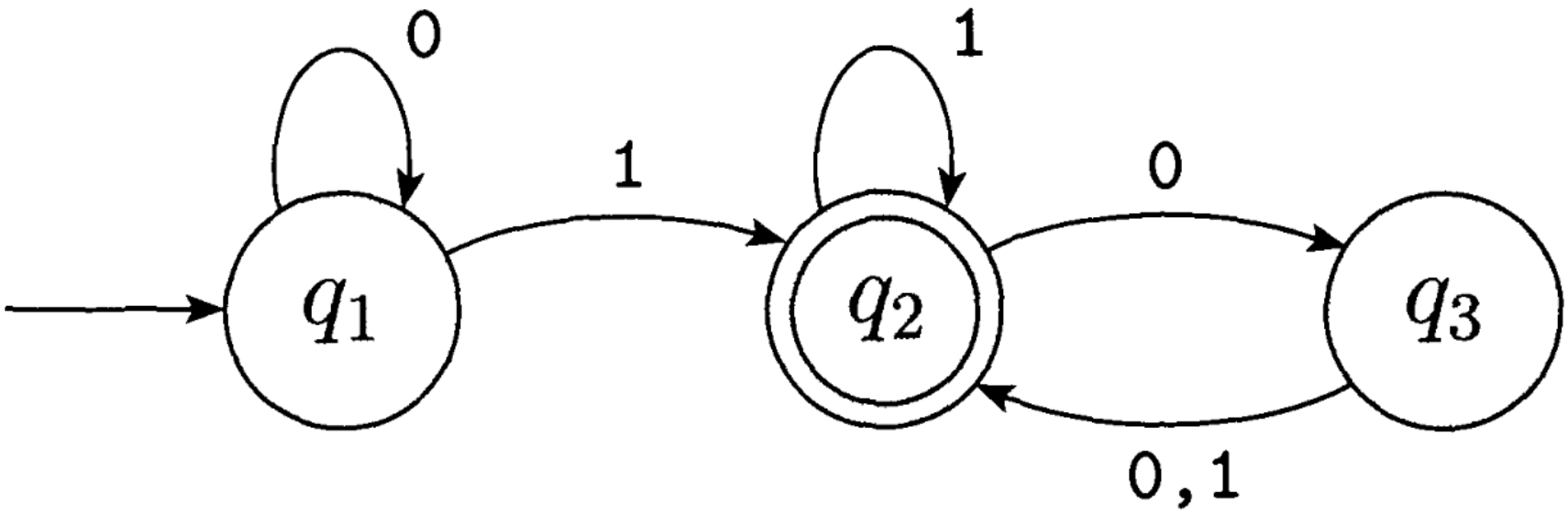

Figura: Diagrama de un autómata finito M_1

¿M_1 acepta la cadena 11012?

Formalidad (yay!)

Lenguaje de un autómata M

Si A es el conjunto de todas las cadenas que M acepta, A es el lenguaje de M y se denota $L(M) = A$.

- $A = \{w | M \text{ acepta } w\}$
- Decimos que M **reconoce** A.
- A es un lenguaje único.

Informalidad :-(

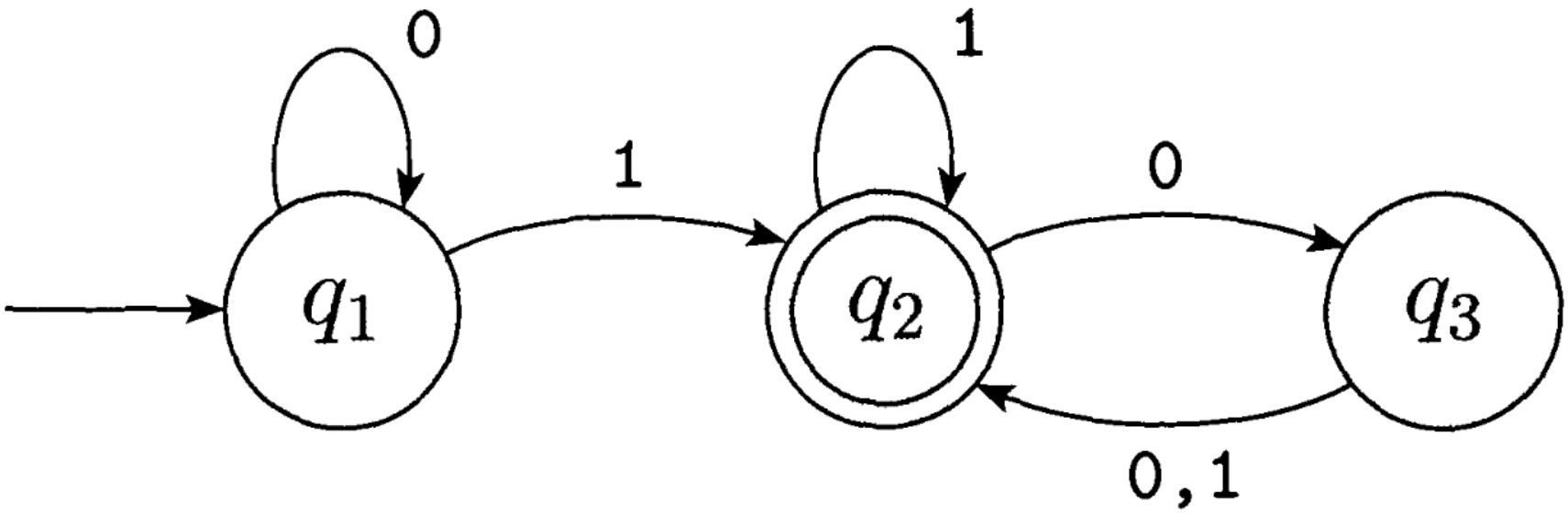

Figura: Diagrama de un autómata finito M_1

¿Qué lenguaje reconoce M_1?

