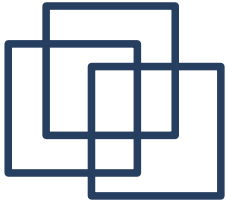


# Module AD

---

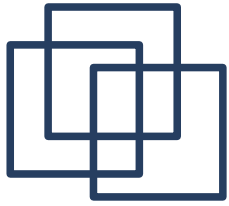
- ✓ Introduction à R
- ✓ Introduction aux méthodes de classification (supervisée)
- ➔ **Introduction aux méthodes de Fouille de Textes**



---

# AIDE A LA DECISION

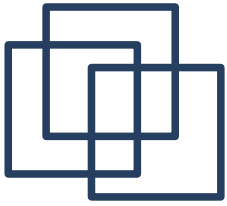
## Fouille de Textes



# Plan du cours FDT

---

- Introduction aux problématiques de la fouille de textes (pourquoi, quoi, comment)
- Les différentes étapes d'un processus de fouille de textes
- Représentation vectorielle des textes pour la classification
- Introduction au package R « tm » (text mining)



# Introduction à la FDT

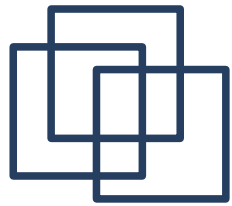
---

## État des lieux

- Entre 3 et 5 exabytes ( $10^9$  Go) de données originales produites chaque année,
- Du texte, de l'audio, de l'image, de la vidéo, ...
- Les 17 millions de livres de la bibliothèque du Congrès américain représenteraient 136 terabytes ( $10^3$  Go) soit 37000 fois moins.

## Accès et exploitation

- L'accès à l'information, à des données pertinentes, à la connaissance, devient un véritable enjeu.
- Rendre les masses de données utilisables.



# FDT vs. Interrogation de BDD

---

## Données non ou peu structurées

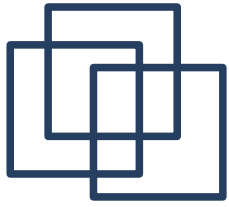
- Par opposition aux bases de données où les données sont structurées et stockées dans des tables avec des champs particuliers
- A priori, pas de travail préalable de réflexion, de structuration et de représentation des données (indexation)

## Recherche d'informations implicites

- La connaissance/information est souvent déduite d'autres données (type de média, contexte historique, etc.)

Exemple Qui a été reçu par N. Sarkozy le 10 décembre 2007?

*...le dirigeant Lybien a été accueilli à l'Elysée avec tous les honneurs ...*



# Définitions

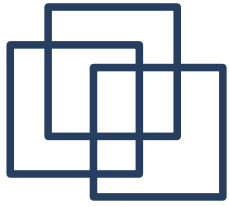
---

## Fouille de textes

- Acquérir des connaissances, des informations (plus généralement des données) enfouies dans des corpus de textes

## Corpus

- Recueil de documents concernant une même discipline (dictionnaire, encyclopédie, archives journalistiques, etc.)
- Ensemble « cohérent » de textes, d'objets
- Exemple de corpus de textes géant : le Web.



# Tâches de FDT

---

## Mesure des audiences Web

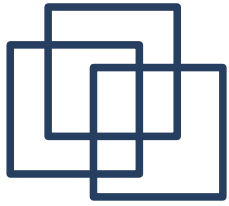
Enregistrer les sites visités par un panel d'internautes et mettre en relation les parcours des internautes avec leur description sociologique.

- Corpus : pages Web visitées
- Objectif : cibler la clientèle, chercher des traits caractéristiques

## Systèmes de question/réponse

Répondre à une question précise à partir de textes électroniques ou du Web.

- Corpus : textes électroniques ou Web
- Exemple d'objectif : dans quelle ville se situe la Tour Eiffel?



# Tâches de FDT

---

## Réseaux sociaux

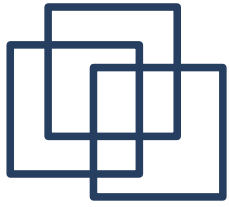
Étudier les co-occurrences (personnes, entreprises) afin d'en extraire des relations

- Corpus : sites Web ciblés d'actualité
- Objectif : qui connaît qui?

## Extraction d'information

- Classification thématique de documents (mails, dépêches d'actualité)
- Détection d'évènements et de nouveauté (forums de discussion)
- Extraction de faits : changement de dirigeant, fusion d'entreprises, opérations boursières, etc.





# Tâches de FDT

---

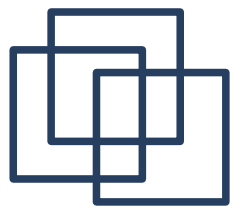
## Fouille d'opinions

Repérer/distinguer les avis positifs et négatifs sur un fait


- Corpus : corpus d'avis (réactions d'articles, enquêtes de satisfaction)
- Objectif : recommandation, sondages

## Recherche d'information

- Classification thématique de documents (pages web de réponse)
- Enrichissement de requêtes
- Résumé automatique de textes



# Tâches de FDT



web news images wikipedia blogs jobs more »

apple

Search

[advanced preferences](#)

clusters sources sites

All Results (250) remix

+ Store (30)

+ Download (18)

+ Reviews (17)

+ Photography (18)

+ Developer, Connection (11)

+ Tablet (14)

+ Mac OS X (11)

• Program (9)

+ Features (11)













• History (9)

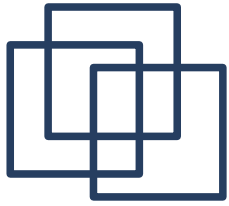
+ Fruit (5)

• Phones (7)

Top 249 results of at least 139,000,000 retrieved for the query **apple** ([definition](#)) ([details](#))

[Site Apple ® Officiel](#) - Découvrez les nouveautés **Apple** et le MacBook Air. Livraison gra

