Théorie de l'information et mesures d'entropie

Damien Nouvel

Plan

Mesures sur des corpus

- **Taille** pour stocker un corpus
  - Nombre de fichiers (80jours : 1)
  - Nombre de documents (80jours : 1)
  - Nombre de mots (80jours : 85K)
  - Espace disque requis (80jours : 776Ko)

  $\Rightarrow$ Quelles mesures pour l' « information » ?
- **Information** contenue dans un corpus
  - Compression de fichier (80jours : zip 192 Ko, bz2 117 Ko...)
  - Nombre de mots distincts (80jours : 9412)
  - ... ?

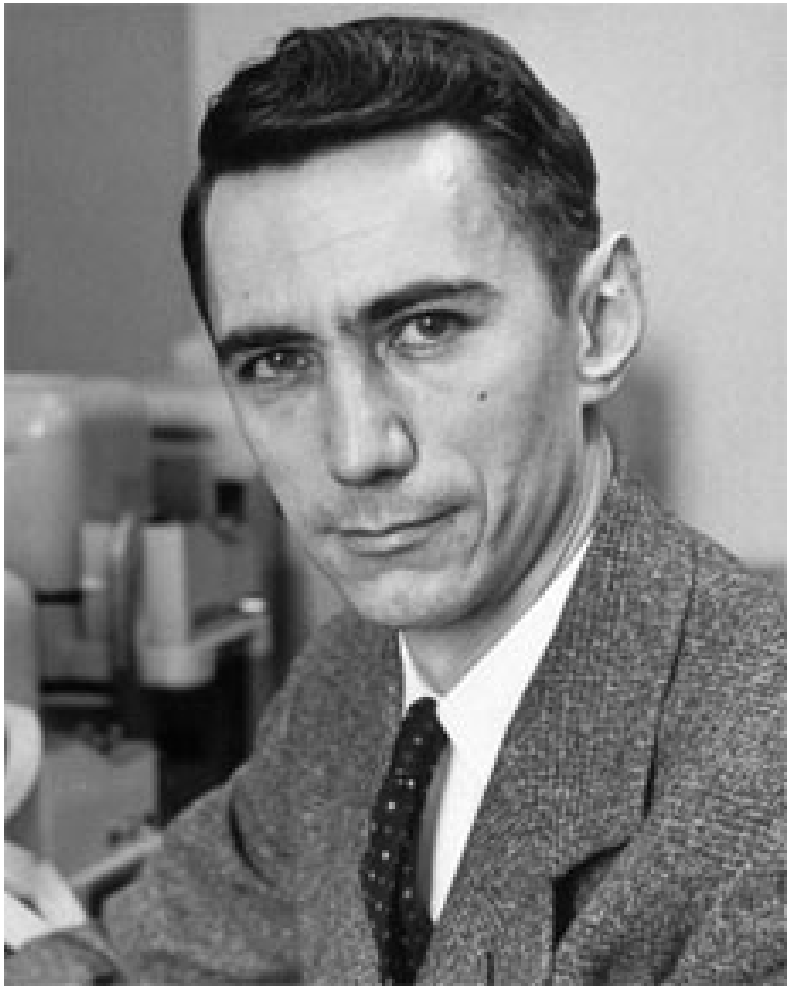
$\Rightarrow$ Nombreuses **mesures** pour **quantifier** un corpus

- Lien entre taille et **information**
  - Comment stocker un document de manière optimale ?
  - Combien de temps pour **lire et comprendre** un texte ?

