PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Introducción a la lingüística computacional

César Antonio Aguilar

Facultad de Lenguas y Letras

05/09/2017

Cesar.Aguilar72@gmail.com

Análisis probabilístico (1)

En lingüística computacional es usual realizar análisis combinando modelos lingüístico y métodos probabilísticos. Al respecto, podemos decir que lo que intentamos es responder a la siguiente pregunta: ¿todo lo que pasa en un lenguaje humano es fruto del azar? Aquí no es casual usar la palabra **azar**. Tratemos de definirla realizando el siguiente experimento:

Primer experimento: supongamos que tenemos un dado y nos preguntamos, ¿cuántas veces puede caer el número 4 en 6 tiradas?

Delimitemos:

1. Un dado tiene 6 caras.
2. Cada cara representa un número del 1 al 6.
3. En teoría, en 6 tiradas me puede salir al menos un número del 1 al 6.

Análisis probabilístico (2)

Lo anterior se puede representar en una tabla del tipo:

Tiradas						
1	1/6	1/6	1/6	1/6	1/6	1/6
2						
3						
4						
5						
6	1	2	3	4	5	6

Caras del dado

Análisis probabilístico (3)

Algunas definiciones:

- A nuestra serie de tiradas del dado (esto es, 6 las veces que lo arrojamos para que nos salga el número 6, conforman un *espacio simple* (o *Es*).

- Cada tirada puede ser vista como un *fenómeno* o *hecho* en donde podemos confirmar o no si sale el número 6.

- Si lanzamos 6 veces nuestro dado, y pensamos que por lo menos 1 vez va a caer el número 6, eso quiere decir que las posibilidades de que salga este número es de 1/6, o lo que es lo mismo, 0.16.

Análisis probabilístico (4)

Segundo experimento: supongamos que ahora jugamos con una moneda, y queremos ver si en tres lanzamientos nos sale dos veces Sol.

Condiciones:

1. Nuestra moneda tiene dos lados: Águila y Sol.
2. Nuestra moneda no está “calada”.
3. En teoría, la posibilidad de que se cumpla el hecho que de caiga Águila o Sol es de un 50% para cada uno, o sea 0.5.

Análisis probabilístico (5)

Para decirlo de una manera más “elegante”:

- Hay dos hechos posibles, Águila (Ag) y Sol (So)
- Espacio simple *Es* es {H,T}
- Al lanzar una moneda al aire, suponemos que hay las mismas probalidades de que caiga Ag o So.
- La suma de estas dos posibilidades nos da 1.
- P({Ag}) = 0.5
- P({So}) = 0.5
- P({Ag,So}) = P({Ag})+P({So}) = 1.0

Análisis probabilístico (6)

- Lanzemos 3 veces nuestra moneda.
- Las combinaciones de lanzamientos conforma nuestro *Es*, y éste es representable como una cadena.

- *Es* ={AgAgAg; AgAgSo; AgSoAg; SoAgAg; AgSoSo; SoAgSo; SoSoAg; SoSoSo}

- Asumimos que cada hecho es *equiprobable*.
- **Pregunta:** ¿cuántas probabilidades tenemos de que caiga 2 veces Sol?
- A = {SoSoAg; SoAgSo; AgSoSo}
- P(A) = P({SoSoAg}) + P({SoAgSo}) + P({SoAgSo})
- = 1/8 + 1/8 + 1/8
- =3/8

Análisis probabilístico (7)

- Una duda: ¿y si nuestra modena está trucada y no permite que caiga Sol más que una sola vez?

