Text Mining


Universidad
de Valparaíso
CHILE


ESCUELA DE INGENIERÍA
INFORMÁTICA
UNIVERSIDAD DE VALPARAÍSO

Unidad 1 - Fundamentos de Text Mining

➔ Contenidos

- Conceptos básicos
- **Leyes del texto**
- Procesamiento de texto
- Problemas en Text Mining

Unidad 1 - Fundamentos de Text Mining

Leyes del Texto

- Ley de zipf
- Ley de heaps

Material basado en: https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Unidad 1 - Fundamentos de Text Mining

Ley de Zipf

- 1940 por el lingüista George Kingsley Zipf
- Como términos están distribuidos en el documento.
- Si t_1 es el término más común, t_2 es el segundo más común, ..., etc. Luego la frecuencia de la colección cf_i del i-ésimo término más común es proporcional a ***1/i*** :

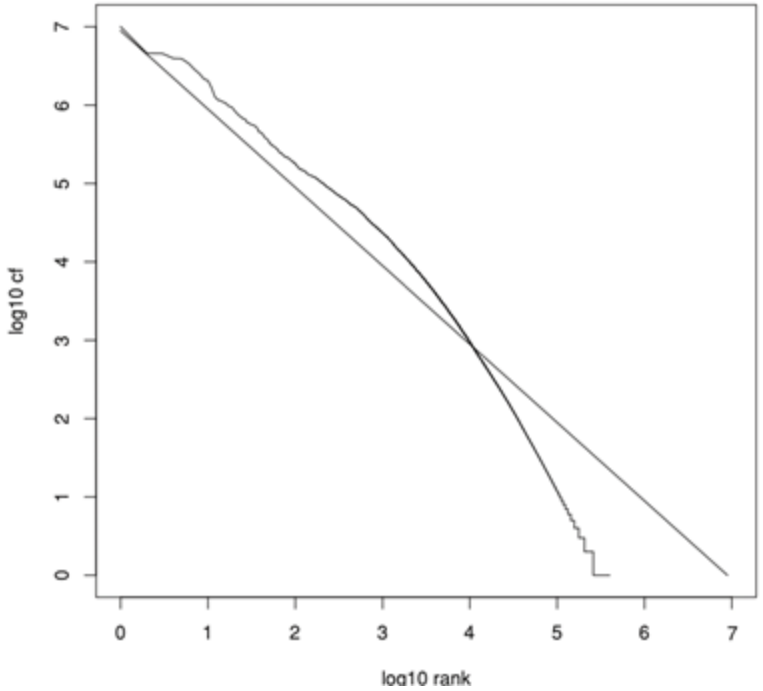
$$cf_i \propto 1/i$$

Si el término más frecuente ocurre cf1 veces, luego el segundo más frecuente ocurre la mitad de veces, el tercer término un tercio, y así... la intuición es que la frecuencia disminuye muy rápidamente.

- De forma equivalente, la ley Zipf se puede escribir como: $cf_i = ci^k$
- O como: $\log cf_i = \log c + k \log i$ donde $k = -1$ y c corresponde a un valor constante. (Power law)

Unidad 1 - Fundamentos de Text Mining

Ley de Zipf



Ley de Zipf para data Reuters-RCV1

Ranking de frecuencia para los términos de la colección.

Imagen: https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Unidad 1 - Fundamentos de Text Mining

Ley de Heaps

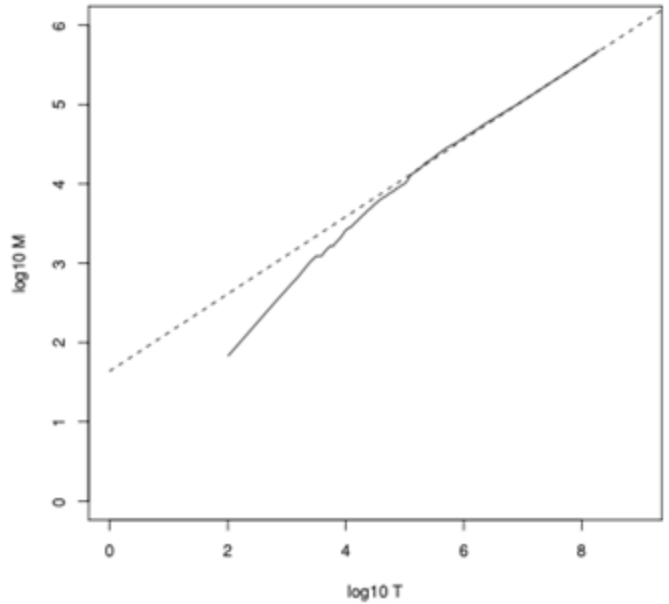
- Permite estimar el tamaño del vocabulario (M) en función del tamaño de la colección.

$$M = k * T^b$$

  - T: número de tokens de la colección.
  - Con k, 30 <= k <=100
  - b ≃ 0,5
- Idea general: existe una relación sencilla entre el tamaño de la colección y el tamaño del vocabulario en el espacio log-log, y la suposición de linealidad suele cumplirse en la práctica.

Unidad 1 - Fundamentos de Text Mining

Ley de Heaps



Ejemplo para dataset Reuters-RCV1.

- Si $T > 10^5 = 100000$, $b = 0{,}49$, $k = 44$.

  Para los 1000020 tokens, se tiene

- $44 \times 1000020^{0.49} \simeq 38{,}323$
- 38,323 términos, y el número real corresponde a 38,365.