Formalidad (yay!)

$$A = \{w | w \text{ contiene al menos un 1 y termina en 1 o en un número par de 0}\}$$

- $L(M_1) = A$ o, lo que es lo mismo, M_1 reconoce A.

Formalidad (yay!)

Lenguaje Regular

Un **lenguaje regular** R es un lenguaje reconocido por un autómata finito.

- A es regular, puesto que $L(M_1) = A$.
- A se puede describir con una expresión regular.

  $\Rightarrow$ Un lenguaje es regular si y sólo si se puede describir con una expresión regular.

Formalidad (yay!)

Operaciones regulares

Sean A y B lenguajes. Las operaciones regulares *unión*, *concatenación* y *estrella* se definen como sigue:

- Unión: $A \cup B = \{x | x \in A \text{ o } x \in B\}$
- Concatenación: $A \circ B = \{xy | x \in A \text{ e } y \in B\}$
- Estrella: $A^* = \{x_1 x_2 \dots x_k | k \geq 0 \text{ y cada } x_i \in A\}$

Notese que $\epsilon \subseteq A^*$.

Formalidad (yay!)

Operaciones regulares (ejemplo)

Sean $A = \{0, 1\}$ y $B = \{a, b\}$:

- Unión: $A \cup B = \{0, a, 1, b\}$
- Concatenación: $A \circ B = \{0a, 0b, 1a, 1b\}$
- Estrella: $A^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 100, 101, 110, 0110, \ldots\}$

Los lenguajes regulares son cerrados bajo unión, concatenación y estrella.

Expresiones regulares en lenguajes de programación

Usualmente, los lenguajes de programación definen estas tres operaciones con los siguientes operadores:

- Unión: $A \cup B$ como $A|B$
- Concatenación: $A \circ B$ como AB
- Estrella: A^* como $A*$
  - $\Rightarrow$ Estrella no vacía: $A+$, que es un álias de $AA*$

Además, no se usa notación de conjuntos, *i.e.* $\{0\}$ se denota cómo 0.

Formalidad (yay!)

$$A = \{w | w \text{ contiene al menos un 1 y termina en 1 o en un número par de 0}\}$$

