Ciencia de Datos y Lenguaje Natural

Leyes empíricas sobre la frecuencia de las palabras

Grupo PLN - INCO

Universidad de la República

Leyes empíricas del lenguaje

- Son leyes deducidas de la observación en conjuntos de texto.
- Tienen que ver con las frecuencias, relacionando cantidad de tipos con cantidad de tokens.

Leyes empíricas del lenguaje

- La ley de Zipf propone una relación entre el rango de una palabra (posición en una lista por frecuencia descendente) y la frecuencia.
- La ley de Heap dice que en un texto a medida que aumenta la cantidad de tokens crece también la cantidad de tipos: siempre aparecen palabras nuevas.

Ley de Zipf

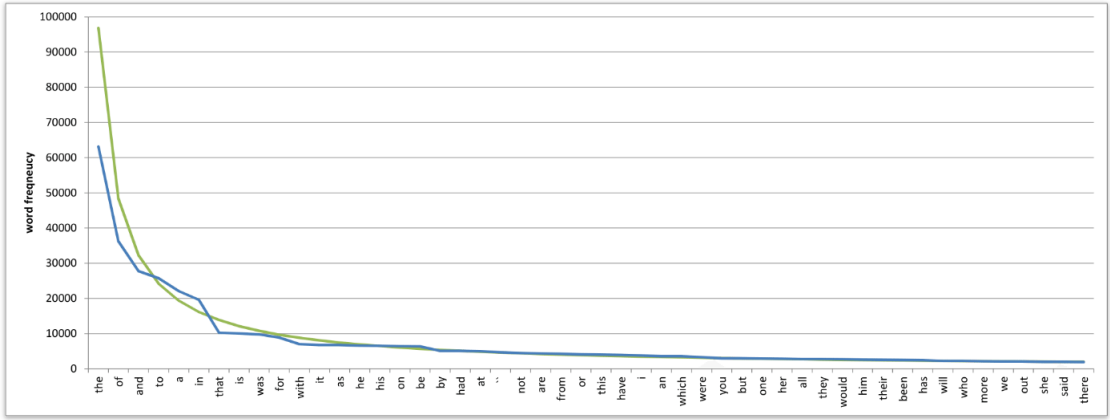
- La ley de Zipf relaciona la frecuencia F_n de una palabra en un texto con su rango n, o sea, la posición que ocupa en una lista por orden decreciente de frecuencia.

- $$f(r) = K\frac{1}{r^{\alpha}}, \alpha \approx 1$$

	frequency (per milln)	cumulative frequency	frequency rank	alphabet rank
the	68351.63	68351.63	1	318525
of	33008.66	101360.29	2	212425
and	28651.11	130011.40	3	11331
to	27599.22	157610.62	4	322312
a	23160.48	180771.10	5	1
in	20670.81	201441.91	6	149032
is	10571.15	212013.06	7	156934
that	10549.02	222562.08	8	318470
was	9939.26	232501.34	9	356587
it	9882.90	242384.23	10	157771
for	9309.44	251693.67	11	114281
on	7636.66	259330.33	12	213645
with	7171.07	266501.39	13	361235
he	7167.84	273669.23	14	134413
be	7153.17	280822.40	15	27945
I	7036.88	287859.28	16	146205
by	5866.89	293726.17	17	44040
as	5793.35	299519.52	18	19178
at	5154.12	304673.64	19	20631
you	5043.27	309716.91	20	364651

https://www.cs.cmu.edu/~cburch/words/top.html

Zipf sobre el Brown corpus, 1 millón de palabras



https://shadycharacters.co.uk/2015/10/zipfs-law/

Observaciones a relaciones de frecuencia

- Las palabras más frecuentes son palabras gramaticales.
- Son palabras muy cortas.
- La frecuencia de la palabra n-ésima multiplicada por su rango y la de la primera son aproximadamente iguales

Desviaciones a la ley de Zipf

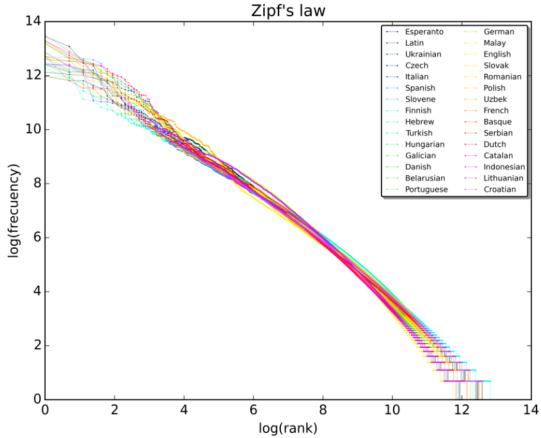
- Si bien la ley de Zipf se comprobó empíricamente, se encontró que sistemáticamente distorsiona en valores extremos (términos muy frecuentes, términos muy poco frecuentes).

$$f(r) = K\frac{1}{r^{\alpha}} \qquad log(f(r)) = log(K) - \alpha log(r)$$

- Si se grafica tomando logaritmos, queda la ecuación de una recta con pendiente negativa.