  $\Rightarrow$ Compromis entre **stockage** et **accessibilité**

Plan

Théorie de l'information de Shannon

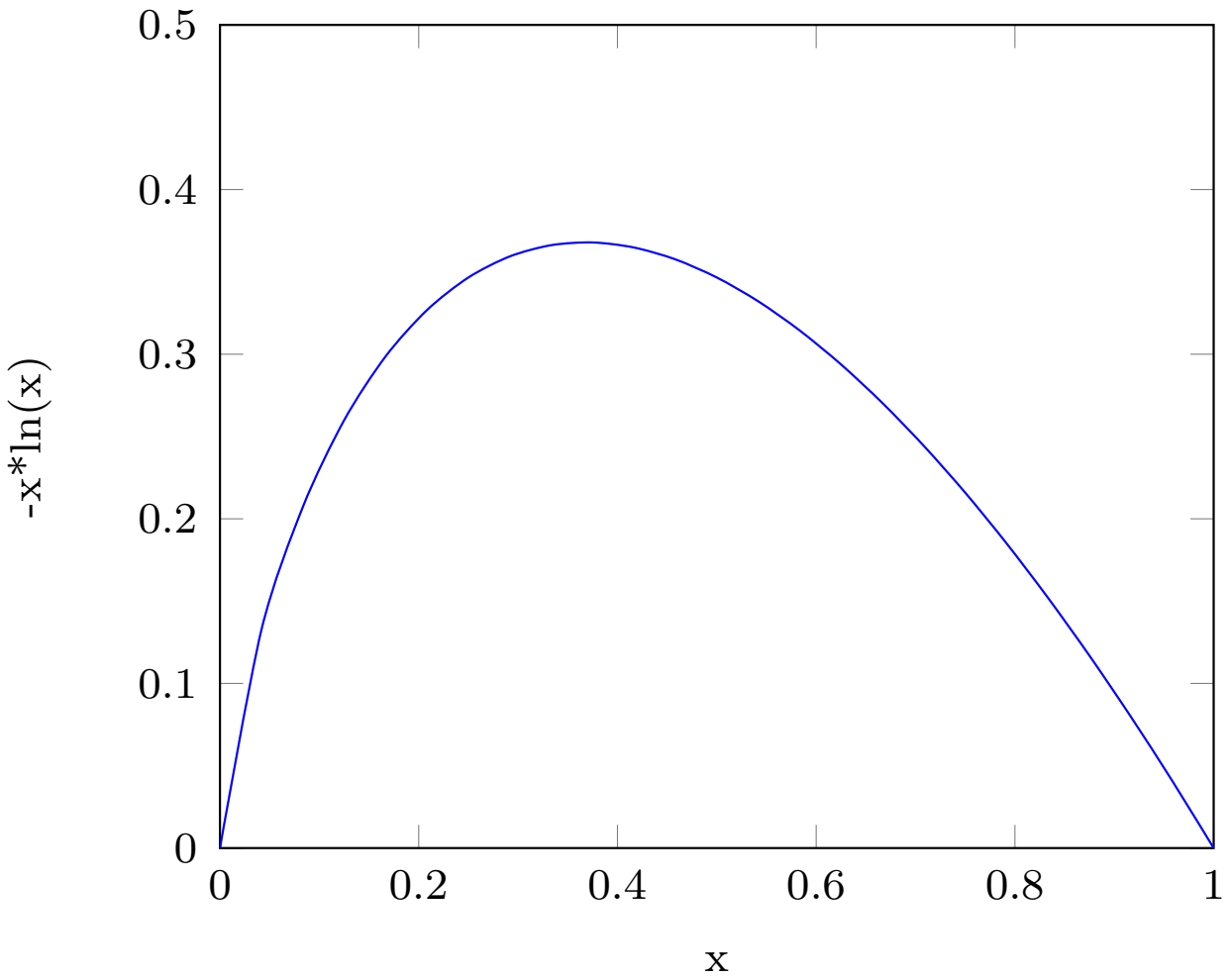


- **Claude Shannon** : **entropie**, th. de l'information (1948)

Entropie de Shannon

- Mesure **thermodynamique** adaptée aux **télécoms**
- Répandue en sciences (néguentropie)
- **Définition**
  - Formule : $H(X) = -\sum_{x \in X} P(X = x) * log_2(P(X = x))$
- **Propriétés**
  - **Positive** : $H(X) \geqslant 0$
  - Entropie **jointe** : $H(X, Y) \leqslant H(X) + H(Y)$
  - Entropie **conditionnelle** : $H(X, Y) = H(X) + H(Y|X)$
- **Comportement**
  - **Augmente** avec le **nombre** d'évènements équiprobables
    - Deux évènements $(P(X = i) = 0.5)$: $H(X) = 1$
    - Quatre évènements $(P(X = i) = 0.25)$: $H(X) = 2$
  - **Augmente** avec l'**équilibre** des probabilités
    - Déséquilibre $(P(X = 1) = 0.1, P(X = 2) = 0.9)$: $H(X) = 0.47$
    - Équilibrée $(P(X = 1) = 0.4, P(X = 2) = 0.6)$: $H(X) = 0.97$

Fonction d'entropie



Interprétation de l'entropie

⇒ L'entropie comme **mesure** de...
- Incertitude...
- Indécidabilité ?!
- Désorganisation #§ !/
- Chaos :s
- Information ?