- $0 * 1(1|01|00)*$

Informalidad :-(

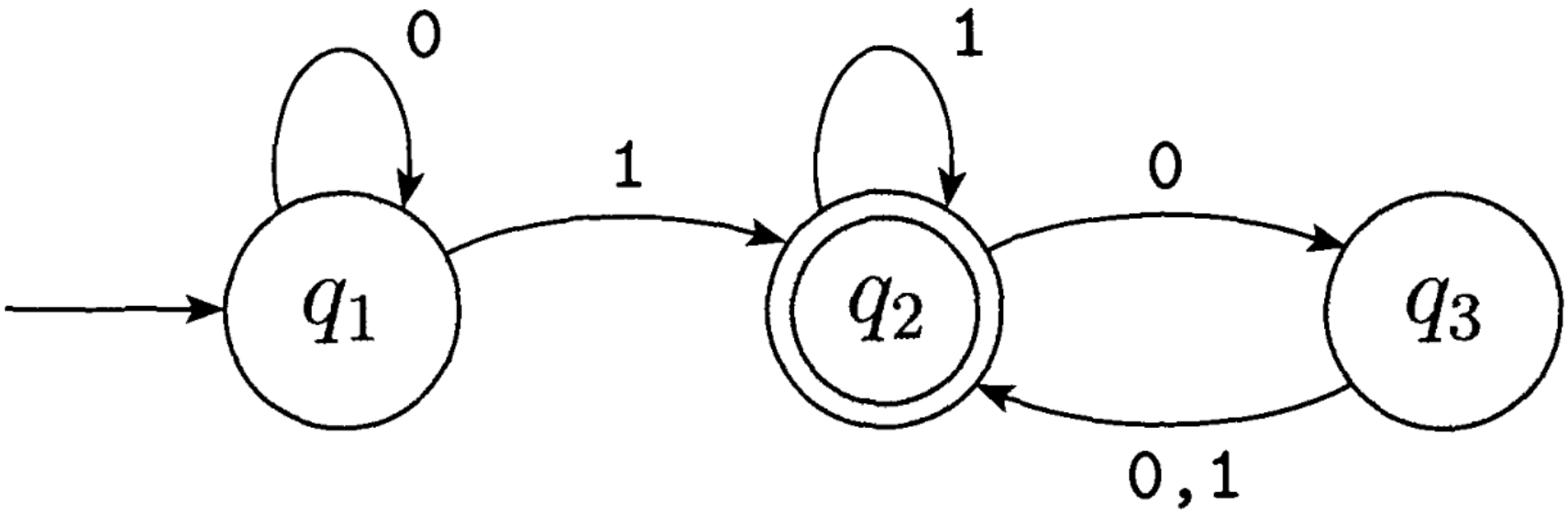

Figura: Diagrama de un autómata finito M_1

$0*1(1|01|00)*$

Limitaciones de las expresiones regulares

- Las expresiones regulares no describen lenguajes libres de contexto.
  - $\Rightarrow$ El lenguaje de parentesis balanceados **es libre de contexto**.
  - $\Rightarrow$ (())()((((())())())))
  - $\Rightarrow$ $S \rightarrow SS \mid (S) \mid \epsilon$

Encuentre la expresión regular

Sobre el alfabeto $\Sigma = \{0, 1\}$:

$$A = \{w | \text{ cada 0 en } w \text{ está seguido de por lo menos un } 1\}$$

$(\epsilon \in A)$

- ¿Cuál es su expresión regular?

Expresiones regulares en Python

```
import re

txt = "000001000"
x = re.search("0*10*", txt)
```

Módulo `re`

- `re.findall`: Regresa una lista de objetos Match, con todas las subcadenas que están en el lenguaje de la expresión regular.
- `re.search`: Regresa el primer Match.
- `re.split`: Corta la subcadena en cada Match y regresa la lista.
- `re.sub`: Reemplaza cada Match por otra cadena.

Módule `re`

Metacaracteres:

[ ]	Conjunto de caracteres	“[a-z]”
\	Secuencia de escape	“\n”
.	Cualquier caracter (salvo \n)	“ce..o”
^	Comienza con	“^La”
$	Termina con	“[.]txt$”
*	Estrella	“a∗”
+	Estrella no vacía	“a+”
{ }	Número exacto de ocurrencias	“cer{2}o”
|	Unión	“a|A”
( )	Grupo	“(a|A)rco”

Módulo `re`

\b	Límite de palabra
\d	Cualquier dígito decimal ([0-9])
\D	Cualquier caracter que no es un dígito ([^0-9])
\s	Cualquier caracter de espacio en blanco ([\t\n\r\f\v])
\S	Cualquier caracter que no es de espacio en blanco ([^\t\n\r\f \v])
\w	Cualquier caracter alfanumérico ([a-zA-Z0-9_])
\W	Cualquier caracter no alfanumérico ([^a-zA-Z0-9_])
\t	Tabulación
\n	Salto de línea

Contenidos

Extracción de información con RegEx

- Encuentra todos los *handles* de Twitter o *hashtags*.
  $\Rightarrow$ `[@#]\w+`
- Encuentra verbos en infinitivo (más de 3 caracteres).
  $\Rightarrow$ `\b\w+(ar|er|ir)\b`
- ¿Cómo podemos encontrar cosas como "*barqueando*"?

Extracción estadística

- Basada en frecuencias de un corpus
  - ⇒ Probabilidad simple
    - → $P(t) = \frac{C(t)}{\sum_{t'} C(t')}$
  - ⇒ Probabilidad condicional (diferencia entre grupos)
    - → $P(t \mid k) = \frac{C'(t,k)}{\sum_{t'} C'(t',k)}$
  - ⇒ Donde:
    - → $t, t' \in T$ son términos.
    - → $k \in K$ es una categoría, *e.g.* $K = \{Hombre, Mujer\}$.
    - → $C : T \to \mathbb{N}$ es la frecuencia de un término.
    - → $C' : T \times K \to \mathbb{N}$ es la frecuencia de un término dada una clase.

tf-idf

- Es una medida que evalúa la *importancia* de un término en un documento, con respecto de una colección de éstos.
  - $\Rightarrow$ Si hablamos de un sólo texto, cada oración puede ser considerada un documento único.
- Frecuencia de termino (tf) - frecuencia inversa de documento (idf)
  - $\Rightarrow$ tf: mide la frecuencia de un término
    - $\rightarrow$ $tf(t,d) = \frac{C(w,d)}{\sum_{w'} C(w',d)}$
  - $\Rightarrow$ idf: mide la "importancia" relativa de un término
    - $\rightarrow$ $idf(t,D) = log(\frac{|D|}{|\{d \in D \mid t \in d\}|})$
  - $\Rightarrow$ $tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$