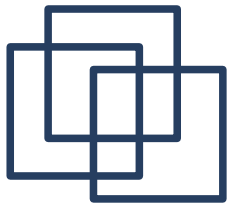
- [Apple](#)     
**Apple** designs and creates iPod and iTunes, Mac laptop and desktop computers, the OS X  
[www.apple.com](#) - [cache] - Ask, Open Directory, Yahoo!
- [The Apple Store](#)     
Buy direct from the **Apple** Store. Online ordering with custom configurations and special de  
[store.apple.com](#) - [cache] - Yahoo!, Ask, Open Directory
- [Apple - Wikipedia, the free encyclopedia](#)     
The **apple** is the pomaceous fruit of the **apple** tree, species *Malus domestica* in the rose fa  
[en.wikipedia.org/wiki/Apple](#) - [cache] - Bing, Ask, Yahoo!
- [Apple Developer Connection](#)     
Provides news and technical information for **Apple** Developers.  
[developer.apple.com](#) - [cache] - Open Directory, Ask



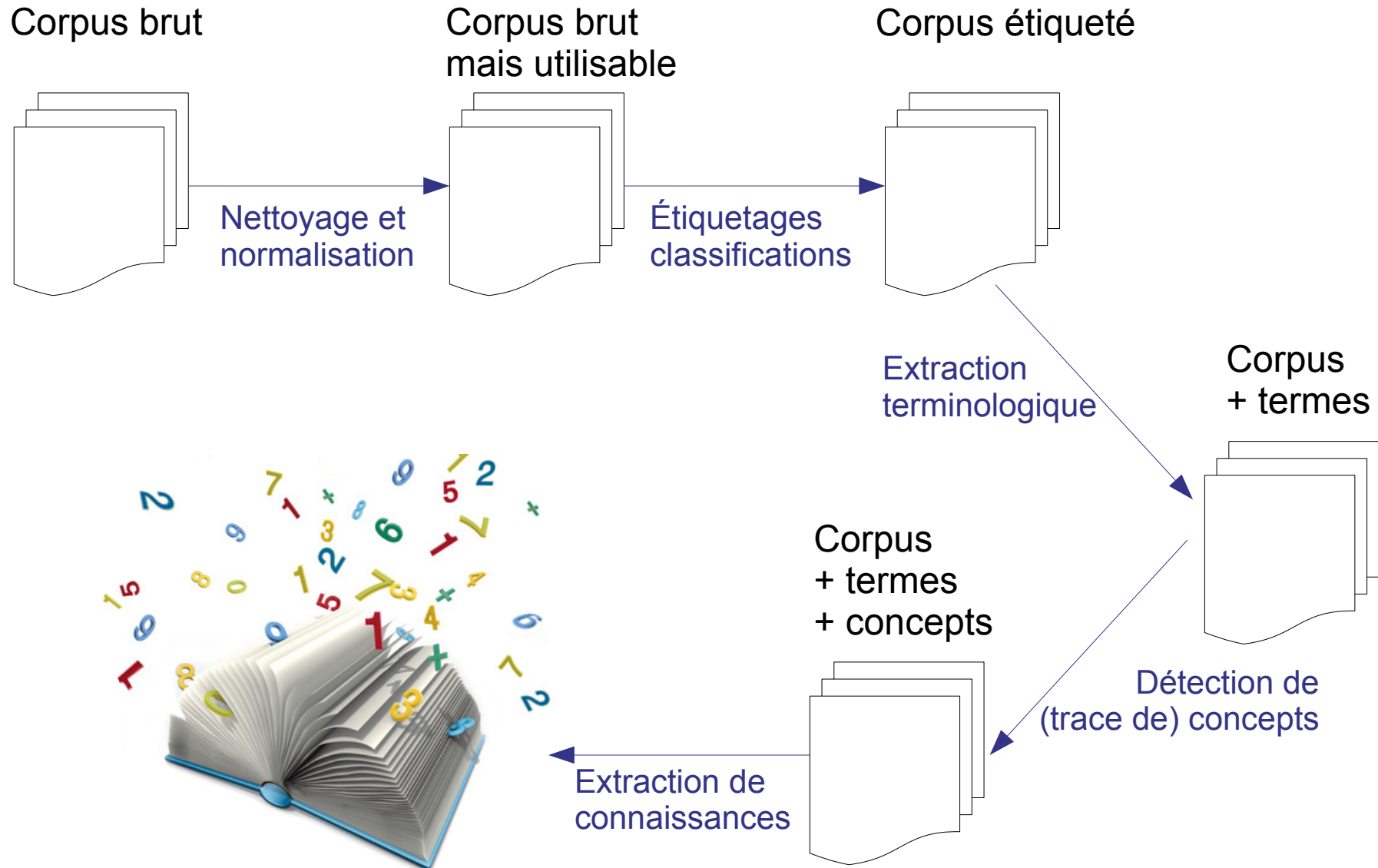
# Plan du cours FDT

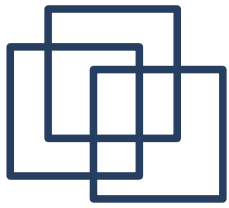
---

- Introduction aux problématiques de la fouille de textes (pourquoi, quoi, comment)
- Les différentes étapes d'un processus de fouille de textes
- Représentation vectorielle des textes pour la classification
- Introduction au package R « tm » (text mining)

