

Extraction d'information

LGI2P Doctoral Research Day 2016 – 2017

---

**TAGNY NGOMPE Gildas<sup>??,??</sup>, Sébastien Harispe<sup>??</sup>, Jacky Montmain<sup>??</sup>, Stéphane Mussard<sup>??</sup>, Guillaume Zambrano<sup>??</sup>**

24 mars 2017

1. LGI2P (École des mines d'Alès)
2. CHROME EA 7352 (Université de Nîmes)





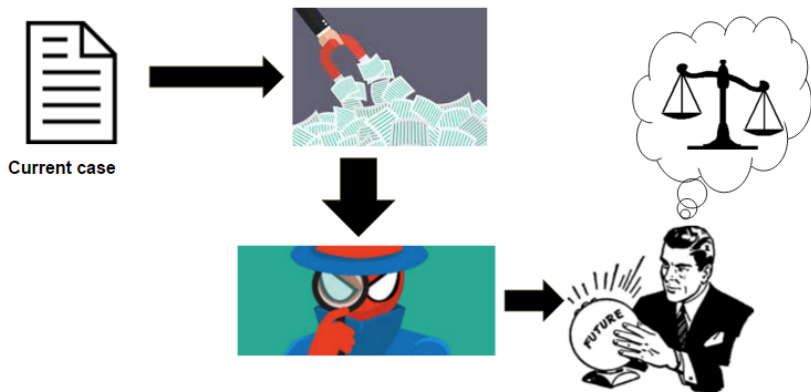
problématiques

---

## Analyse Sémantique d'un Corpus Exhaustif de Décisions Jurisprudentielles pour l'Élaboration d'un Modèle Prédictif du Risque Judiciaire



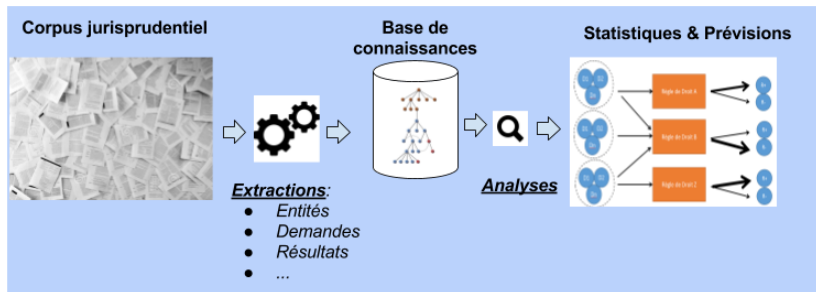
# Les juristes analysent les décisions afin d'anticiper



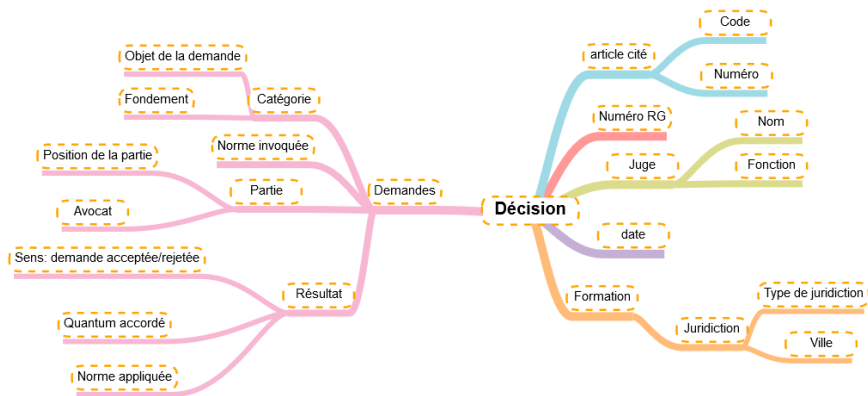
# Défis liés à la recherche et à l'analyse

- Grande quantité des décisions
- Documents non-structurés
- Complexité de l'organisation de la justice
- Compréhension difficile du langage juridique