tf-idf

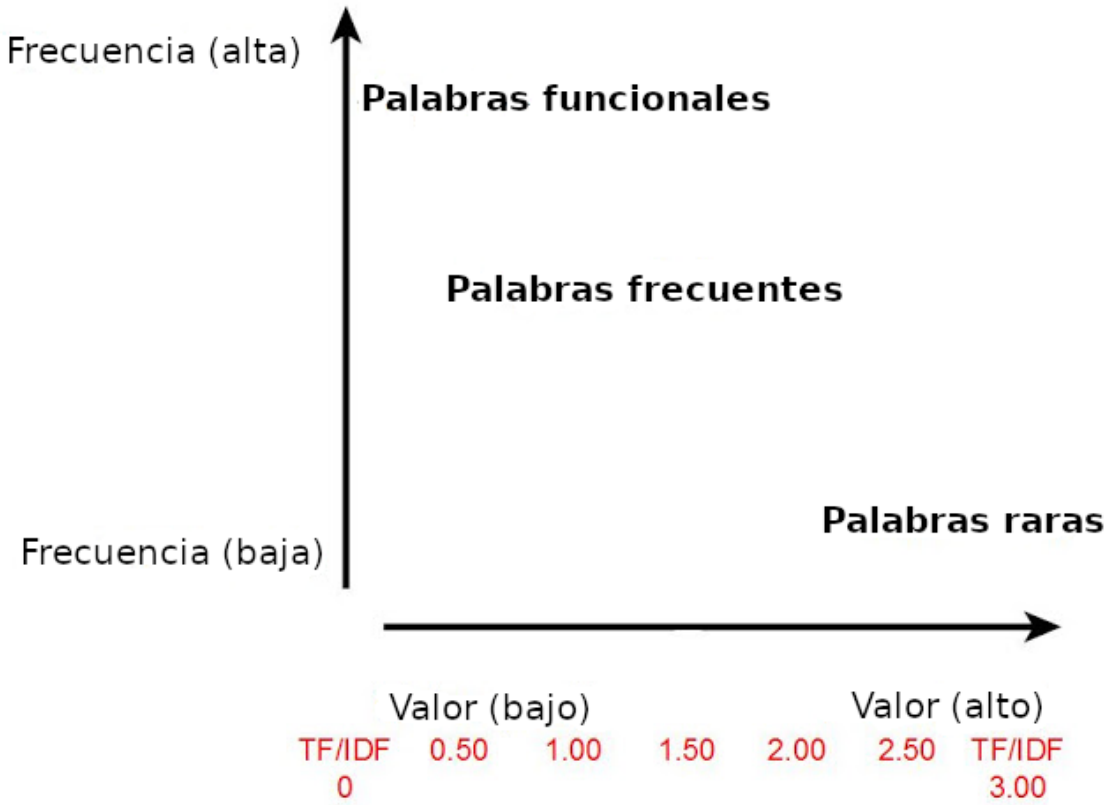

Figura: Transformación tf-idf

PageRank

- El algoritmo original de Google.
- Asigna una importancia a un sitio web con respecto de:
  - ⇒ sus links de entrada/salida
  - ⇒ la *calidad* de los links (*i.e.* el *PR* de los sitios web vecinos)
- Es iterativo.
  - ⇒ Todos los sitios empiezan con el mismo valor de *PR* (la suma de *PR* de todos los nodos es 1,0).
  - ⇒ Con cada reindexación, se recalculan los *PR*.

PageRank (versión simple)

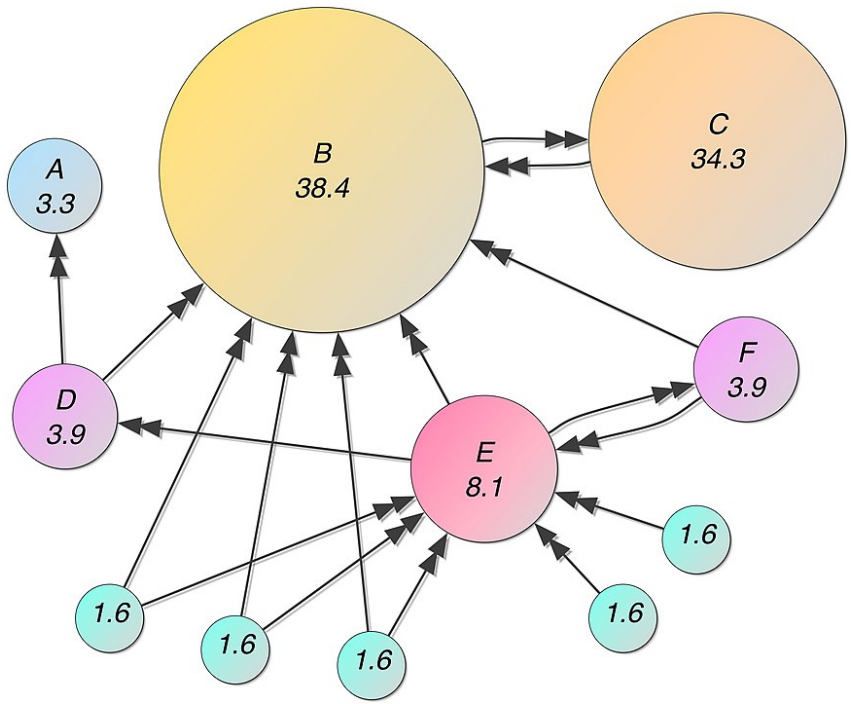

Figura: Ejemplo de PageRank[1]

[1] https://en.wikipedia.org/wiki/PageRank

PageRank (versión simple)

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

- u es un sitio web
- B_u es el conjunto de todos los sitios web que tienen hipervínculos a u
- $L(v)$ es el número total de hipervínculos que tiene v.