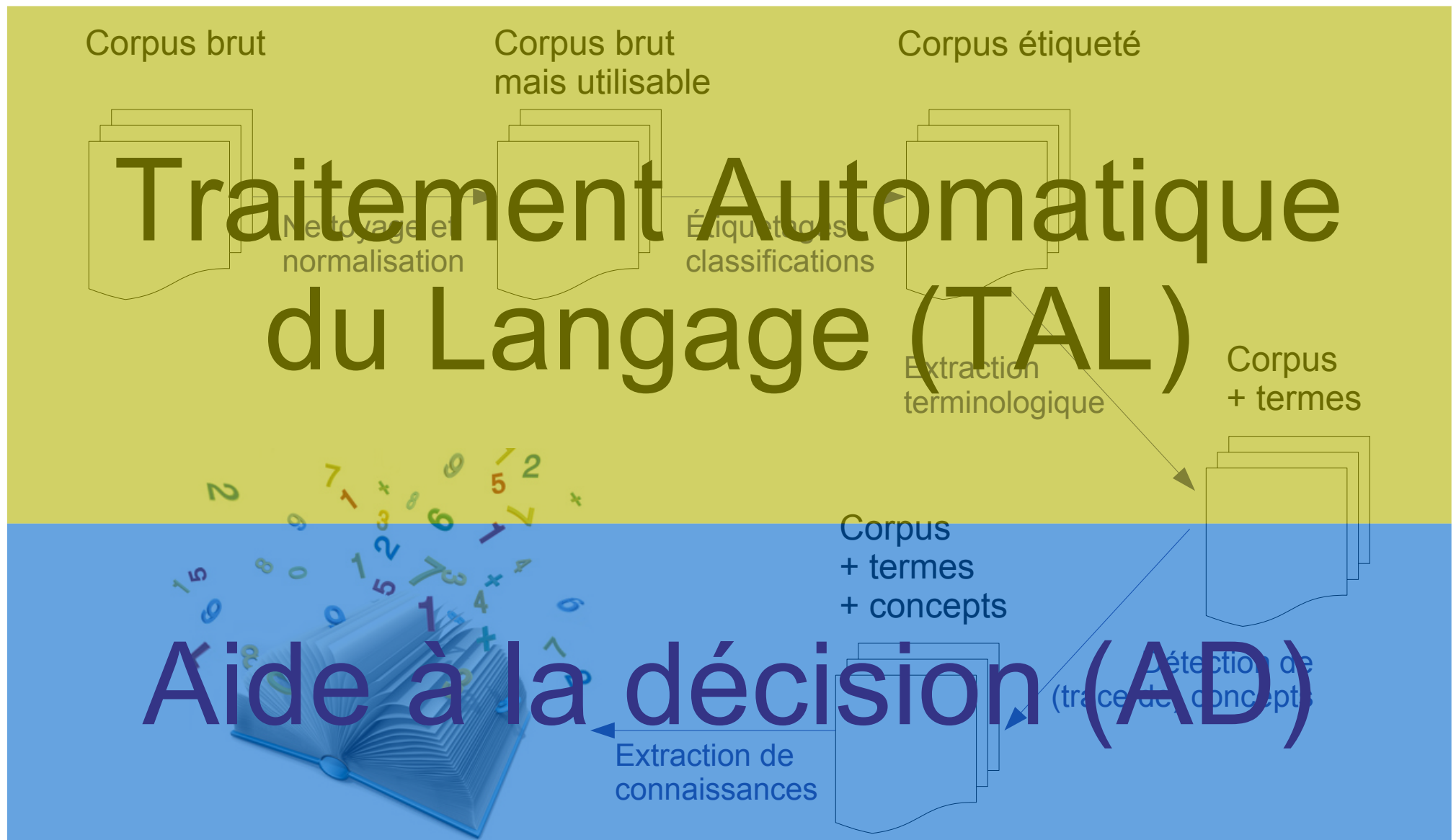


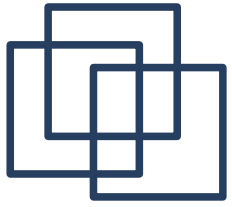
# Processus de FDT





# Processus de FDT





# Traitement de type TAL

---

## Objectif

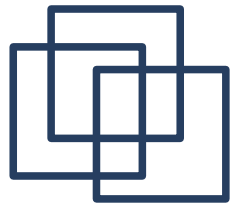
Extraire des caractéristiques sous une forme manipulable pour représenter l'information

Par quoi est portée l'information?

- Le mot, la ponctuation, la phrase, le paragraphe, ...
- Les balises.

## Les étapes de TAL

- Nettoyage et normalisation
- Étiquetages classifications
- Extraction terminologique



# Nettoyage et normalisation

---

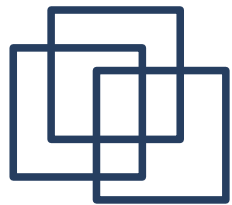
Création d'une forme intermédiaire des textes respectant l'information utile, proche de l'originale, prête pour l'extraction des caractéristiques.

## Nettoyage

- Extraire les parties utiles (balises HTML ou XML)
- Structurer le document si nécessaire
- Traiter l'encodage

## Normalisation

- Définir et segmenter les éléments d'information (e.g. mots)
- Limiter les variantes tout en respecter les nuances (e.g. l'Histoire)



# Nettoyage et normalisation

---

## Qu'est-ce qu'un mot?

Un token = des caractères entre deux espaces?

- C'était ou était mots. ou mots

Un token = des caractères entre deux signes de ponctuation?

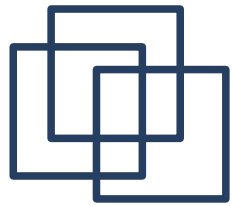
- Le ou le Livre ou livre vice-président ou vice et président

Un token = des caractères minuscules entre deux signes de ponctuation dans { , ; . ! ? « » ' ... } ?

- mots ou mot histoire ou Histoire dix milles ou 10,000

Un token = ... ?





# Nettoyage et normalisation

---

## Racinisation (stemming)

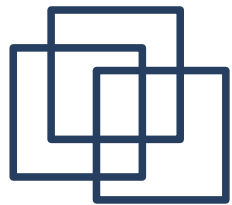
Consiste à réduire chaque mot à sa racine (traitement brutal).

$\{station, stationnement, stationnaire, stationner\} \rightarrow station$

## Lemmatisation (lemme)

Consiste à réduire chaque mot à sa forme canonique : retrait des marques de genre et de nombre (traitement plus léger).

$\{stations, station\}$	$\rightarrow station$
$\{stationnement, stationnements\}$	$\rightarrow stationnement$
$\{stationnait, stationnons\}$	$\rightarrow stationner$
...	