# Objectif : un pipeline d'analyse de corpus jurisprudentiels



# Informations pertinentes à extraire





# Structure dans la base de connaissances

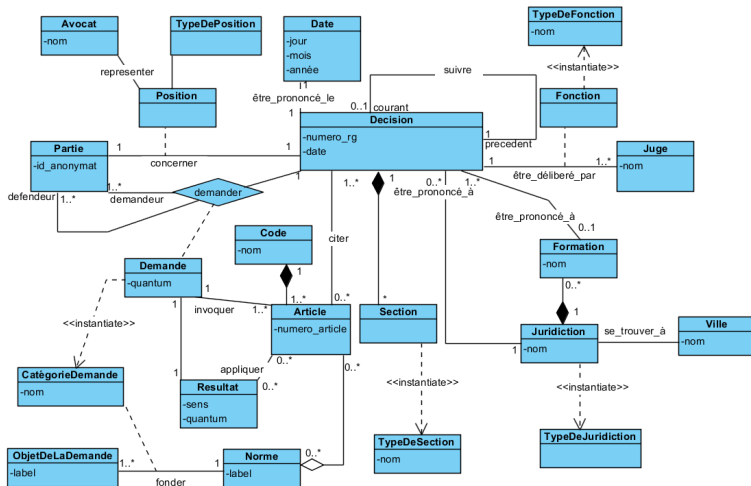


FIGURE – Modèle des données



# Informations noyées dans des textes non-structurées

ARRÊT N°

R.G : 11/03924

...

COUR D'APPEL DE NÎMES

CHAMBRE CIVILE

1ère Chambre A

ARRÊT DU 20 MARS 2012

APPELANTE :

Madame Michèle A. ...

assistée de la SELARL VAJOU, ...

INTIMES :

Monsieur Martial B ...

assisté de la SCP MARION GUIZARD

PATRICIA SERVAIS, ...

COMPOSITION DE LA COUR LORS DU

DÉLIBÉRÉ :

M. Dominique BRUZY, Président

M. Serge BERTHET, Conseiller

...

FAITS, PROCEDURE, ...

Madame Michèle A. demande :

...

- de condamner Madame JONES-B. à lui  
payer la somme de 2.500 euros au titre de  
l'article 700 du Code de Procédure Civile,

PAR CES MOTIFS, LA COUR :

...

Vu l'article 809 du Code de Procédure  
Civile,

...

Déboute Madame A. de sa demande de  
provision sur dommages-intérêts.

...

Vu l'article 700 du Code de Procédure  
Civile,

Condamne Madame JONES-B. à verser à  
Madame A. la somme de 2.500 euros.

# Sectionner les décisions pour organiser l'extraction

ARRÊT N°

R.G : 11/03924

COUR D'APPEL DE NÎMES  
CHAMBRE CIVILE

1ère Chambre A

ARRÊT DU 20 MARS 2012

APPELANTE :

Madame Michèle A. ...

assistée de la SELARL VAJOU, ...

INTIMES :

Monsieur Martial B ...

assisté de la SCP MARION GUIZARD  
PATRICIA SERVAIS, ...

COMPOSITION DE LA COUR LORS  
DU DÉLIBÉRÉ :

M. Dominique BRUZY, Président

M. Serge BERTHET, Conseiller

...

FAITS, PROCEDURE, ...

Madame Michèle A. demande :

...

- de condamner Madame JONES-B. à lui payer  
la somme de 2.500 euros au titre de l'article 700  
du Code de Procédure Civile,

## Corps : demandes et normes

PAR CES MOTIFS, LA COUR :

...

Vu l'article 809 du Code de Procédure Civile,

...

Déboute Madame A. de sa demande de provi-  
sion sur dommages-intérêts.

...

Vu l'article 700 du Code de Procédure Civile,  
Condamne Madame JONES-B. à verser à Ma-  
dame A. la somme de 2.500 euros.