PageRank con términos (versión simple)

- Cada palabra es un nodo.
- Los "hipervínculos" se modelan por co-ocurrencia.
  - $\Rightarrow$ El texto se divide en oraciones.
  - $\Rightarrow$ Se lematiza/remueven stopwords.
  - $\Rightarrow$ Dos palabras que aparecen juntas en la misma oración, tienen un vínculo bi-direccional (gráfica no dirigida).
- Se aplica PageRank en la gráfica de co-ocurrencia.
  - $\Rightarrow$ El criterio de parada de la iteración suele ser el cambio global del sistema.

PageRank con términos (versión simple)

Ejemplo

Venus es el segundo planeta del sistema solar en orden de distancia desde el Sol, y el tercero en cuanto a tamaño, de menor a mayor. Al igual que Mercurio, carece de satélites naturales. Recibe su nombre en honor a Venus, la diosa romana del amor. Se trata de un planeta de tipo rocoso y terrestre, llamado con frecuencia el planeta hermano de la Tierra, ya que ambos son similares en cuanto a tamaño, masa y composición, aunque totalmente diferentes en cuestiones térmicas y atmosféricas.

PageRank (avanzado)

- Los algoritmos más modernos usan aristas con pesos.
  - $\Rightarrow$ Para texto, se representan los nodos con vectores y los pesos son por distancia.
- La ecuación tiene pesos para equilibrar la importancia
  - $\Rightarrow$ Evita que un sitio web con PR excesivamente grande cobre demasiada importancia.

Test de Independencia de χ^2

- ¿La palabra México es significativamente más frecuente en un texto sobre historia de México?
  - $\Rightarrow$ Intuitivamente si.
  - $\Rightarrow$ ¿Cómo lo medimos?

Test de Independencia de χ^2

El test de independencia de χ^2 permite comparar dos variables categóricas (frecuencias) en una tabla de contingencia.

- Intenta determinar si están relacionadas.
- Permite saber si sus distribuciones difieren.
  - Un valor de χ^2 muy pequeño: las distribuciones son muy similares (✓ relación).
  - Un valor de χ^2 muy grande: las distribuciones son muy distintas (✗ relación).

Test de Independencia de χ^2

Estadística χ^2

$$\chi^2 = \sum_{i \in C}^{|C|} \frac{(O_i - E_i)^2}{E_i}$$

dónde C es la categoría.

Test de Independencia de χ^2

- Hipótesis del test:
  - H_0: No hay relacion entre las dos variables (son independientes).
    - Una variable "no dice nada" de la otra (tienen los mismos niveles).
  - H_1: Hay relación entre las dos variables.

Test de Independencia de χ^2

- χ^2 pregunta:
  - $\Rightarrow$ ¿Qué tan diferentes son las frecuencias observadas contra las esperadas si asumimos independencia?
  - $\Rightarrow$ Revisa dos variables:
    - $\rightarrow$ Historia o $\sim$Historia
    - $\rightarrow$ Palabra o $\sim$Palabra

Distribución de χ^2

OBS	México	~México	Total
Historia	75	22679	22754
General	32084	152530921	152563005
Total	32159	152553600	152585759
EXP	México	~México	
Historia	4.795636833	22749.20436	
General	32154.20436	152530851	
OBS-EXP	México	~México	
Historia	70.20436317	-70.2043632	
General	-70.2043632	70.20436317	
(OBS-EXP)^2	México	~México	
Historia	4928.652608	4928.652608	
General	4928.652608	4928.652608	
((OBS-EXP)^2)/EXP	México	~México	Total
Historia	1027.736832	0.216651648	1027.953484
General	0.153281747	3.23125E-05	0.153314059
			1028.106798

Distribución de χ^2

OBS	de	~de	Total
Historia	1673	21081	22754
General	9999518	142563487	152563005
Total	10001191	142584568	152585759
EXP	de	~de	
Historia	1491.404581	21262.59542	
General	9999699.595	142563305	
OBS-EXP	de	~de	
Historia	181.5954187	-181.595419	
General	-181.595419	181.5954187	
(OBS-EXP)^2	de	~de	
Historia	32976.89609	32976.89609	
General	32976.89609	32976.89609	
((OBS-EXP)^2)/EXP	de	~de	Total
Historia	22.11130132	1.550934655	23.66223598
General	0.003297789	0.000231314	0.003529103
			23.66576508