# Nettoyage et normalisation

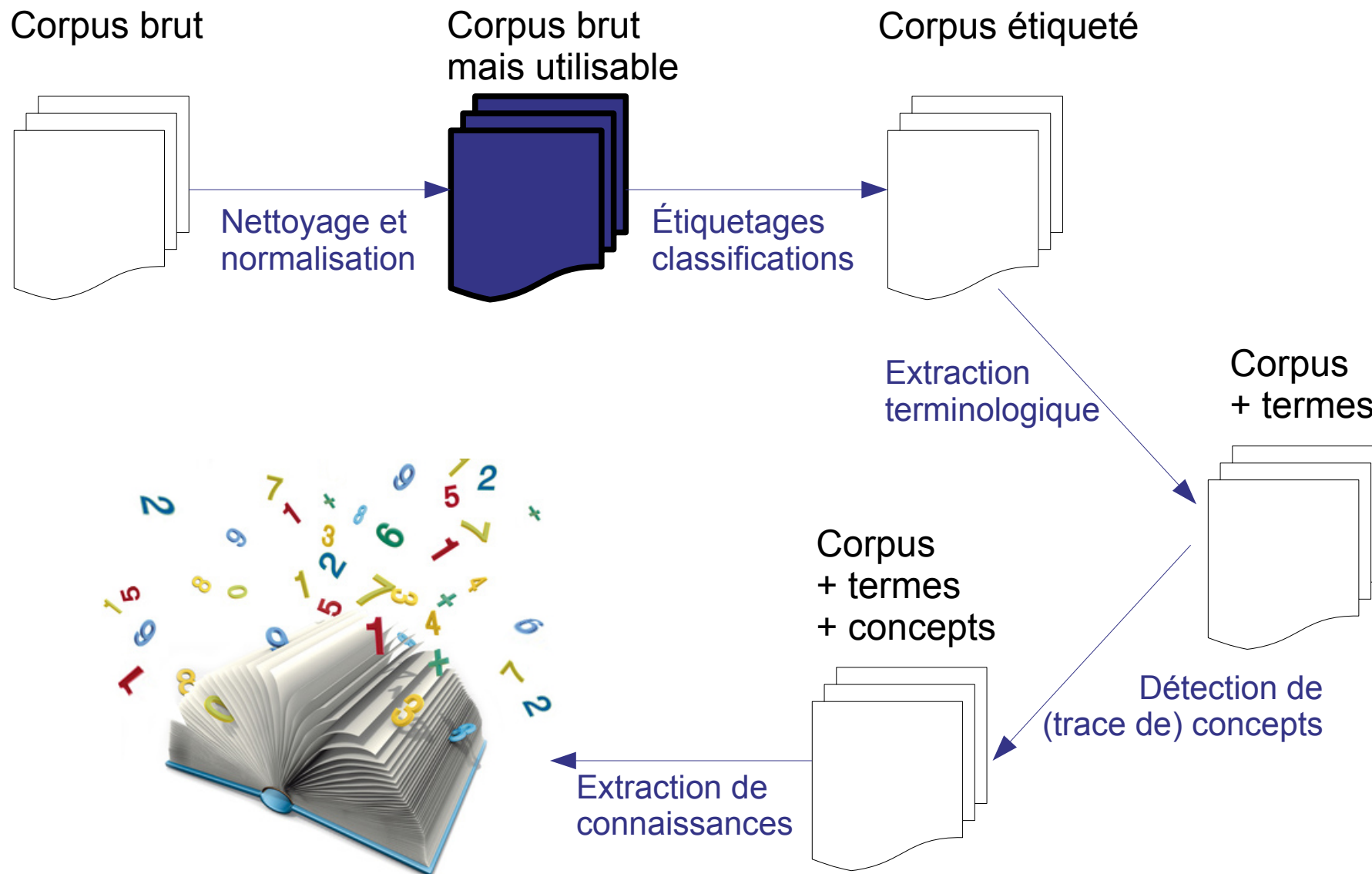
---

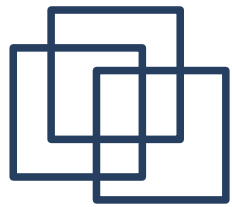
## Que peut-on utiliser d'autre que les mots?

- Les caractères ou suites de caractères (n-grammes de caractères)
- Les multi-mots :
  - entités nommées : "*Barack Obama*" "*vivendi universal*"
  - Collocations : "*pomme de terre*" "*faim de loup*"
  - N-grammes de mots : séquence de mots consécutifs
- Les balises (XML, HTML, liens hypertextes, etc.) et séquences de balises
- Des critères statistiques de plus bas niveau :
  - Longueur des mots
  - Longueur des phrases
  - Densité des documents (vocabulaire différentiel)
  - ...



# Processus de FDT





# Étiquetages et classifications

---

## Enrichir le texte par un étiquetage approprié

- Étiquetage grammatical

<u>Exemple</u>	Luc	mange	du	pain
	<NP>	<V>	<ART>	<NC>

- Étiquetage morphologique

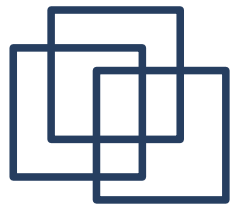
<u>Exemple</u>	Luc	mange	du	pain
	<NP>	<V>	<ART>	<NC>
		<Ind.Prés,1PS>	<masc. Sing>	<masc. Sing>

- Étiquetage syntaxique

<u>Exemple</u>	Luc	mange	du pain
	<Suj>	<V>	<Complément>

- Étiquetage sémantique

<u>Exemple</u>	Luc	mange	du pain
	<Personne>	<Action>	<nourriture>



# Étiquetages et classifications

---

## Problème de classification

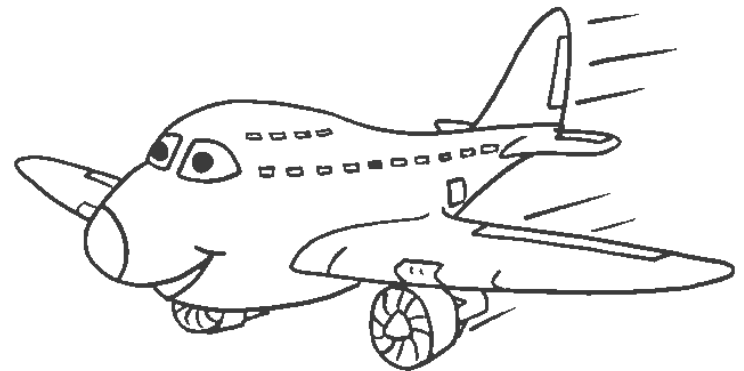
- Attribuer une étiquette à un mot ou groupe de mots

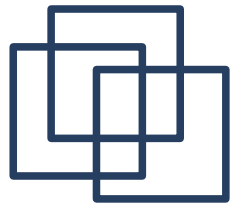
Exemple individu = « pain » → étiquettes = {NC, NP, V, Adj, etc...}

## Pas toujours facile (ambiguïté)

- Étiquetage grammatical

Exemple avions → V ou NC ?





# Étiquetages et classifications

---

## Problème de classification

- Attribuer une étiquette à un mot ou groupe de mots

Exemple individu = « pain » → étiquettes = {NC, NP, V, Adj, etc...}

## Pas toujours facile (ambiguïté)

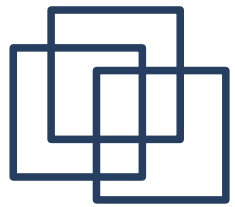
- Étiquetage grammatical

Exemple *avions* → V ou NC ?