## Dispositif : résultats et normes

## Entêtes : méta-données

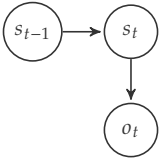
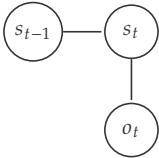
# Entités et sections à détecter

Entités	Labels	Exemples
<b>Section entête (E)</b>		
Numéro R.G.	<b>RG</b>	"10/02324", "60/JAF/09"
Ville	<b>VL</b>	"NÎMES", "Agen", "Toulouse"
Type de juridiction	<b>JR</b>	"COUR D'APPEL"
Formation	<b>FM</b>	"1re chambre", "Chambre économique"
Date	<b>DT</b>	"01 MARS 2012", "15/04/2014"
Partie appelante	<b>AP</b>	"SARL K.", "Syndicat ...", "Mme X ..."
Partie intimée	<b>IM</b>	- // -
Partie intervenante	<b>IV</b>	- // -
Avocat	<b>AV</b>	"Me Dominique A., avocat au barreau de Papeete"
Juge	<b>JG</b>	"Monsieur André R.", "Mme BOUSQUEL"
fonction du juge	<b>FT</b>	"Conseiller", "Président"
<b>Corps (T) et dispositif (D)</b>		
Norme	<b>NO</b>	"l' article 700 NCPC", "articles 901 et 903"
Élément à éviter	<b>O</b>	<i>tout élément ne faisant partie d'aucune entité ciblée</i>

TABLE – Entités et leurs labels par section.

# Approches probabilistes d'étiquetage de séquence

## Modèles probabilistes à états et observations

HMM	CRF
un seul descripteur par observation	plusieurs descripteurs complexes par observation
	
$P_{\lambda}(S O) = \prod_{t=1}^T P(s_t s_{t-1}) * P(o_t s_t)$ <p>[?]</p>	$P_{\lambda}(S O) = \frac{1}{Z(O)} \exp \left( \sum_{t=1}^T \sum_k \lambda_k f_k(s_{t-1}, s_t, o_t) \right)$ <p>[?]</p>

Objectif : Trouver la séquence la plus probable d'étiquetage pour l'ensemble du texte

**Entrainement fait sur des séquences préalablement étiquetées**

# Introduire des descripteurs discriminants dans les modèles

Exemple : Soit l'annotation manuelle :

« ... l' @NO article 700 du code de procédure #NO ... »

Introduction des caractéristiques au niveau de  $t_i = \text{« 700 »}$

$$f_1(l_{i-1}, l_i, t_{1:n}, i) = \begin{cases} b_1(T, i) & \text{si } l_{i-1} = \text{NORME} \wedge l_i = \text{NORME} \\ 0 & \text{sinon} \end{cases}$$

$$f_2(l_{i-1}, l_i, t_{1:n}, i) = \begin{cases} b_2(T, i) & \text{si } l_i = \text{NORME} \\ 0 & \text{sinon} \end{cases}$$

avec

$$b_1(T, i) = \begin{cases} 1 & \text{si } (t_{i-1} = \text{article}) \wedge (POS_{i-1} = \text{NOM}) \\ & \wedge (NP1_{i-1} = \text{<unknown>}) \wedge (NS1_{i-1} = \text{@card@}) \\ 0 & \text{sinon} \end{cases}$$

$$b_2(T, i) = \begin{cases} 1 & \text{si } (t_i = 700) \wedge (POS_i = \text{NUM}) \wedge (NP1_i = \text{article}) \wedge (NS1_i = \text{code}) \\ 0 & \text{sinon} \end{cases}$$

[?]