Distribución de χ^2

	CorpusHist	O	E	O-E	(O-E)^2	(O-E)^2/E
de	1673	73525.533972049	65545.55	7979.98397	63680144.2	971.540313
la	1076	47288.388854707	41148.59	6139.79885	37697130	916.122034
el	772	33928.100553749	29953.48	3974.62055	15797608.5	527.404781
y	769	33796.25560341	27401.19	6395.0656	40896864.1	1492.52146
en	732	32170.16788257	27755.16	4415.00788	19492294.6	702.294442
a	468	20567.812252791	21375.03	-807.217747	651600.491	30.4841907
los	453	19908.587501099	17164.95	2743.6375	7527546.74	438.541722
del	416	18282.499780259	12173.87	6108.62978	37315357.8	3065.20094
que	413	18150.65482992	30688.85	-12538.1952	157206338	5122.58811
se	318	13975.564735871	13257.31	718.254736	515889.866	38.9136156
las	289	12701.063549266	11056.37	1644.69355	2705016.87	244.656869
por	236	10371.802759954	10238.07	133.73276	17884.4511	1.74685767
con	231	10152.061176057	9711.74	440.321176	193882.738	19.9637488
al	169	7427.26553573	6234.03	1193.23554	1423811.04	228.393358
gobierno	163	7163.5756350532	740.77	6422.80564	41252432.2	55688.5838
como	158	6943.8340511558	5069.96	1873.87405	3511403.96	692.590072
su	150	6592.2475169201	7234.06	-641.812483	411923.263	56.9421961
un	149	6548.2992001406	10879.95	-4331.6508	18763198.7	1724.56663
más	131	5757.2294981102	4337.33	1419.8995	2016114.58	464.828497
una	123	5405.6429638745	8833.36	-3427.71704	11749244.1	1330.09909
no	107	4702.469895403	9606.18	-4903.7101	24046372.8	2503.21905
para	106	4658.5215786235	6962.26	-2303.73842	5307210.71	762.282752
era	94	4131.1417772699	1441.63	2689.51178	7233473.6	5017.56595
lo	87	3823.5035598137	5682.77	-1859.26644	3456871.7	608.307515
federal	86	3779.5552430342	76.77	3702.78524	13710618.6	178593.442
méxico	75	3296.1237584601	210.3	3085.82376	9522308.27	45279.6399
país	74	3252.1754416806	685.42	2566.75544	6588233.5	9611.96565

Distribución de χ^2

	CorpusHist	O	E	O-E	(O-E)^2	(O-E)^2/E
federal	86	3779.5552430342	76.77	3702.78524	13710618.6	178593.442
gobierno	163	7163.5756350532	740.77	6422.80564	41252432.2	55688.5838
méxico	75	3296.1237584601	210.3	3085.82376	9522308.27	45279.6399
país	74	3252.1754416806	685.42	2566.75544	6588233.5	9611.96565
que	413	18150.65482992	30688.85	-12538.1952	157206338	5122.58811
era	94	4131.1417772699	1441.63	2689.51178	7233473.6	5017.56595
del	416	18282.499780259	12173.87	6108.62978	37315357.8	3065.20094
no	107	4702.469895403	9606.18	-4903.7101	24046372.8	2503.21905
un	149	6548.2992001406	10879.95	-4331.6508	18763198.7	1724.56663
y	769	33796.25560341	27401.19	6395.0656	40896864.1	1492.52146
una	123	5405.6429638745	8833.36	-3427.71704	11749244.1	1330.09909
de	1673	73525.533972049	65545.55	7979.98397	63680144.2	971.540313
la	1076	47288.388854707	41148.59	6139.79885	37697130	916.122034
para	106	4658.5215786235	6962.26	-2303.73842	5307210.71	762.282752
en	732	32170.16788257	27755.16	4415.00788	19492294.6	702.294442
como	158	6943.8340511558	5069.96	1873.87405	3511403.96	692.590072
lo	87	3823.5035598137	5682.77	-1859.26644	3456871.7	608.307515
el	772	33928.100553749	29953.48	3974.62055	15797608.5	527.404781
más	131	5757.2294981102	4337.33	1419.8995	2016114.58	464.828497
los	453	19908.587501099	17164.95	2743.6375	7527546.74	438.541722
las	289	12701.063549266	11056.37	1644.69355	2705016.87	244.656869
al	169	7427.26553573	6234.03	1193.23554	1423811.04	228.393358
su	150	6592.2475169201	7234.06	-641.812483	411923.263	56.9421961
se	318	13975.564735871	13257.31	718.254736	515889.866	38.9136156
a	468	20567.812252791	21375.03	-807.217747	651600.491	30.4841907
con	231	10152.061176057	9711.74	440.321176	193882.738	19.9637488
por	236	10371.802759954	10238.07	133.73276	17884.4511	1.74685767

Información Mutua Puntual (PMI)

- Es una medida de asociación entre una característica (término) y una clase (categoría).

$$pmi(t;c) = log(\frac{p(t,c)}{p(t)p(c)})$$

- Contesta a la pregunta ¿cuánta información da t sobre la clase c?
  $\Rightarrow$ Mientras mayor $pmi(t;c)$, menor incertidumbre de que t ocurra en c.

Información Mutua Puntual (PMI)

$$pmi(\text{México;Historia}) = log(\frac{\frac{75}{152585759}}{\frac{32159}{152585759} \cdot \frac{22754}{152585759}}) = 2{,}7635$$

$$pmi(\text{México;General}) = log(\frac{\frac{32084}{152585759}}{\frac{32159}{152585759} \cdot \frac{152563005}{152585759}}) = -0{,}0021$$

Contenidos

Relaciones léxicas

- Se refiere a las relaciones que existen entre los significados de las palabras léxicas.
- Pueden ser de diferentes tipos:
  - ⇒ Homonímia. Relación entre diferentes significados que comparten una forma.
    - → *e.g.* banco
  - ⇒ Sinonimia. Relación entre diferentes formas que comparten significado.
    - → *e.g.* barco y buque
  - ⇒ Meronimia. Relación que denota los constituyentes o miembros de algo.
    - → *e.g.* automovil → volante
    - → son relaciones "tiene un"
  - ⇒ Troponimia. Relación que denota la "manera" entre verbos.
    - → *e.g.* comer → atragantarse
    - → son relaciones "es una forma de"

Relaciones de hiponimia

- Hiperonimia e hiponimia. Definición y pertenencia a un campo semántico, respectivamente
  - $\Rightarrow$ *e.g.* color (hiperónimo) $\rightarrow$ rojo (hipónimo)
  - $\Rightarrow$ son relaciones "es un"
- Nos dan una idea sobre como representar diferencias en el significado de las palabras.
- Definen similitud por medio de la distancia en la jerarquía.