- Étiquetage syntaxique

Exemple *la bonne soupe* → Groupe nominal ou Sujet+Verbe ?





# Étiquetages et classifications

---

## Problème de classification

- Attribuer une étiquette à un mot ou groupe de mots

Exemple individu = « pain » → étiquettes = {NC, NP, V, Adj, etc...}

## Pas toujours facile (ambiguïté)

- Étiquetage grammatical

Exemple avions → V ou NC ?

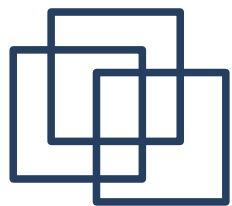
- Étiquetage syntaxique

Exemple la bonne soupe → Groupe nominal ou Sujet+Verbe ?

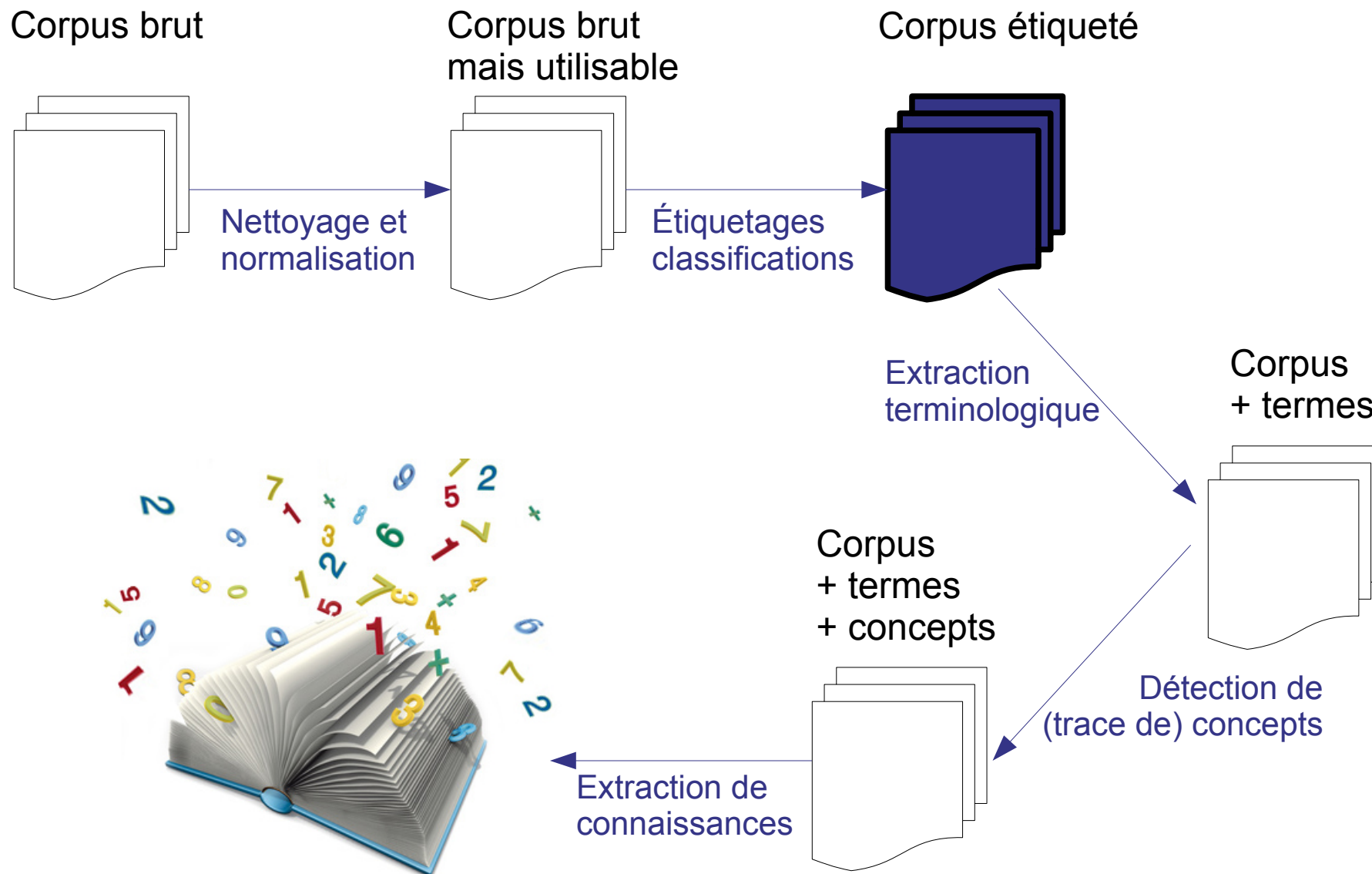
- Étiquetage sémantique

Exemple l'avocat est pourri → le fruit ou l'homme de loi ?





# Processus de FDT







# Extraction terminologique

---

N.B. Les traitements qui suivent concernent des tâches particulières de Fouille de Textes, telles que la classification thématique de documents.

## Hypothèse du sac de mots

### Une représentation simple

- Un document est un sac
- Ce sac contient des «mots» (le plus souvent des lemmes ou tokens) qui apparaissent une ou plusieurs fois (fréquences d'apparition)

### Hypothèse simplificatrice

- L'ordre des mots est ignoré
- La structure du texte est mise à plat.

### Choix

- Les mots représentent-ils le texte? Doit-on préférer des lemmes ou des tokens? Comment choisir les mots pertinents pour la tâche visée?



# Extraction terminologique

---

## Fréquence des mots dans « Le Cid »

1	429	de	40	67	trop
2	264	<b>l'</b>	41	67	<b>Rodrigue</b>
3	259	<b>?</b>	42	65	j'
4	258	et	43	65	du
5	245	un	44	65	Mais
6	230	en	45	64	au
7	229	le	46	63	honneur
8	220	que	47	63	ai
9	201	mon	48	62	bien
10	198	est	49	61	des
11	191	<b>Et</b>	50	59	fait
12	189	d'	51	58	ta
13	187	je	52	57	Que
14	177	la	53	55	te
15	167	il	54	54	<b>amour</b>
16	155	vous	55	54	<b>Chimène</b>
17	151	qu'	56	53	ton
18	142	ma			





# Extraction terminologique

---

## Solution (partielle) 1 : utiliser une Stop List

Utiliser une liste de mots dits « mots-outils » définie linguistiquement dans chaque langue (articles, coordinations, pronoms, etc.)