# Approche d'évaluation des modèles

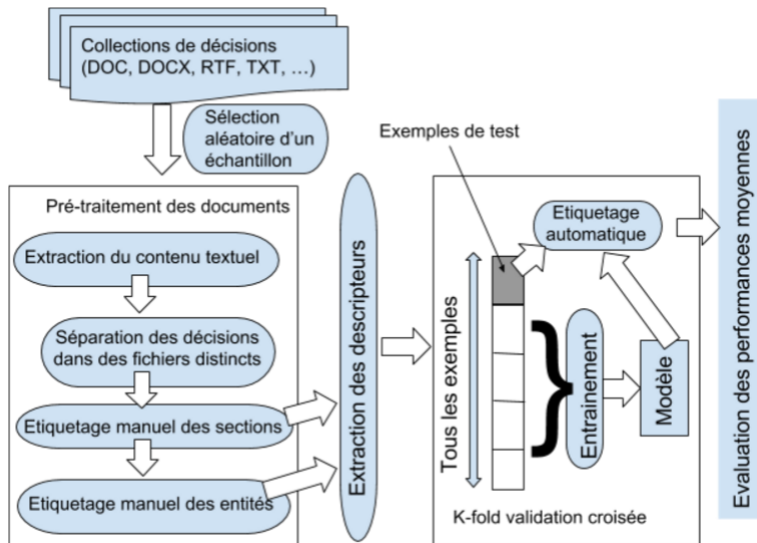


FIGURE – Évaluation des modèles.



# Conditions de tests

- 505 décisions de cour d'appel annotées manuellement
- Implémentation java (Mallet [?])
- TreeTagger : extraction de lemmes et rôles grammaticaux
- Implémentation de l'extraction de descripteurs (page ??)
- 5-fold validation croisée
- Précision (P), rappel (R), F1-mesure (F1) :

$$P_l = \frac{\text{nombre d'éléments correctement étiquetés par le modèle avec } l}{\text{nombre d'éléments étiquetés par le modèle avec } l}$$

$$R_l = \frac{\text{nombre d'éléments correctement étiquetés par le modèle avec } l}{\text{nombre d'éléments manuellement étiquetés avec } l}$$

$$F1_l = 2 \times \frac{P_l \times R_l}{P_l + R_l}$$

avec  $l$ =label

# Evaluation de la détection des sections

	HMM			CRF-			CRF+		
<i>labels</i>	P	R	F1	P	R	F1	P	R	F1
E (Entete)	84.2	91.8	87.8	93.8	85.4	89.3	99.3	99.6	99.5
T (Corps)	88.4	63.9	74.1	86.3	98.2	91.8	99.8	99.5	99.7
D (Dispositif)	15.4	47.0	23.0	100.0	8.5	15.6	98.0	100.0	98.9
<i>Moyenne</i>	62.7	67.6	67.6	93.3	64.0	64.0	99.7	99.8	99.8

TABLE – Précision (P), rappel (R), F1-mesure (F1) au niveau des lignes (%).

# Evaluation de la détection des entités

	HMM			CRF-			CRF+		
<i>labels</i>	P	R	F1	P	R	F1	P	R	F1
<i>Section Entête (E)</i>									
AP	35.3	14.1	20.1	64.9	48.8	55.6	92.0	86.7	89.3
AV	83.8	98.3	90.5	96.4	97.5	96.9	97.6	98.1	97.9
DT	70.9	72.6	71.7	94.4	86.8	90.4	98.8	97.7	98.2
FM	87.6	93.7	90.5	98.8	98.4	98.6	98.9	99.3	99.1
FT	88.8	59.8	71.3	94.2	92.3	93.3	97.1	95.5	96.3
IM	53.1	57.4	55.1	67.2	64.6	65.8	89.3	88.1	88.7
IV	-	2.2	-	25.9	26.5	26.2	67.3	41.4	46.4
JG	68.0	85.7	75.7	96.2	95.7	96.0	98.1	97.7	97.9
JR	75.8	99.5	86.0	98.6	99.4	99.0	99.3	99.4	99.4
RG	-	0	-	83.7	46.1	59.4	98.6	97.4	98.0
VL	93.1	27.9	42.6	98.2	98.4	98.3	99.0	99.0	99.0
<i>Sections inférieures (T &amp; D)</i>									
NO	92.9	90.9	91.9	96.0	93.8	94.9	97.9	96.5	97.2