Redes semánticas

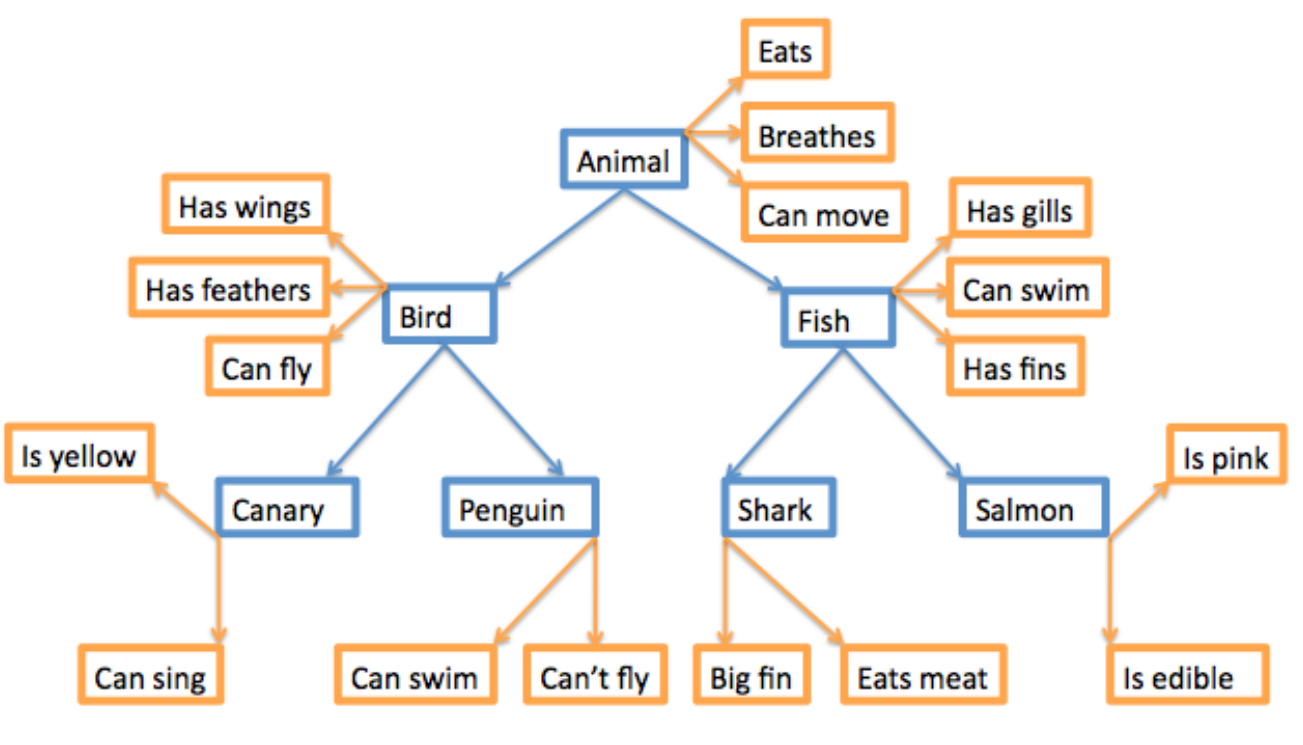

Figura: Red semántica[2]

[2] https://en.wikipedia.org/wiki/Semantic_lexicon

Patrones de Hearst[3]

- Método clásico de extracción de relaciones de significado.
  - $\Rightarrow$ Se usan patrones léxico-sintácticos para definir relaciones.
- Requiere conocimiento lingüístico.
  - $\Rightarrow$ Análisis sintáctico superficial.
  - $\Rightarrow$ Conocimiento de estructura discursiva.

[3] Hearst, M. A. "Automatic acquisition of hyponyms from large text corpora." Proceedings of the 14th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1992.

Patrones de Hearst

(S1) The bow lute, such as the Bambara ndang,
is plucked and has an individual
curved neck for each string.

Queremos establecer relaciones de hiponimia.

Patrones de Hearst

(1a) NP_0 *such as* $\{NP_1, NP_2 \ldots, (and \mid or)\}$ NP_n

are such that they imply

(1b) *for all* NP_i, $1 \leq i \leq n$, *hyponym*(NP_i, NP_0)

Thus from sentence (S1) we conclude

hyponym("*Bambara ndang*", "*bow lute*").

Patrones de Hearst

(2) *such NP as {NP ,}* {(or | and)} NP*

... works by such authors as Herrick, Goldsmith, and Shakespeare.

