

# Statistiques sur les termes

Soit un document  $d$  :

- constitué de  $K$  mots  $d[1], \dots, d[i], \dots$
- appartenant au vocabulaire  $V = \{t_1, \dots, t_m\}$  constitué de  $m$  termes.

**Fréquence d'un terme  $t$  :**

Soit  $t \in V$  un terme de vocabulaire. On note  $P(X=t)$  la fréquence d'apparition de ce terme dans le langage  $\mathcal{L}$  considéré, soit  $\sim: P(X=t) = \frac{|\{\omega \in \Omega : X=t\}|}{|\Omega|}$  où  $\Omega$  représente l'ensemble des productions de termes.

On a par définition  $\sim: \sum_{t \in A} P(X=t) = 1$

La fréquence empirique du symbole  $t$  dans le document  $d$  est donnée par  $\sim$ :



$$f_d(t) = \frac{|\{i:d[i]=t\}|}{|d|}$$

où  $|d|$  est le nombre de mots dans le document.

## Corpus de documents

Soit  $B$  un corpus de documents, constitué de  $n$  documents.



La fréquence empirique du terme  $t$  dans le corpus  $B$  est donnée par  $\sim: f_B(t) = \frac{|\{(i,j):d_i \in B, d_i[j]=t\}|}{|B|}$  où  $|B|$  est le nombre total de mots dans le corpus.

**Fréquence locale :**

Le fréquence empirique locale  $f_B(t,d)$  est donnée par :  $f_B(t,d) = p(X=t|Y=d) = p(t|d)$   
 $f_B(t,d) = \frac{|\{j:d \in B, d[j]=t\}|}{|d|}$  où  $|d|$  est le nombre de mots dans le document  $d$ .

**Fréquence documentaire**

On appelle **fréquence documentaire**  $g(t)$  d'un terme  $t$  la fréquence d'apparition du terme dans les différents documents de la base :

$$g(t) = p(t \in d)$$

Fréquence documentaire empirique :

$g(t) = \frac{|\{d:t \in d\}|}{|B|}$  avec:

- $n = |B|$  : nombre de documents
- $|\{d:t \in d\}|$  : nombre de documents contenant  $t$

### Information documentaire

$I(t) = -\log_2 g(t)$

- $I(t) = 0 \Rightarrow$  aucune information documentaire.

Ainsi, les termes apportant  $I$  bits d'information permettent de réaliser  $I$  partitions de la base (pour extraire des sous-ensembles de taille  $|B| / 2^I$ )

On remarque que :

- si le terme est présent dans tous les documents, son information documentaire est nulle.
- si le terme est présent dans un seul document, son information documentaire est maximale

On peut de même calculer l'**entropie (documentaire) croisée** de la base comme  $E(I(t))$  :  $H(B) = -E(\log_2 p(t \in d)) = -\sum_{t \in V} p(t) \log_2 p(t \in d)$  où  $p(t)$  représente la probabilité d'apparition du terme  $t$  sur tous les documents de la base.

On note  $h(t)$  la **contribution documentaire** du terme  $t$  :  $h(t) = -p(t) \log_2 p(t \in d)$

---

On calcule de même l'**entropie conditionnelle** d'un document  $d$  comme  $E(I(t) | d)$  :  $H(d) = -E(\log_2 p(t \in d) | d) = -\sum_{t \in V} p(t|d) \log_2 p(t \in d) = -\sum_{t \in d} p(t|d) \log_2 p(t \in d)$

On note  $h(t|d)$  la **contribution documentaire conditionnelle** du terme  $t$  dans le document  $d$  :  $h(t | d) = -p(t|d) \log_2 p(t \in d)$

Cette contribution est également appelée : **TF-IDF** ("Term frequency - Inverse document frequency")

From:  
<https://wiki.centrale-med.fr/informatique/> - **WiKi informatique**

Permanent link:  
[https://wiki.centrale-med.fr/informatique/public:algo-txt:statistiques\\_sur\\_les\\_termes](https://wiki.centrale-med.fr/informatique/public:algo-txt:statistiques_sur_les_termes)

Last update: **2020/04/20 17:27**