TABLE – Précision (P), rappel (R), F1-mesure (F1) au niveau des mots (9%)



## Informations pertinentes à extraire

- **Position de la partie** : Intimé
- **Catégorie de demande** : Dommages-intérêts pour procédure abusive
  - **Objet** : Dommages-intérêts
  - **Fondement** : Articles 1382 code civil et 32-1 code de procédure civile
- **Quantum demandé** : 20 000 euros
- **Résultat** : Rejet
- **Quantum accordé** : 0 euros

# Expression plus ou moins explicite

## EXPRESSION DE DEMANDE (EXPLICITE / IMPLICITE ?)

La société A. conclut à la confirmation du jugement entrepris sauf à former appel incident sur la disposition du jugement l'ayant déboutée de sa demande de **dommages intérêts pour abus de procédure** et elle demande à la cour de condamner l'appelante à lui payer la somme de **20 000 euros** à titre de dommages intérêts ...

...

## EXPRESSION IMPLICITE DE RESULTAT

La cour, ...

**Confirme la décision entreprise en toutes ses dispositions,**

# Extraire les demandes suivants leur catégorie

## Exemples de catégorie de demande

- dommages et intérêts pour procédure abusive,
- dommages et intérêts pour concurrence déloyale,
- dommages et intérêts pour trouble de voisinage,
- prestation compensatoire,
- torts exclusifs,
- droit de visite,
- etc.

## Définition d'une classe de décision

Soit  $C$  une catégorie de demande et  $D$  une décision,  
s'il existe dans  $D$  une demande  $d$  de catégorie  $C$ , alors  $C$  est  
une classe de  $D$

- une décision comprend plusieurs demandes de catégories variées
- toutes les catégories ne sont pas connues d'avance



# Catégorisation semi-supervisée des décisions

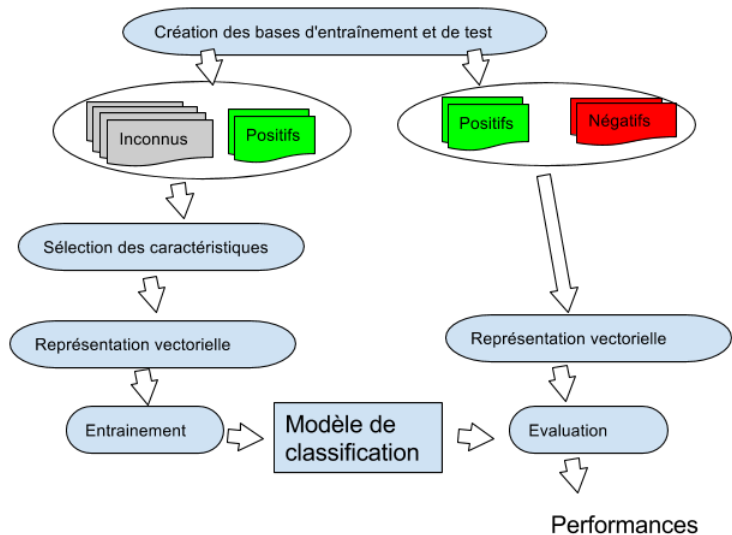
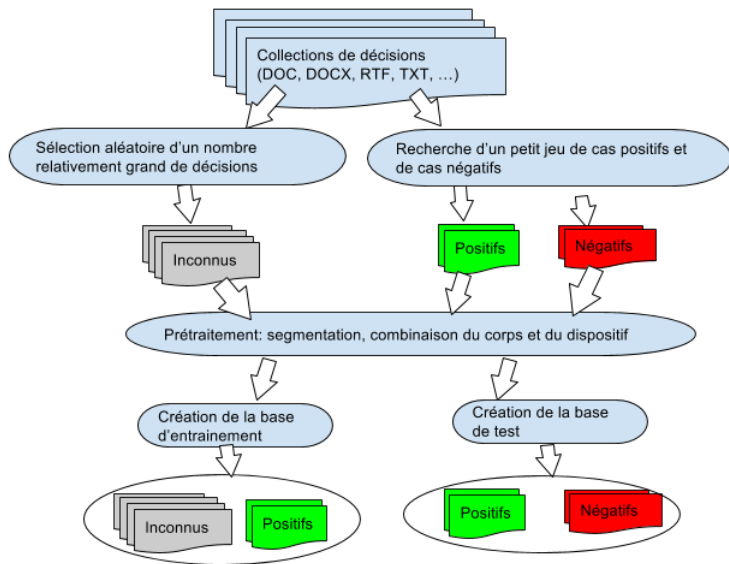