⇒ difficile à interpéter...

- Intérêt de l'entropie
  - Mesure la **quantité** d'information
    - Un signal peu informatif est **redondant**
    - Un signal informatif est très **diversifié** et **peu prédictible**

    ⇒ En télécommunications : bande passante est nécessaire ?

  ⇒ Relation entre **données** et **modèle statistique**
- Se mesure en nombre de **bits** (logarithme base 2)

Calcul de l'entropie en python

⇒ Utilisation de la fonction `log` de la librairie `math`

```
import math
probas = [0.2, 0.3, 0.5]
entropie = 0
for proba in probas:
  entropie -= proba*math.log(proba, 2)
print('Entropie:', entropie)
```

⇒ Utilisation de la fonction `entropy` de la librairie `scipy.stats`

```
from scipy import stats
vals = [2, 3, 5]
print('Entropie:', stats.entropy(vals))
```

Information mutuelle

$\Rightarrow$ Mesure de la **corrélation** entre deux variables

- **Formule**

$$I(X, Y) = \sum_{x \in X, y \in Y} P(X = x, Y = y) * log_2 \left(\frac{P(X = x, Y = y)}{P(X = x) * P(Y = y)} \right)$$

- **Propriétés**
  - **Positive** : $I(X, Y) \geqslant 0$
  - En cas d'indépendance : $I(X, Y) = 0$
  - Lien / entropie : $H(X, Y) = H(X) + H(Y) + I(X, Y)$
  - Lien / entropie conditionnelle : $I(X, Y) = H(X) - H(X|Y)$

Divergence de Kullback-Leibler

⇒ Mesure la perte d'information par **approximation** d'une loi

- **Formule**

$$D_{KL}(P||Q) = \sum_{x \in X} P(X = x) * log_2 \left(\frac{P(X = x)}{Q(X = x)} \right)$$

- **Propriétés**
  - **Positive** : $D_{KL}(P, Q) \geqslant 0$
  - Les lois ne divergent pas si $D_{KL}(P||Q) = 0$
  - Comparaison sur les **mêmes données**

⇒ Aussi : **gain d'information**, **entropie relative**

Plan

Critères sur des données

- Tâche de **classification**
  - Recueil et examen des **données**
  - Recherche de **critères « utiles »**
  - Focalisation sur les **sous-ensembles** de données

⇒ Quelle importance accorder à chaque **critère**

⇒ Prise de décision

jour	*température*	*pluie*	*travail*	**sortir**
lundi	27	non	oui	oui
jeudi	12	oui	non	non
samedi	10	oui	oui	oui
mercredi	23	non	oui	non
lundi	27	oui	non	oui
mercredi	15	oui	non	oui

Critères sur des données

⇒ **L'arbre de décision** évalue les critères pour **classifier**

- Structure de l'arbre
  - Les nœuds contiennent les variables
  - Les arcs contiennent une décision sur les valeurs
  - Les feuilles contiennent les données
- Évaluation de l'apport d'une décision par **entropie**
  - Pour chaque **feuille**, pour chaque **critère** différence entre
    - Entropie du nœud n
      $-\sum_{x \in X} P(X = x|n) * log_2(P(X = x|n))$
    - Somme pondérée des entropie des nœuds enfants $e \in child(n)$
      $-\sum_{e \in enfant(n)} \frac{|e|}{|n|} \sum_{x \in X} P(X = x|e) * log_2(P(X = x|e))$