- Nuestras opciones de combinación se reducen, por ejemplo:

- *Es* ={AgAgAg; AgAgSo; AgSoAg; SoAgAg; AgSoSo; SoAgSo; SoSoAg; SoSoSo}

- **Pregunta:** ¿cuántas probabilidades tenemos de que caiga Sol por lo menos una sola vez?
- B = {AgAgSo; AgSoAg; SoAgAg}
- P(B) = P({AgAgSo}) + P({AgSoAg}) + P({SoAgAg})
- = 2/8 + 2/8 + 2/8
- =6/8

Análisis probabilístico (8)

Tiradas

Ag								
Ag								
Ag								
Ag			2/8					
Ag		2/8						
Ag	2/8							
Ag								
Ag	So	So	So	So	So	So	So	So

Caras del dado

Probabilidades independientes y condicionales (1)

Para nuestros análisis, vamos a tomar en cuenta dos tipos de probabilidades: una en donde suponemos que no hay una relación de ningún tipo entre dos (o más variables) para que se dé un fenómeno. Decimos que tales fenómenos están regidos por el azar (y de manera absoluta).

Algo tan simple como decir: *me duelen las muelas del juicio, seguro es que va a llover hoy*, es postular justo un fenómeno que, si se llegara a cumplir, no veríamos una relación clara entre las variables que intervienen en él. La fórmula para representar esto es:

$$P(A \text{ and } B) = P(A) \times P(B)$$

Probabilidades independientes y condicionales (2)

En cambio, si hay una relación subyacente entre dos variables, y ésta condiciona la manifestación de un fenómeno, entonces hablamos de **probabilidades condicionales**. En otras palabras, lo que queremos señalar es que cada vez que se presenta una asociación entre dos variables (A y B), observamos una alta probabilidad de que ocurra un fenómeno. Así, las dos variables se influyen mutuamente. La fórmula para presentar esto es:

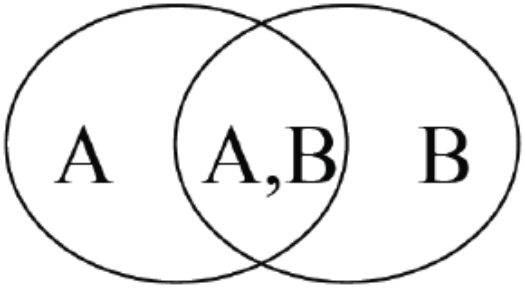
$$p(B|A) = p(A \cap B) / p(A)$$

Volviendo a nuestro juego de cartas, podríamos decir que conforme resolvamos quién tiene las Reinas que nos faltan para nuestro póker, eso condiciona que podamos formar la secuencia ganadora, en la menor cantidad de manos posibles.

Probabilidades independientes y condicionales (3)

Si nos ponemos más formales, una probabilidad condicional equivale a decir que:

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$



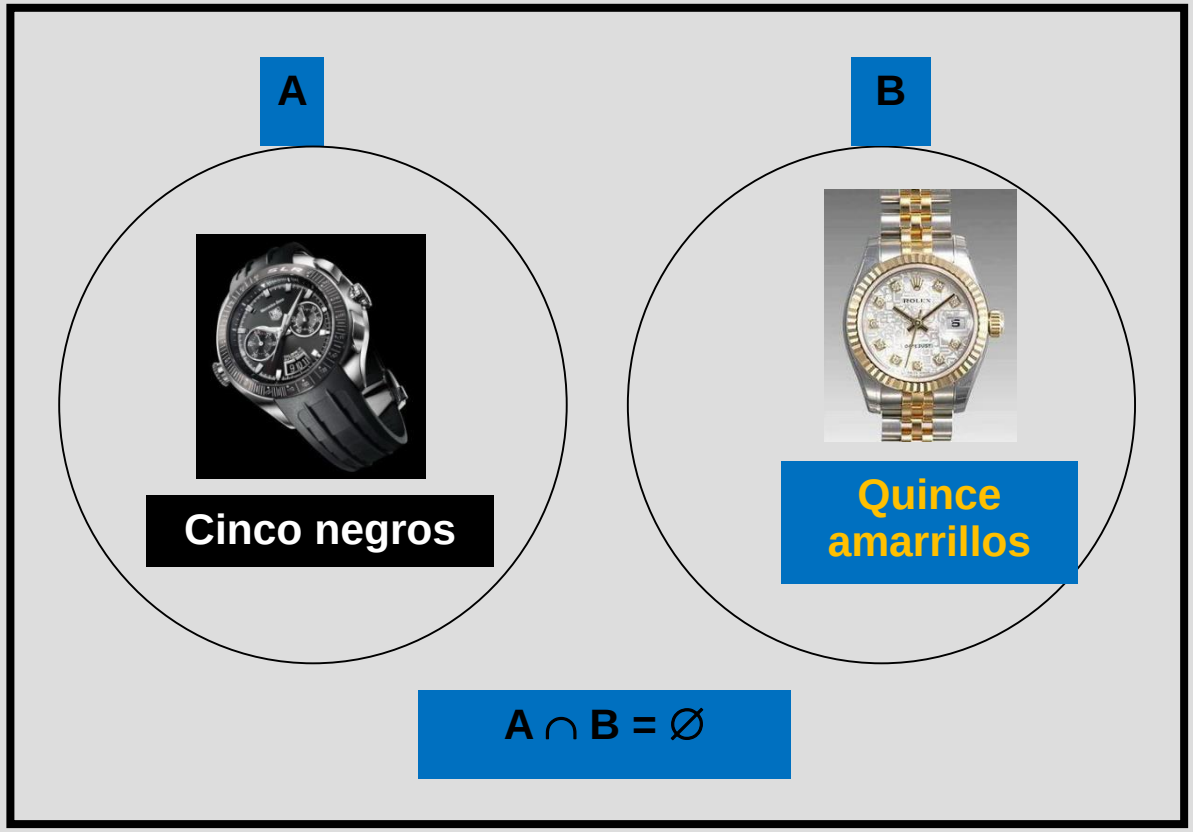
Recordando las matemáticas de la secundaria, esto se puede representar con un diagrama de Venn:

Hay que notar que: $P(A,B)=P(A|B) \cdot P(B)$

También: $P(A,B) = P(B,A)$

Probabilidades independientes y condicionales (4)

Para entender mejor estas probabilidades, veamos el siguiente ejemplo: supongamos que tenemos un conjunto de relojes de la siguiente forma:



Probabilidades independientes y condicionales (5)

Si calculamos su probabilidad independiente, decimos que:

$$P(A \cup B) = P(A) + P(B)$$

Esto quiere decir que *la probabilidad la unión de los conjuntos A y B es igual a la suma de la probabilidad de A y la probabilidad de B*.

Si esto es así, lo que estamos suponiendo es que entre los conjuntos A y B no tienen elementos en común, aunque forman parte de un mismo universo. En otras palabras, nuestro conjunto Universal son 20 relojes, pero los separamos a partir de un rasgo: su color, lo que nos permite generar dos conjuntos específicos.

Probabilidades independientes y condicionales (5)

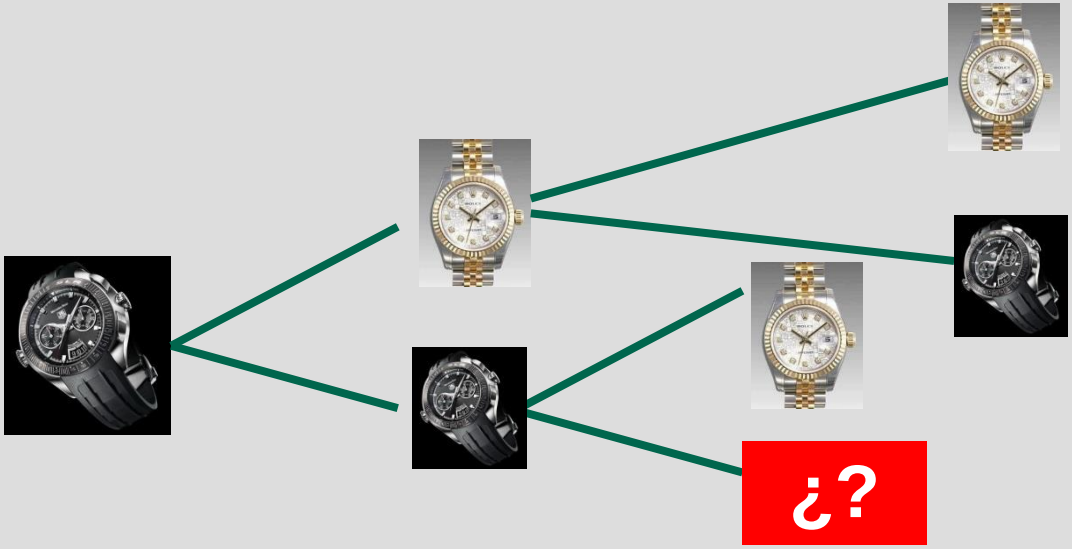
Veamos ahora el caso contrario: supongamos que los eventos A y B tienen un punto de cruce entre sí, de tal manera que forman una intersección.