- En la gráfica siguiente se nota claramente la desviación en valores iniciales y finales.

Rango versus frecuencia para los 1eros 10 millones de tokens en 30 Wikipedias en escala logarítmica



Zipf para wikipedias en 30 lenguas, 2015

Ley de Zipf-Mandelbrot

- Mandelbrot (1953) presenta una modificación a la ley de Zipz, tratando de lograr mejor desempeño en valores de los extremos.
- Se agrega un parámetro, β, que suele tomar valores pequeños. Cuando es 0, equivale a Zipf.
- Se ha reportado que la ley de Zipf-Mandelbrot presenta un mejor ajuste a los datos lingüísticos particularmente en la región de palabras de rango bajo (aprox menor a 20)
-

$$f(r) = K\ 1/(r+\beta)^{\alpha} \qquad log(f(r)) = log(K) - \alpha log(r+\beta)$$

Otras manifestaciones de Zipf

- Las palabras más frecuentes tienen muchos significados.
- Hay una tendencia a que las palabras más frecuentes sean más cortas.

Intentos de explicaciones

- Economía de esfuerzos. Tanto el hablante como el interlocutor tratan de minimizar el esfuerzo que realizan
- El hablante tiende a utilizar un repertorio pequeño de palabras
- El interlocutor a manejar un repertorio más amplio de palabras menos comunes (de modo de reducir ambigüedad).
- El compromiso de economía máxima daría la relación entre frecuencia y rango que se advierte en Zipf.

Ley de Heap

- Ley empírica que describe la cantidad de palabras distintas en un corpus en función de la cantidad de tokens.

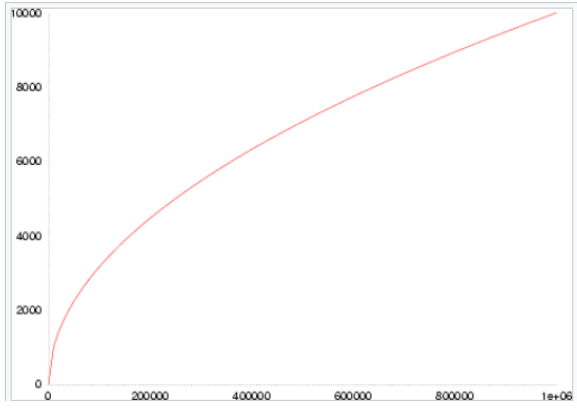
$$V_R(n) = Kn^\beta V_R(n) = Kn^\beta$$

V_R – cantidad de palabras distintas en una porción de texto desde el inicio hasta el n-ésimo token.
K y β son parámetros, β cercano a 0.5

- Implica que, con colecciones suficientemente grandes, sigue aumentando el tamaño del vocabulario. O sea, el vocabulario no es finito.

- Importa aclarar que se incluyen en el vocabulario nombres propios de diversa índole.

Ley de Heap



Las x representan la cantidad de tokens del texto desde el inicio y las y el tamaño del vocabulario.

Wikipedia, ley de Heap

Frecuencias de frecuencias

Frequencies of frequencies in *Tom Sawyer*

Word Frequency	Frequency of Frequency		
1	3993	71,730	word tokens
2	1292	8,018	word types
3	664		
4	410		
5	243		
6	199		
7	172		
8	131		
9	82		
10	91		
11–50	540		
51–100	99		
> 100	102		

Consecuencias de las leyes de Zipf y Heap para el Procesamiento Automático del Lenguaje

- Casi el 50 % de las palabras de un corpus ocurren solo una vez. Esto implica que un modelo estadístico se tendrá que enfrentar al problema de datos dispersos sin instancias suficientes como para estimar probabilidades.

- Gran parte de las ocurrencias de las palabras corresponde a palabras con varias acepciones. Esto plantea el problema de la ambigüedad léxica.