  ⇒ Choix du critère qui **diminue le plus l'entropie**

⇒ Séquence de **décisions** guidées par l'**entropie**

⇒ Possibilité de visualiser les décisions sous forme d'**arbre**

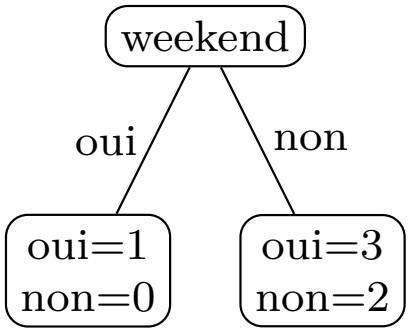
Exemple

weekend	*temp.*	*pluie*	*travail*	**sortir**
non	chaud	non	oui	oui
non	froid	oui	non	non
oui	froid	oui	oui	oui
non	chaud	non	oui	non
non	chaud	oui	non	oui
non	doux	oui	non	oui

$$
\begin{aligned}
&H(sortir) \\
&= -4/6 * log(4/6) - 2/6 * log(2/6) \\
&= 0.92
\end{aligned}
$$

Exemple

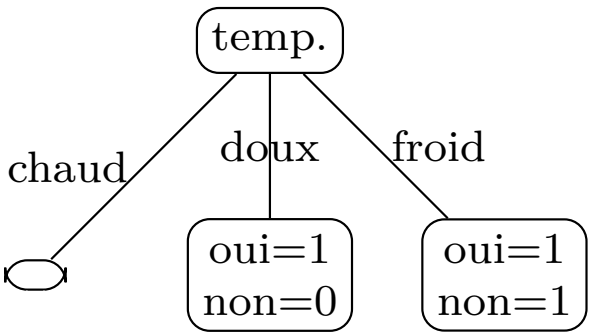
weekend	*temp.*	*pluie*	*travail*	**sortir**
non	chaud	non	oui	oui
non	froid	oui	non	non
oui	froid	oui	oui	oui
non	chaud	non	oui	non
non	chaud	oui	non	oui
non	doux	oui	non	oui



$$
\begin{aligned}
&H(sortir) \\
&= \ 1/6 * (-1 * log(1)) \\
&+ 5/6 * (-3/5 * log(3/5) - 2/5 * log(2/5)) \\
&= 0.81
\end{aligned}
$$

Exemple

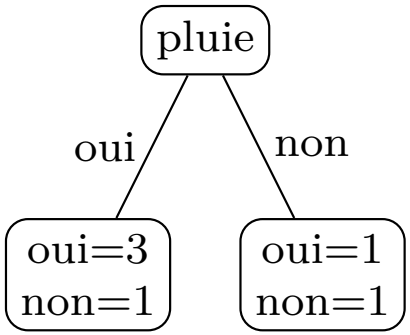
weekend	*temp.*	*pluie*	*travail*	**sortir**
non	chaud	non	oui	oui
non	froid	oui	non	non
oui	froid	oui	oui	oui
non	chaud	non	oui	non
non	chaud	oui	non	oui
non	doux	oui	non	oui



$$
\begin{aligned}
&H(sortir) \\
&= \quad 3/6 * (-2/3 * log(2/3) - 1/3 * log(1/3)) \\
&+ 1/6 * (-1 * log(1)) \\
&+ 2/6 * (-1/2 * log(1/2) - 1/2 * log(1/2)) \\
&= 0.79
\end{aligned}
$$

Exemple

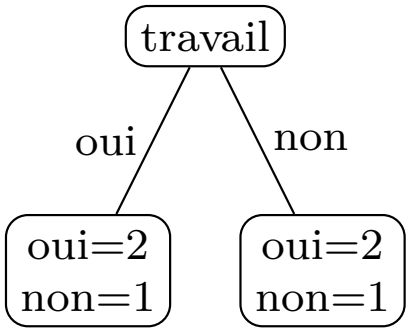
weekend	*temp.*	*pluie*	*travail*	**sortir**
non	chaud	non	oui	oui
non	froid	oui	non	non
oui	froid	oui	oui	oui
non	chaud	non	oui	non
non	chaud	oui	non	oui
non	doux	oui	non	oui



$$
\begin{aligned}
&H(sortir) \\
&= \quad 4/6 * (-3/4 * log(3/4) - 1/4 * log(1/4)) \\
&+ 2/6 * (-1/2 * log(1/2) - 1/2 * log(1/2)) \\
&= 0.87
\end{aligned}
$$