## Solution (partielle) 2 : supprimer les mots fréquents

Ne pas considérer les mots les plus fréquents car peu informatifs. Traitement statistique permettant d'isoler des « mots-outils ».

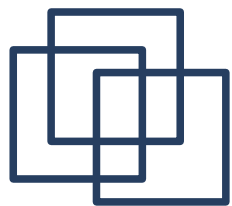
- Quel seuil de fréquence utiliser?

Exemple (Le petit prince) :

*le, de, je, il, et, les, un, la, petit, pas, à, prince,...*

Les mots-outils sont-ils vraiment inutiles?

- **Non** pour les tâches d'étiquetage, de reconnaissance de la langue ou de la parole
- **Oui** pour l'indexation de documents



# Extraction terminologique

---

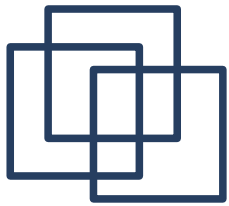
## Solution (partielle) 3 : supprimer les mots infréquents

Ne pas considérer les mots qui apparaissent moins de X fois ou dans moins de Y documents différent.

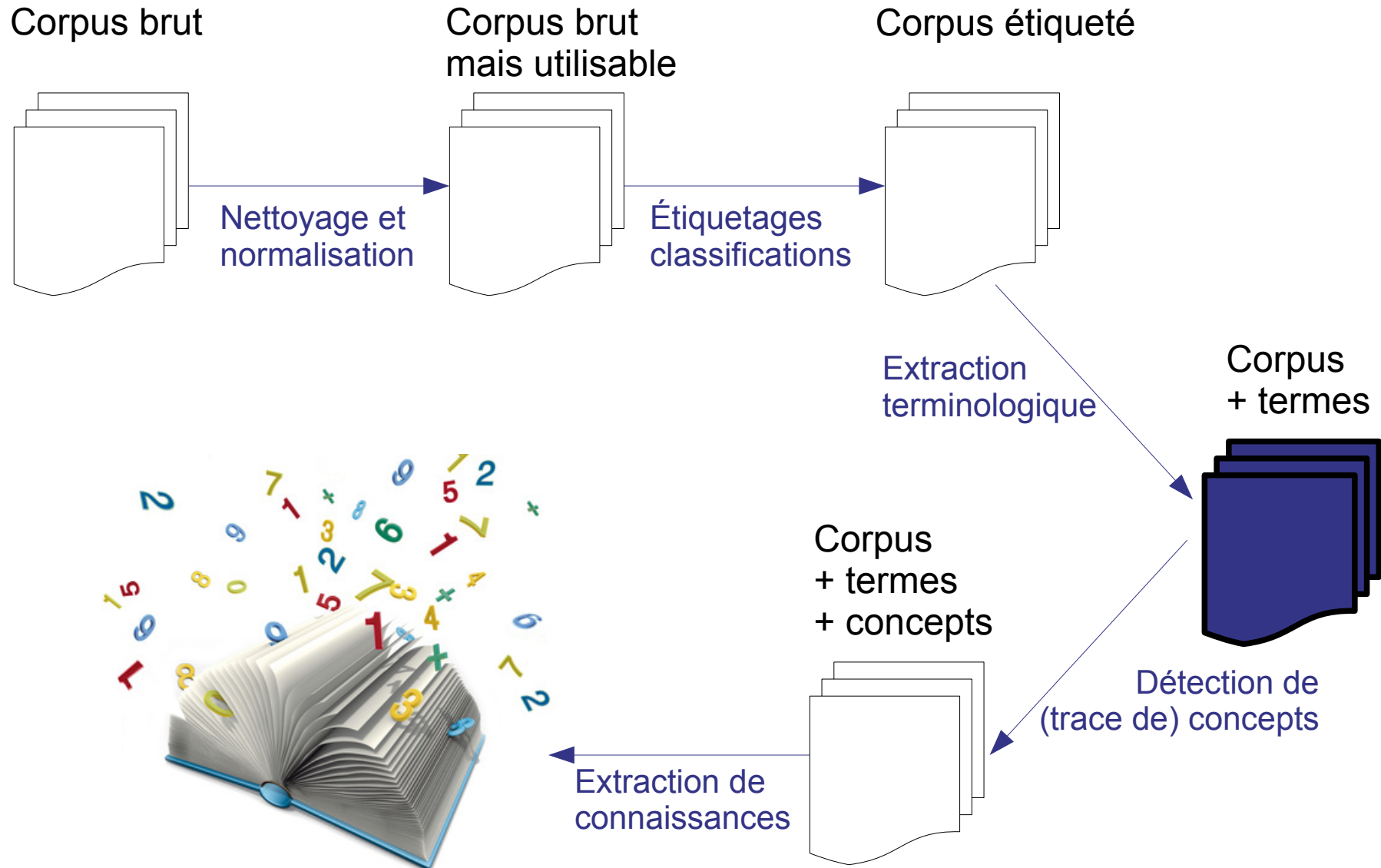
## Solution (partielle) 4 : utiliser l'étiquetage grammatical

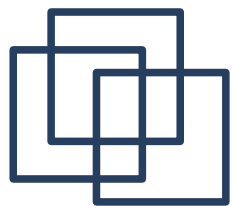
Ne considérer, à l'inverse, que certaines étiquettes grammaticales jugées informatives : NC, NP, V, Adj, etc.

NB. Les quatre solutions envisagées sont non-exclusives et leur utilisation dépend des informations disponibles (étiquetage).



# Processus de FDT





# Représentation vectorielle

---

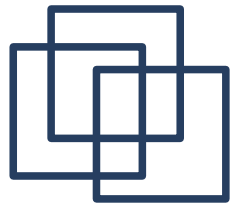
## Représentation exploitable par un système d'Aide à la décision

Chaque document sera représenté par un vecteur dont :

- Les attributs/descripteurs sont les mots du vocabulaire du corpus
- Les valeurs sont binaires (présence/absence) ou entières (occurrences)

Exemple construire la représentation vectorielle du corpus suivant (3 doc.)