FIGURE – Approche d'expérimentation de la classification

# Création d'une base d'apprentissage



# Conditions d'évaluation de la catégorisation des décisions

- Représentation vectorielle :

$$poids(w*, t) = poids_{local}(w*, t) * poids_{global}(w*) * facteur_{normalisation}$$

- Évaluation de différentes configurations :
  - dimensions des vecteurs : 2, 10, ..., 250, ...
  - méthodes de sélection de termes discriminants (p. ?? & ??) :  $\chi^2$ ,  $\Delta_{DF}$ , *Marascuilo*, *NGL*, *GSS* ...
  - méthodes de classification : SVM, arbre de décision, KNN, naïf bayésien (avec Weka[?])
  - méthodes de pondération locale (p. ??) : TF, LogTF, ATF, TP
- environ 2000 cas inconnus,
- dommages-intérêts pour abus de procédure : entraînement 152 positifs, test 39 positifs + 157 négatifs
- prestation compensatoire : entraînement 100 positifs, test 100 positifs + 100 négatifs
- ...

$$P_C = \frac{\text{nombre de décisions test correctement classées par le modèle dans } C}{\text{nombre de décisions test effectivement dans } C}$$

$$R_C = \frac{\text{nombre de décisions test correctement classées par le modèle dans } C}{\text{nombre de décisions test effectivement dans } C}$$

$$F1_C = 2 \times \frac{P_C \times R_C}{P_C + R_C}$$

avec  $C$  = une classe de décision

# Premières évaluation de la classification des décisions

Est-ce l'effet de la méthode de :

- constitution des exemples d'apprentissage/test
- sélection de caractéristiques

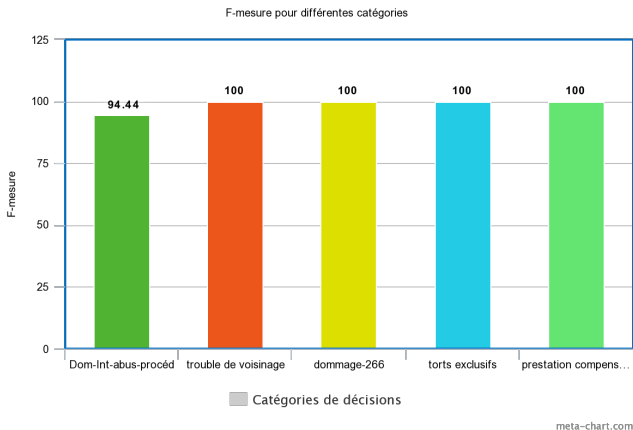


FIGURE – Performance actuelles de classification



## SITUATION CONCRETE

mon mari demande le divorce,

nous avons un prêt en commun qui finance une maison construite sur mon terrain (donation).