⇒ *hyponym("author", "Herrick"), hyponym("author", "Goldsmith"), hyponym("author", "Shakespeare")*

Patrones de Hearst

(3) *NP* {, *NP*}* {,} *or other NP*

Bruises, wounds, broken bones or other injuries ...

$\Longrightarrow$ *hyponym*("*bruise*", "*injury*"),
hyponym("*wound*", "*injury*"),
hyponym("*broken bone*", "*injury*")

Patrones de Hearst

(4) *NP* {, *NP*}* {,} *and other NP*

... temples, treasuries, and other important civic buildings.

$\Longrightarrow$ *hyponym*(“*temple*”, “*civic building*”), *hyponym*(“*treasury*”, “*civic building*”)

Patrones de Hearst

(5) *NP* {,} *including* {*NP* ,}* {*or* | *and*} *NP*

All common-law countries, including Canada and England ...

$\Longrightarrow$ *hyponym*(“*Canada*”, “*common-law country*”), *hyponym*(“*England*”, “*common-law country*”)

Patrones de Hearst

(6) *NP* {,} *especially* {*NP* ,}* {*or* | *and*} *NP*

... most European countries, especially France, England, and Spain.

$\Longrightarrow$ *hyponym*(“*France*”, “*European country*”),
hyponym(“*England*”, “*European country*”),
hyponym(“*Spain*”, “*European country*”)

Patrones de Hearst

Consideraciones para encontrar nuevos patrones:

1. Decidir una relación léxica de interés, *e.g.* grupo/miembro.
2. Colectar una lista de terminos para los cuales esta relación se cumple, *e.g.* Inglaterra-país.
   - $\Rightarrow$ Lexicon
   - $\Rightarrow$ Base de conocimiento
   - $\Rightarrow$ Observaciones previas.
3. Encontrar ejemplos en el corpus donde se encuentren estos términos.
   - $\Rightarrow$ Queremos identificar marcadores discursivos, verbos, etc. que formen patrones comunes.
4. Encontrar los patrones comunes entre muchos pares de términos.
5. Al encontrarse los nuevos patrones, se pueden usar para identificar nuevos pares automáticamente.

Contenidos

Modelo Estadístico de Lenguaje

- Un modelo estadístico de lenguaje asigna una distribución de probabilidad a una secuencia de elementos.
  - $\Rightarrow$ Sean tipos de un vocabulario o símbolos de un alfabeto.
- $P(\text{“el gobierno mexicano”}) = \frac{9}{22668} = 0{,}0003970$
- $P(\text{“otros empresarios formaron”}) = \frac{1}{22668} = 0{,}0000441$

Modelo Estadístico de Lenguaje

- Traducción
  - $\Rightarrow$ “blood pressure”: $P(\text{“presión sanguínea”}) > P(\text{“presión de la sangre”})$
- Corrección ortográfica
  - $\Rightarrow$ $P(\text{“presión sanguínea”}) > P(\text{“preción sanguínea”})$
- Corrección de estílo
  - $\Rightarrow$ $P(\text{“la sangre, hace presión”}) > P(\text{“la sangre hace presión”})$
- Generación de texto
  - $\Rightarrow$ $P(\text{“la presión sanguínea”}), P(\text{“la presión ejercida”}),$ $P(\text{“la presión hidrostática”})$
- $\vdots$

Modelo Estadístico de Lenguaje

$$P(W) = P(w_1, w_2, w_3, \ldots, w_n)$$

dónde:

- W es una secuencia de w_i
- $w_1, \ldots, w_n \in V$, $|V| = n$

Unigramas

$$P(w_1, w_2, w_3, \ldots, w_n) = \Pi_i P(w_i)$$

- La probabilidad de cada $w_1, w_2, \ldots$ es independiente.
- Asume que las oraciones con palabras frecuentes tienen alta probabilidad.

$$P(w) = \frac{c(w, D)}{|D|}$$

Modelo de n-gramas

$$P(w_n|w_1, w_2, \ldots, w_{n-1}) = P(w_n)\Pi_i P(w_i|w_{i-1})$$

- Los unigramas asumen distribuciones de palabras irreales.
  $\Rightarrow$ Las palabras tienen contexto.
- Los n-gramas permiten capturar ese contexto.
- Asume la propiedad de Markov.
  $\Rightarrow$ El valor de $P(w_n)$ depende sólo de los $n-1$ elementos anteriores.

$$P(w_n|w_1, \ldots, w_{n-1}) = \frac{c(w_1, \ldots, w_{n-1}, w_n, D)}{c(w_1, \ldots, w_{n-1}, D)}$$

Smoothing

$$P_{Laplace}(w_n|w_1,\ldots,w_{n-1}) = \frac{c(w_1,\ldots,w_{n-1},w_n,D)+1}{c(w_1,\ldots,w_{n-1},D)+|V|}$$

- $P(\text{“futuro es ahora”}) = \frac{c(\text{“futuro es ahora”, D})}{c(\text{“futuro es”, D})} = \frac{0}{0} = \infty$
- $P_{Laplace}(\text{“futuro es ahora”}) = \frac{c(\text{“futuro es ahora”, D})+1}{c(\text{“futuro es”, D})+22670} = \frac{1}{22670}$
  $= 0{,}00004411$