Forme brute	Forme normalisée
<i>Je suis passé à la radio.</i> <i>J'ai passé une radio.</i> <i>Les radios sont passées au numérique.</i>	<i>Je être passer à la radio</i> <i>Je avoir passer une radio</i> <i>Le radio être passer au numérique</i>
<u>vocabulaire</u>  (être, passer, radio, ,numérique)	$d_1 = (1, 1, 1, 0)$ $d_2 = (0, 1, 1, 0)$ $d_3 = (1, 1, 1, 1)$

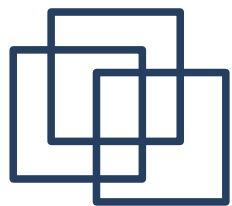


# Représentation vectorielle

---

## Remarques sur la représentation vectorielle

- Représentation de type « sac de mots » : perte de la séquentialité
- Si un document contient  $n$  mots différents, il y a  $n$  composantes non-nulles et  $V-n$  composantes nulles!
- Vecteurs très « creux » (matrices documents x mots très éparses)
- Prévoir une structure de données adaptée



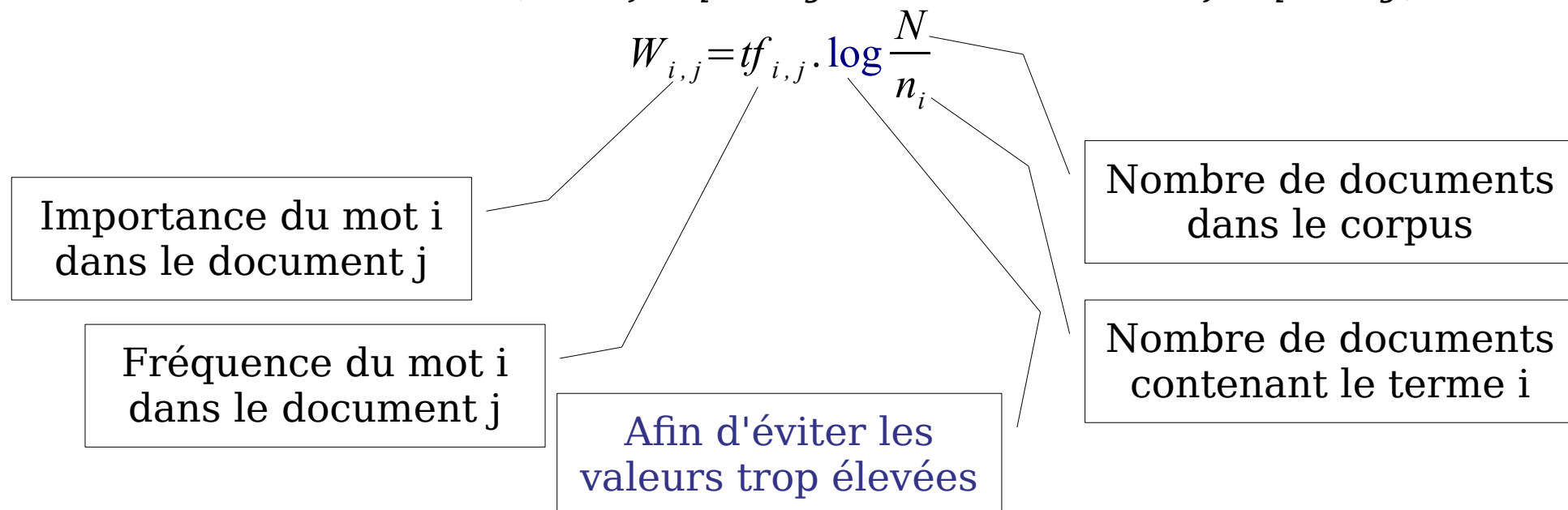
# Représentation vectorielle

## Calculer l'importance d'un mot dans un document

Une alternative aux représentations de type présence/absence ou nombre d'occurrences : associer un poids à un mot dans un document.

Un mot sera jugé important pour un document parcequ'il est fréquent dans ce document et rare dans les autres documents.

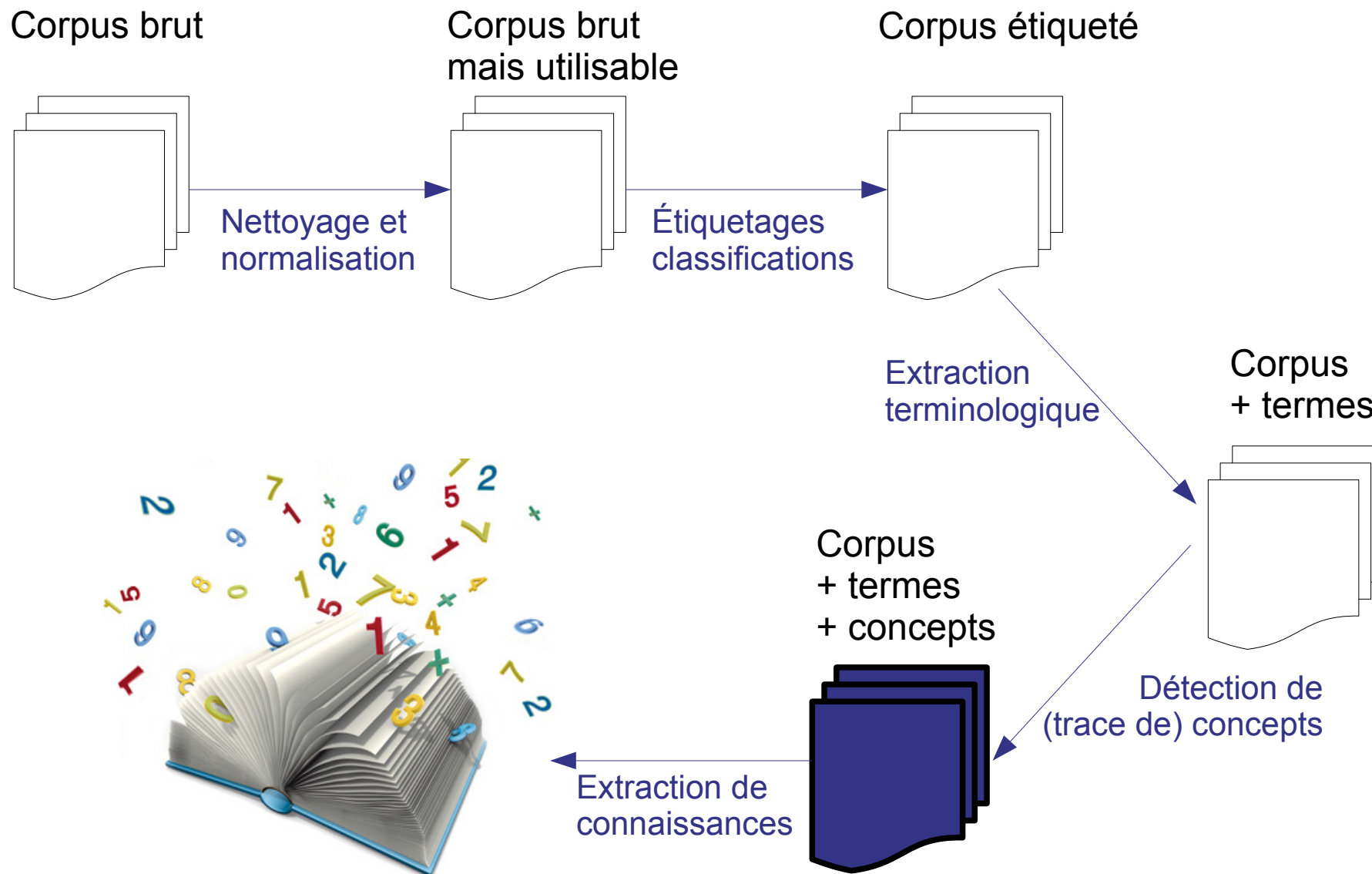
- La mesure du tf x idf (*term frequency inverse document frequency*)