Exemple

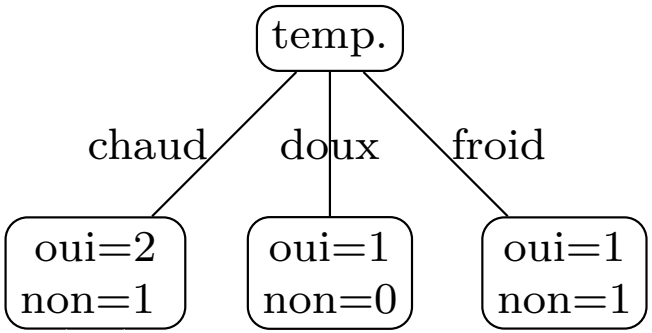
weekend	*temp.*	*pluie*	*travail*	**sortir**
non	chaud	non	oui	oui
non	froid	oui	non	non
oui	froid	oui	oui	oui
non	chaud	non	oui	non
non	chaud	oui	non	oui
non	doux	oui	non	oui



$$
\begin{aligned}
&H(sortir) \\
&= \quad 3/6 * (-2/3 * log(2/3) - 1/3 * log(1/3)) \\
&+ 3/6 * (-2/3 * log(2/3) - 1/3 * log(1/3)) \\
&= 0.92
\end{aligned}
$$

Exemple

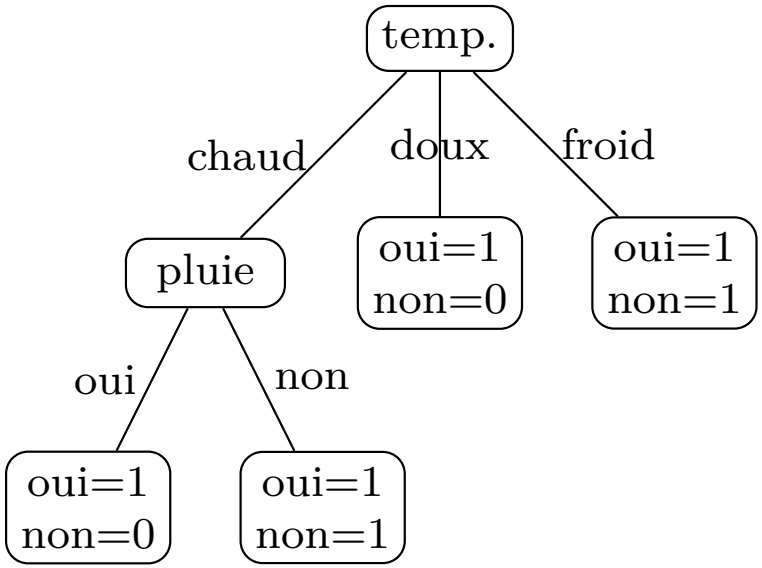
weekend	*temp.*	*pluie*	*travail*	**sortir**
non	chaud	non	oui	oui
non	froid	oui	non	non
oui	froid	oui	oui	oui
non	chaud	non	oui	non
non	chaud	oui	non	oui
non	doux	oui	non	oui



$$H(sortir|temp = chaud)$$
$$= -2/3 * log(2/3) - 1/3 * log(1/3)$$
$$= 0.92$$

Exemple

weekend	*temp.*	*pluie*	*travail*	**sortir**
non	chaud	non	oui	oui
non	froid	oui	non	non
oui	froid	oui	oui	oui
non	chaud	non	oui	non
non	chaud	oui	non	oui
non	doux	oui	non	oui



$$H(sortir|temp = chaud)$$
$$= 1/3 * (-1 * log(1))$$
$$+ 2/3 * (-1/2 * log(1/2) - 1/2 * log(1/2))$$
$$= 0.67$$

$$H(sortir)$$
$$= 3/6 * H(sortir|temp = chaud)$$
$$+ 1/6 * (-1 * log(1))$$
$$+ 2/6 * (-1/2 * log(1/2) - 1/2 * log(1/2))$$
$$= 0.66$$

Exercice

- Réalisez un arbre de décision sur le tableau suivant décrivant les mots présents dans des textes et leurs catégories associées :

ballon	*président*	*euro*	*équipe*	**catégorie**
oui	non	non	oui	sport
oui	non	oui	oui	sport
non	oui	oui	non	sport
non	oui	non	non	pol
non	oui	oui	oui	pol
non	non	non	non	eco
oui	non	oui	non	eco
non	oui	oui	non	eco