RI- Chapter 04 – Processing Text — Notes

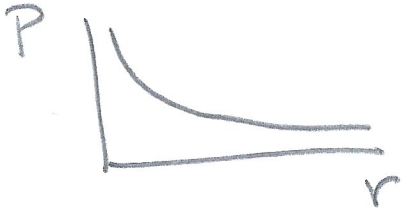
Ley Zipf

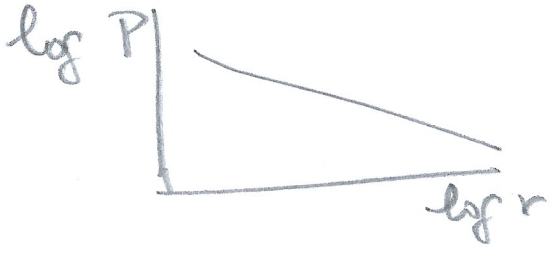
f: frecuencia = nº ocurrencia de la palabra

P: probabilidad = $\frac{1}{\text{nº ocurrencias totales en la colección}}$

$f = \frac{n}{r}$, equiv. $P = \frac{c}{r}$ | $P(w) = \frac{C}{r(w)}$

$\log P(w) = \log(C) - \log(r(w))$





$C \geq$ 0'1 Zipf's

APF9 asistant → $P = \frac{5.095}{39.749.175} = 0'00013$ → 0'013 %

$r \cdot f = k$

rango · frecuencia $= k$

$r_n \cdot n = k$ $\qquad r_n = k/n$

$r_{n+1} \cdot (n+1) = k$ $\qquad r_{n+1} = k/(n+1)$

ranking ultima palabra $\cdot\ 1 = k$

k = ranking ultima palabra = número total de palabras

Número de palabras con frecuencia n:

$$r_n - r_{n+1} = \frac{k}{k} - \frac{k}{n(n+1)} = \frac{k}{n(n+1)}$$

$$\frac{\text{Num. palabras frecuencia } n}{\text{Num. total de palabras}} = \frac{1}{k}\,\frac{k}{n(n+1)} = \frac{1}{n(n+1)}$$

En Ejemplo: AP89

$n = 5099$ veces $\qquad r_n = 1006$

$n+1 = 5100$ " $\qquad r_{n+1} = 1002$

$r_n - r_{n+1} = 1006 - 1002 = 4$ = nº palabras con frecuencia 5099

Por la predicción

$$\frac{1}{5099 \times 5100} \times 39.799.179 = 1'53$$

39.799.179 → palabras

k = Número total de palabras en AP89

Ley Heap

$$V = k \cdot n^{\beta}$$

APB9 $k = 62.95$
$\beta = 0.455$

GOV2 $k = 7.34$
$\beta = 0.648$

Estimating Result Set Size

f_a = nº documentos en los que ocurre el término a

Prob. ocurrencia término a ← $P(a) = \frac{f_a}{N}$ → Suponemos independencia términos

$$f_{abc} = P(a \cap b \cap c) \cdot N = P(a) \cdot P(b) \cdot P(c) \cdot N =$$

$$= \frac{f_a}{N} \cdot \frac{f_b}{N} \cdot \frac{f_c}{N} \cdot N = \frac{f_a \cdot f_b \cdot f_c}{N^2}$$

$$P(c \mid (a \cap b)) = \frac{P(c \cap a \cap b)}{P(a \cap b)}$$

$$P(a \cap b \cap c) = P(a \cap b) \cdot P(c \mid (a \cap b))$$

Esto lo aproxima por: $P(c \mid a)$ o $P(c \mid b)$

a = tropical
c = fish
b = aquarium

$$P(a \cap c \cap b) = P(a \cap b) \cdot P(c \mid b)$$

$$\frac{P(c \cap b)}{P(b)}$$

$$\frac{f_1}{N} = \frac{f_2}{N} \cdot \frac{f_3/N}{f_4/N}$$

$$P(1) = P(2) \cdot \frac{P(3)}{P(4)} \Rightarrow f_1 = f_2 \cdot \frac{f_3}{f_4}$$

Result set Elmotore

26480 docs provenids contens 'aquarium'

De 3000 provenids de 30.26480) 258 contenen 'tropical' i/o 'aquarium'

3000 → 258

26480 → x

$$x = \frac{258 \cdot 26480}{3000} = \frac{c}{s} = f_{\text{tropical i aquarium}}$$

c num de documents que contenen les 3 paraules

$s = \frac{3000}{264800}$ → proporció del total de documents que tenen realment

Estimator collection size

Suposición de palabras independientes

$$P(a \cap b) = P(a) \cdot P(b)$$

$$\frac{f_{ab}}{N} = \frac{f_a}{N} \cdot \frac{f_b}{N}$$

$$N = \frac{f_a \cdot f_b}{f_{ab}}$$

Estimar collection size Google, Bing

lincoln → f_a

tropical → f_b

Búsqueda avanzada, f_{ab} lincoln AND tropical

$$N = \frac{f_a \cdot f_b}{f_{ab}}$$

Stemming

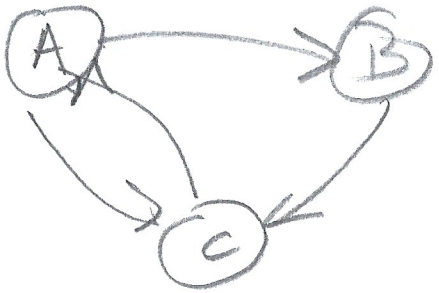
Supply falso negativo: el stemmer no lo detecta, i.e., dice que no es stem y realmente supply es la raiz

up falso positivo: el stemmer sí lo detecta pero no es la raiz de UPS

Falsos positivos: ej: organization/organ, el stemmer detecta el stem organ que no es la raiz de organization

Falsos negativos: ej: matrices/metrix, el stemmer produce distintos stems pero realmente tienen la misma raiz.

Page Rank



$$PR(C) = \frac{PR(A)}{2} + \frac{PR(B)}{1}$$

$$PR(A) = PR(C)$$

$$PR(B) = \frac{PR(A)}{2}$$

$$\left(PR(A)^0 \; PR(B)^0 \; PR(C)^0\right) \cdot MTP = \left(PR(A)^1 \; PR(B)^1 \; PR(C)^1\right)$$

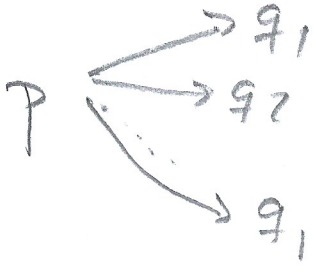
$$\begin{pmatrix} 0.33 & 0.33 & 0.33 \end{pmatrix} \cdot \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0.33 & 0.17 & 0.5 \end{pmatrix}$$

...

converge a $\begin{pmatrix} 0.4 & 0.2 & 0.4 \end{pmatrix}$

PageRank Pseudocódigo

l.11: $\Pi_i \leftarrow \frac{\lambda}{|P|}$ probabilidad de cada página por teleporting



l.17 $R_q = R_q + (1-\lambda)\frac{I_p}{|Q|}$ ⇝ propaga el PR de p, I_p, a las págs q a las que apunta p de forma equiprobable y en la cantidad total ponderada por $(1-\lambda)$

l.21. Si p no tiene páginas salientes, $|Q|=0$, su PageRank se propaga de forma equiprobable a todas las páginas del grafo.