Si quisiera calcular de probabilidad dos eventos no disjuntos A y B, debo excluir su punto de intersección, esto es:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probabilidades independientes y condicionales (6)

Volviendo a la extracción de relojes, si sacara un reloj negro del conjunto A, y me quedo con 4, ¿conservo las mismas probabilidades para realizar mi experimento, o éstas se ven afectadas por tal decisión?



Probabilidades independientes y condicionales (7)

Cuando realizamos este experimento, suponiendo que tenemos las mismas condiciones entre la primera y la segunda extracción, los escenarios que tenemos son:

Negros	Amarillos	P(negros)	P(amarillos)	# Experimento
5	15	5/20	15/20	Exp. 1 = 1N + 2 A
4	13	4/17	13/17	Exp. 2 = 1N + 2 A
3	11	3/14	11/14	Exp. 3 = 1N + 2 A

A este hecho de forzar las probabilidades de los resultados esperados en cada experimento es lo que llamamos condicionalidad, lo que da pie a las llamadas probabilidades condicionales.

Probabilidades independientes y condicionales (8)

Una forma de definir las probabilidades condicionales es la siguiente:

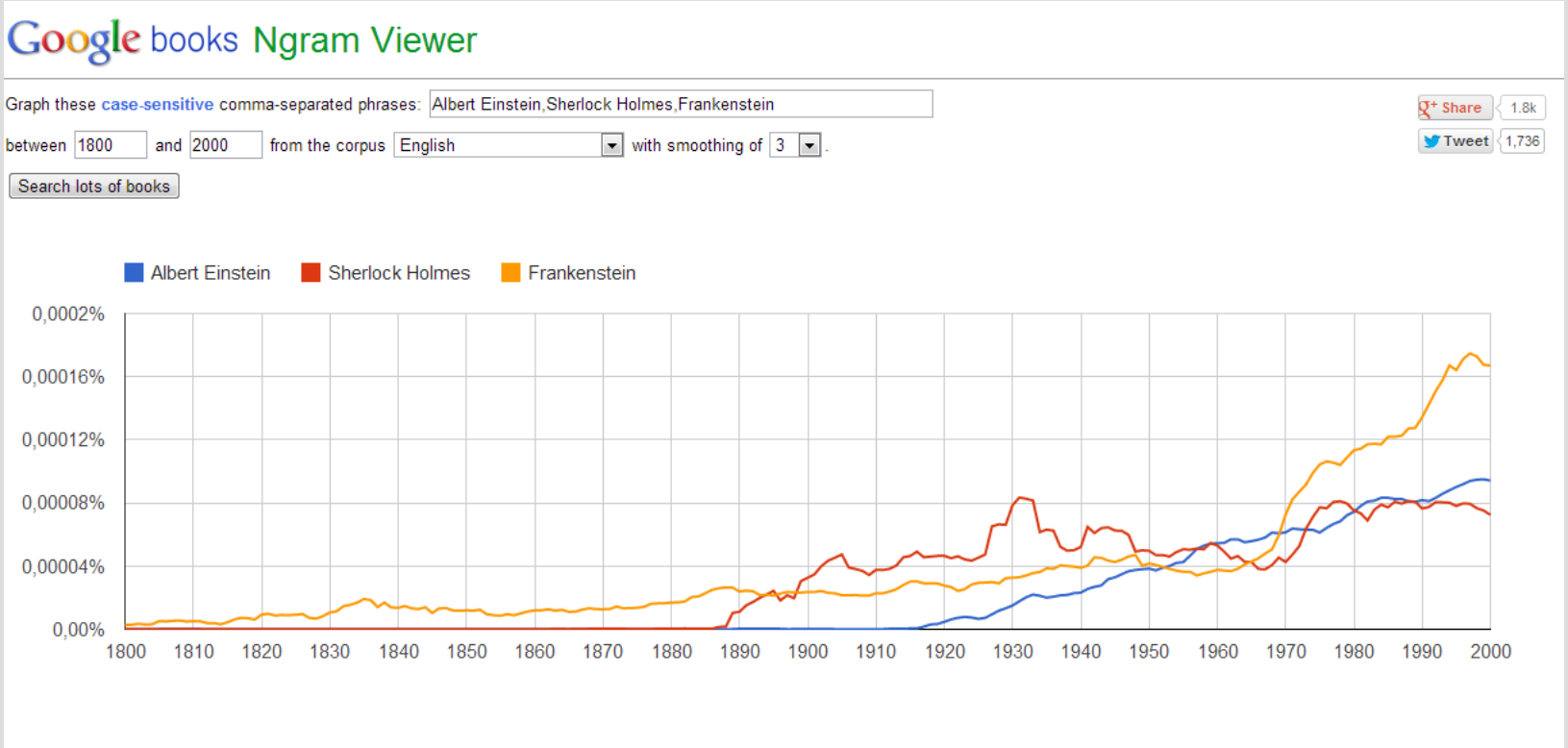
- Sean A y B dos eventos estadísticamente dependientes, entonces, la probabilidad condicional de A dado B, denotado por P(A / B), es la probabilidad de que suceda A dado que se sabe que el evento B ocurrió.

$$P(A / B) = \frac{P(A \cap B)}{P(B)}$$

P(A/B) = probabilidad de que ocurra el evento A, dado que el evento B ya ocurrió.
$P(A \cap B)$ = probabilidad de que ocurra A y B
P(B) = probabilidad de que ocurra el primer evento B.

