Lingüística computacional

Fundamentos estadísticos

Escuela Nacional de Antropología e Historia (ENAH)
Agosto – diciembre de 2015

Frecuencia de palabras

- ¿Cuáles son las palabras más frecuentes en un corpus?
- Palabras funcionales (function words)

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Table 1.1 Common words in ***Tom Sawyer.***

Frecuencia de palabras

- La mayoría de las palabras aparecen muy pocas veces
- De pocas palabras hay muchos ejemplos
- ¿Cómo estudiar elementos tan poco frecuentes?
- Hapax legomena = frecuencia 1

Word Frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
> 100	102

Table 1.2 Frequency of frequencies of word types in *Tom Sawyer.*

Frecuencia de palabras

Tamaño de un corpus

- Ocurrencias (tokens): cada una de las palabras de un corpus
- Tipos de palabra (types): cada palabra distinta de un corpus
- Razón tipos/tokens: diversidad léxica

TF-IDF

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t.$$

- Asigna un pesa a cada palabra (término) de un documento
- Se calcula en una colección de documentos
- Alta cuando la palabra ocurre en pocos documentos
- Baja cuando la palabra ocurre poco o en muchos documentos
- Cero si la palabra ocurre en todos los documentos

TF-IDF

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}.$$

	Doc1	Doc2	Doc3			
	TF			N	DF	IDF
car	27	4	24	3	3	0
auto	3	33	0	3	2	0.17609
insurance	0	33	29	3	2	0.17609
best	14	0	17	3	2	0.17609

	Doc1	Doc2	Doc3
	TF-IDF		
car	0	0	0
auto	0.52827	5.81101	0
insurance	0	5.81101	5.10665
best	2.46528	0	2.99355

Probabilidad

- La Teoría de la probabilidad trata de hacer predicciones de qué tan probable es que algo ocurra
- Las probabilidades son números entre 0 y 1
- 0 = imposibilidad
- 1 = certeza

Probabilidad

- $P(A) =$ probabilidad del evento *A*
- Si una moneda se lanza tres veces, ¿cuál es la probabilidad de que salgan dos soles?
- $\Omega = \{SSS, SSA, SAS, SAA, ASS, ASA, AAS, AAA\}$ = espacio de posibilidades
- Probabilidad de cada posibilidad = $\frac{1}{8}$
- $P(A) = \frac{|A|}{|\Omega|}$
- $|A| =$ número de elementos en el conjunto *A*

Probabilidad

- Si una moneda se lanza tres veces, ¿cuál es la probabilidad de que salgan dos soles?
- $\Omega = \{SSS, SSA, SAS, SAA, ASS, ASA, AAS, AAA\}$ = espacio de posibilidades
- $|A| = \{SSA, SAS, ASS\}$
- $P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8}$

Probabilidad

Table 5.2. *Numbers of monolingual or bilingual adults in two hypothetical populations cross-tabulated by sex*

	Male	Female	Total
Population A			
Bilingual	2 080	1 920	4 000
Monolingual	3 120	2 880	6 000
	5 200	4 800	10 000
Population B			
Bilingual	2 500	1 500	4 000
Monolingual	2 700	3 300	6 000
	5 200	4 800	10 000

- $P(male) = \frac{5{,}200}{10{,}000} = 0.52$

Population A:

- $P(male\ and\ bilingual) = \frac{2{,}080}{10{,}000} = 0.208$

Probabilidad

Table 5.2. *Numbers of monolingual or bilingual adults in two hypothetical populations cross-tabulated by sex*

	Male	Female	Total
Population A			
Bilingual	2 080	1 920	4 000
Monolingual	3 120	2 880	6 000
	5 200	4 800	10 000
Population B			
Bilingual	2 500	1 500	4 000
Monolingual	2 700	3 300	6 000
	5 200	4 800	10 000

- $P(male) = \frac{5{,}200}{10{,}000} = 0.52$

Population B:

- $P(male\ and\ bilingual) = \frac{2{,}500}{10{,}000} = 0.25$

Probabilidad condicional

- Probabilidad de un evento dado cierto conocimiento
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Probabilidad

Table 5.2. *Numbers of monolingual or bilingual adults in two hypothetical populations cross-tabulated by sex*

	Male	Female	Total
Population A			
Bilingual	2 080	1 920	4 000
Monolingual	3 120	2 880	6 000
	5 200	4 800	10 000
Population B			
Bilingual	2 500	1 500	4 000
Monolingual	2 700	3 300	6 000
	5 200	4 800	10 000

- $P(male) = \frac{5{,}200}{10{,}000} = 0.52$

Population B:

- $P(male\ and\ bilingual) = \frac{2{,}500}{10{,}000} = 0.25$
- $P(bilingual|male) = \frac{P(bilingual\ and\ male)}{P(male)} = \frac{0.25}{0.52} = 0.48$

Algunos conceptos fundamentales

- Tokenizar (tokenization)
- Lematizar (lemmatization)
- Truncar (stemming)
- Etiquetado de categorías gramaticales (Part of Speech Tagging, POST)
- Análisis sintáctico superficial (Shallow parsing o chunking)
- Análisis sintáctico (Parsing, full parsing)

Fin