Je veux bien reprendre le prêt à mon nom au moment du divorce car il ne veut plus payer cette maison, mais il me réclame tout ce qu'il a payé durant notre mariage :prêt+travaux. **en a t'il le droit ?**

j'estime ne rien lui devoir ce qui a été fait est un héritage pour nos enfants. Ce qui est sur mon terrain m'appartient je pense.

en vous remerciant

Source : <http://www.documentissime.fr/questions-droit/question-46724-emprunt-et-divorce.html>

## Recherche des affaires/décisions « *similaires* »

- ☐ catégorie de demandes
- ☐ éléments factuels

# Similarité avec le challenge COLIEE 2017

## REQUETE : UNE SITUATION ABSTRAITE

There is a limitation period on pursuance of warranty if there is restriction due to superficies on the subject matter, but there is no restriction on pursuance of warranty if the seller's rights were revoked due to execution of the mortgage.

## 1. QUELS ARTICLES PERMETTENT DE JUGER CETTE SITUATION ?

**Article 566 (1)**In cases where the subject matter of the sale ...

**Article 567(1)**If the buyer loses his/her ownership of ...

## 2. EST-CE UNE SITUATION EN ACCORD AVEC LES ARTICLES ?

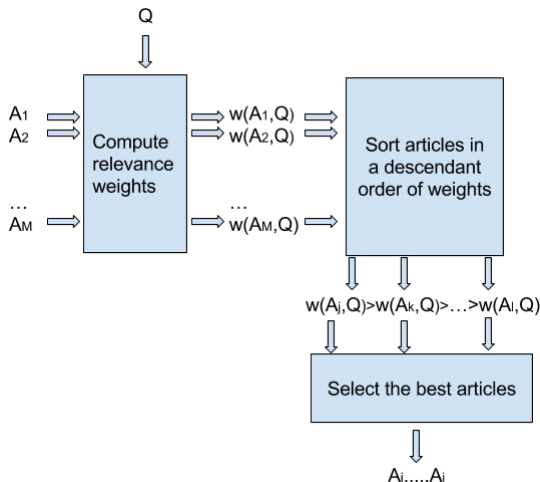
Oui

Source : <http://webdocs.cs.ualberta.ca/~miyoung2/COLIEE2017/>



# Recherche des articles

Utilisation d'une méthode d'estimation de la "pertinence"  
 $w(A, Q)$  d'un texte  $A$  pour une requête  $Q$  (ex. BM25)





- Détection d'entités et de sections
  - Performances encourageantes du CRF
  - Difficultés :
    - Annotation manuelle d'un jeu suffisant d'exemples
    - Identification de bons descripteurs
  - Limite de l'approche :
    - descripteurs définis manuellement (portabilité des modèles)
- Classification des décisions :
  - à partir d'un petit nombre de décisions d'une catégorie, il est possible de retrouver les termes caractéristiques des catégories

- Détection d'entités et de sections
  - Amélioration des performances (ex. plus d'exemples d'entêtes avec les intervenants)
  - Apprendre automatiquement une représentation (ex. deep learning)
  - Résolution et désambiguïsation des entités

*article 700 = article 700 du Code de Procédure Civile*

- Etendre l'étude à d'autres juridictions
- Extraction de demandes :
  - Comment exploiter les termes clés d'une catégorie pour retrouver les demandes dans les décisions ?

Objectif : Mettre sur pied des moyens efficaces :

- de structuration d'un corpus exhaustif des décisions (en général)
- d'analyses dans ce corpus

Défis divers à relever en informatique :

- extraction d'information
- représentation des connaissances
- recherche d'information





# Descripteurs de lignes pour le sectionnement

1. HMM : numéro de ligne
2. CRF :
  - numéro de ligne
  - toute la ligne
  - les 3 premiers termes
  - le nombre de termes
  - 1er terme des lignes suivantes et précédentes
  - ...



Pour les entités de l'entête :

1. HMM : le mot
2. CRF :
  - mot
  - lemme
  - rôle grammatical
  - commence-t-il par une lettre majuscule ?
  - le texte contient-il la chaîne « intervenant » ?
  - ...