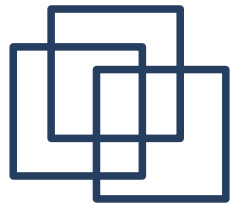






# Processus de FDT



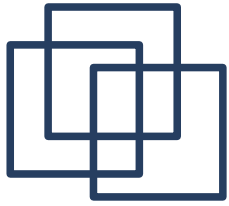


# Classification de documents

---

A partir d'une représentation vectorielle des documents on peut...

- Utiliser un classifieur de Bayes (représentations binaires)
- Calculer des distances entre documents (plus proches voisins)
- Construire un arbre de décision
- Calculer des séparateurs linéaires (réseaux de neurones, SVM)
- ...



# Calcul de distances

---

## Documents similaires

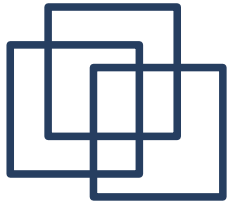
- Les documents qui se ressemblent contiennent les mêmes mots ou des mots similaires
- Hypothèse distributionnelle de Harris : les mots qui ont des contextes identiques sont similaires

## Vecteurs similaires

- Dans l'espace vectoriel ils correspondent à des vecteurs proches

## Représentation dans l'espace

- Dans l'espace vectoriel de dimension  $V$ , les vecteurs représentant les textes forment un faisceau de même origine
- Les vecteurs proches ont des **directions quasi-identiques** ou des **extrémités proches**



# Distance euclidienne

---

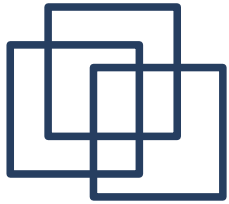
## Distance géométrique entre les extrémités des vecteurs

- Utiliser les vecteurs d'occurrences ou de pondération de type tf.idf
- Normaliser les vecteurs par la norme  $L_2$

$$\|d_i\|_{L_2} = \sqrt{d_{i,1}^2 + \dots + d_{i,V}^2}$$

- Calculer la distance euclidienne entre deux vecteurs (après normalisation)

$$d(x_i, x_j) = \sqrt{(x_{i,1} - x_{j,1})^2 + \dots + (x_{i,V} - x_{j,V})^2}$$



# Similarité du cosinus

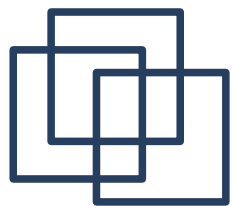
---

## Mesure de l'angle formé par deux vecteurs

- Utiliser les vecteurs d'occurrences ou de pondération de type tf.idf
- Calculer le cosinus de l'angle formé par les vecteurs  $d_i$  et  $d_j$

$$\cos(d_i, d_j) = \frac{\langle d_i, d_j \rangle}{\|d_i\|_{L_2} \cdot \|d_j\|_{L_2}}$$

- Valeurs dans  $[0,1]$  : 0 pour des vecteurs superposés, 1 pour des vecteurs orthogonaux
- On se ramène à une distance par l'opération : distance = 1 - similarité (car similarité dans  $[0,1]$ )



# Divergence de Kullback-Leibler

---

## Distance entre 2 distributions (probabilistes)

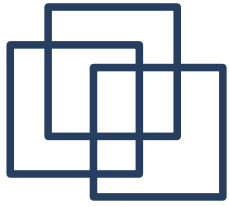
- Utiliser les vecteurs d'occurrences
- Normaliser les vecteurs par la norme  $L_1$  afin d'obtenir des distributions

$$\|d_i\|_{L_1} = d_{i,1} + \dots + d_{i,V}$$

- Calculer la « divergence » entre deux distributions  $p_i$  et  $p_j$

$$D_{KL}(p_i \| p_j) = \sum_{k=1}^V p_{i,k} \log \frac{p_{i,k}}{p_{j,k}}$$

- ATTENTION !! cette divergence n'est pas symétrique



# Indice de Jaccard

---

## Similarité lexicale entre deux textes

- Utiliser les vecteurs binaires : présence/absence
- Calculer les quantités :

$a_{i,j}$  = nombre de composantes où  $d_{i,k} = d_{j,k} = 1$

$b_{i,j}$  = nombre de composantes où  $d_{i,k} = 1$  et  $d_{j,k} = 0$

$c_{i,j}$  = nombre de composantes où  $d_{i,k} = 0$  et  $d_{j,k} = 1$

- Calculer l'indice de similarité de Jaccard :

$$Jaccard(d_i, d_j) = \frac{a_{i,j}}{a_{i,j} + b_{i,j} + c_{i,j}}$$