Probabilidades y *N*-gramas (2)

Un *n-grama* es una subsecuencia de ***n*** elementos de una secuencia dada. En trando más en detalles, podemos definirlo como un modelo probabilístico que permite hacer una predicción estadística del próximo elemento de cierta secuencia de elementos sucedida hasta el momento, es decir, es capaz de predecir x_i basándose en x_{i-1}, . x_{i-2}, x_{i-3}, x_{i-n}.



Probabilidades y *N*-gramas (2)

Cardie (2006) explica que, básicamente, los n-gramas permiten hacer conteos de palabras bajo la hipótesis de que, si identificamos una palabra con la cual inicie una cadena sintáctica (frase, oración, párrafo, etc.), podemos inferir probabilísticamente cuál es la que le sigue, usando una fórmula como:

- $P(w_1, w_2, \ldots, w_{n-1}, w_n)$

$$P(w_1^n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_1^2)\ \ldots P(w_n|w_1^{n-1})$$

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

Probabilidades y *N*-gramas (3)

Volviendo con el ejemplo de Jurafsky y Martin (2000), y considerando entonces que, dependiendo de la palabra que hayamos identificado dentro de una cadena sintáctica, ya se puede proponer cierto grado de probabilidad:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Probabilidades y *N*-gramas (4)

En términos simples, lo que esperamos es que nuestros *n-gramas* se puedan representar del siguiente modo:

<s> I
I want
want to
to eat
eat Chinese
Chinese food
food </s>

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/introduccioacuten-a-la-linguumliacutestica-computacional.html