Pour les normes :

1. HMM : le mot
2. CRF :
  - mot
  - lemme
  - rôle grammatical
  - le mot est-il un terme clé des normes ? (*article, code, loi, contrat, règlement, convention, décret*)
  - noms ou adjectifs voisins
  - ...

# Sélection des caractéristiques d'une catégorie de décisions

## Notations (Pour le corpus d'entraînement)

$w$  : un terme

$t$  : un texte

$L_w$  : longueur de  $w$  (nombre de mots)

$c$  : la classe cible ou positive (catégorie de demande)

$\bar{c}$  : la classe complémentaire ou négative (inconnue)

$N_c$  et  $N_{\bar{c}}$  : resp. nombre de textes de  $c$  et de  $\bar{c}$

$N_{w,c}$  : nombre de textes de  $c$  contenant  $w$

$N$  : nombre total de textes dans le corpus ( $N = N_c + N_{\bar{c}}$ )

$DF_c$  : proportion de textes du corpus appartenant à  $c$  ( $\mathbb{P}(c)$  : probabilité qu'un texte pris au hasard soit de la classe  $c$ )

$DF_w$  : proportion de documents du corpus contenant  $w$  ("Document frequency")

$DF_{w|c}$  : proportion de documents de  $c$  contenant  $w$

$DF_{c|w}$  : proportion de documents contenant  $w$  qui appartiennent à  $c$  ( $\mathbb{P}(c|w)=$ )

$Occ_{w,t}$  : nombre d'occurrences de  $w$  dans  $t$

$Occ_t$  : somme des nombres d'occurrences des termes dans  $t$

$TF_{w,t}$  : fréquence d'observation de  $w$  dans le texte  $t$

$SI_w$  : score d'importance de  $w$  pour  $c$

# Sélection des caractéristiques d'une catégorie de décisions

$$\Delta_{DF}(w, c) = DF_{w,c} - DF_{w,\bar{c}}$$

$$\chi^2(w, c) = \frac{N((N_{w,c}N_{\bar{w},\bar{c}}) - (N_{w,\bar{c}}N_{\bar{w},c}))^2}{N_w N_{\bar{w}} N_c N_{\bar{c}}}$$

$$n\,gl(w, c) = \frac{\sqrt{N}((N_{w,c}N_{\bar{w},\bar{c}}) - (N_{w,\bar{c}}N_{\bar{w},c}))}{\sqrt{N_w N_{\bar{w}} N_c N_{\bar{c}}}}$$

$$gss(w, c) = (N_{w,c}N_{\bar{w},\bar{c}}) - (N_{w,\bar{c}}N_{\bar{w},c})$$

Marascuilo :

$$M(w, c) = \frac{(N_{w,c} - N_w N_c / N)^2 + (N_{w,\bar{c}} - N_w N_{\bar{c}} / N)^2 + (N_{\bar{w},c} - N_c N_{\bar{w}} / N)^2 + (N_{\bar{w},\bar{c}} - N_{\bar{w}} N_{\bar{c}} / N)^2}{N}$$

# Méthode de pondération locale

Méthode	Formule
Fréquence de $w^*$	$TF_{w^*,t} = \frac{Occ_{w^*,t}}{\sum_{w_i \in W} Occ_{w_i,t}}$
Présence de $w^*$	$TP_{w^*,t} = \begin{cases} 1 & \text{si } TF_{w^*,t} > 0 \\ 0 & \text{sinon} \end{cases}$
Fréquence augmentée de $w^*$	$ATF_{w^*,t} = k + (1 - k) \frac{TF_{w^*,t}}{\max_{w_i \in W} TF_{w_i,t}}$
Logarithme de la fréquence de $w^*$	$LogTF_{w^*,t} = \log(TF_{w^*,t} + 1)$

TABLE – Méthodes de pondération locale