

# Méthodes d'analyse sémantique de corpus de décisions jurisprudentielles

Soutenance de thèse

---

présentée par Gildas TAGNY NGOMPÉ

le 24 janvier 2020

devant le jury composé de:

- Stéphane MUSSARD, Professeur, Université de Nîmes (Directeur de thèse)
- Jacky MONTMAIN, Professeur, IMT Mines Alès (Co-directeur de thèse)
- Sandra BRINGAY, Professeur, Université Paul Valéry Montpellier (Rapporteur)
- Mohand BOUGHANEM, Professeur, Université Toulouse III Paul Sabatier (Rapporteur)
- Françoise SEYTE, Maître de Conférences (HDR), Université de Montpellier (Examineur)
- Fabrice MUHLENBACH, Maître de Conférences, Université Jean Monnet de Saint-Étienne (Examineur)
- Guillaume ZAMBRANO, Maître de Conférences, Université de Nîmes (Encadrant de proximité)
- Sébastien HARISPE, Maître Assistant, IMT Mines Alès (Encadrant de proximité)



1. Introduction
2. Annotation des sections et entités judiciaires
3. Identification des demandes
4. Identification du sens du résultat
5. Découverte des circonstances factuelles
6. Conclusions

# 1. Introduction

## 1.1 Contexte

## 1.2 État de l'art

## 1.3 Objectif de la thèse

## 2. Annotation des sections et entités judiciaires

## 3. Identification des demandes

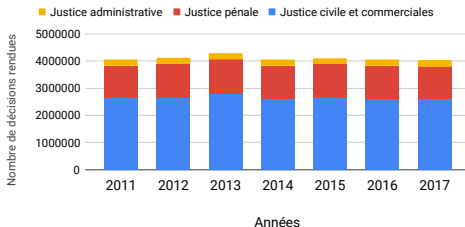
## 4. Identification du sens du résultat

## 5. Découverte des circonstances factuelles

## 6. Conclusions

- Utilité de la jurisprudence pour les juristes
  - elle est analysée pour comprendre l'application de la loi
- Motivation : limite de l'analyse manuelle

## 1. Gros volume de décisions



## 2. Limite des moteurs de recherche juridique :

- Pas de critère de recherche sémantique (catégorie de demande, type de faits, etc.)
- pas d'analyse synthétique de corpus

# Activités en analyse automatique de décisions judiciaires

- Extraction d'information dans les décisions
  - entités juridiques [Waltl et al., 2016, Andrew and Tannier, 2018]
  - faits [Wyner, 2010, Wyner and Peters, 2010, Shulayeva et al., 2017]
  - définitions de concept juridiques [Waltl et al., 2016, Waltl et al., 2017]
  - arguments [Moens et al., 2007]
- Classification de décisions
  - Prédiction des décisions de justice [Ashley and Brüninghaus, 2009, Aletras et al., 2016]
  - identification de la formation et la période [Şulea et al., 2017b, Şulea et al., 2017a]
  - identifier la sentence prononcée (Chine) [Ma et al., 2018]
- Similarité entre décisions
  - décisions qui citent les mêmes lois et précédents [Nair and Wagh, 2018]
  - recherche d'affaires antérieures pertinentes [Thenmozhi et al., 2017]
  - identifier la sentence prononcée (Chine) [Ma et al., 2018]
  - similarité basée sur la question discutée et les faits sous-jacents (Inde) [Kumar et al., 2011]
  - regroupement non-supervisé [Ravi Kumar and Raghuveer, 2012]

# Objectif de la thèse

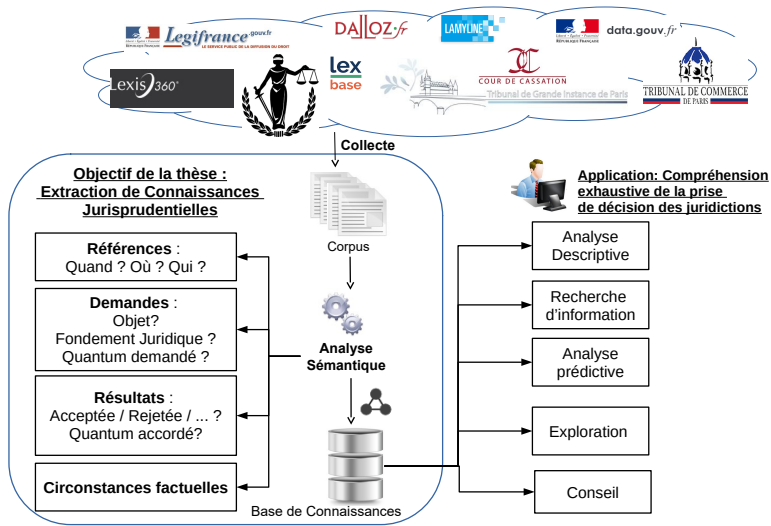


FIGURE – Objectifs et exemples d'application de la thèse.

# Positionnement en fouille de texte

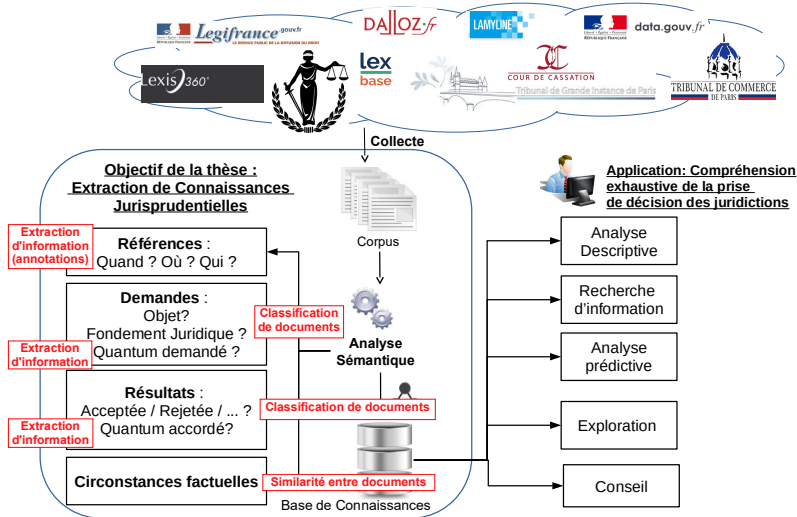


FIGURE – Tâches abordées en analyse de données textuelles.

# Difficultés rencontrées par l'automatisation de ces tâches

- Les décisions sont des textes non-structurés
- Le langage juridique est complexe

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.



## 1. Introduction

## 2. Annotation des sections et entités judiciaires

### 2.1 Objectif

### 2.2 Approches probabilistes de détection d'entités

### 2.3 Sélection de modèle

### 2.4 Discussions

## 3. Identification des demandes

## 4. Identification du sens du résultat

## 5. Découverte des circonstances factuelles

## 6. Conclusions

# Références dans l'entête, normes dans le reste

Cour d'appel  
Lyon  
6e chambre  
17 Mars 2016  
Répertoire Général : 14/06777  
APPELANTE :  
Mme Monique V. ...  
Représentée par Me Chrystelle P. , avocat au ...  
INTIMES :  
Mme Sylvianne C. ...  
Composition de la Cour ... :  
- Claude VIEILLARD , président ...

FAITS, PROCÉDURE, MOYENS ET ...  
Suite à un prêt de 10.000 € ...  
Par jugement en date du 4 avril 2013, ...  
Dans leurs conclusions ..., Mme Sylvianne C. , M. ...  
demandent à la cour de :  
- condamner Mme V. à leur payer ... au titre de l'article 700 du  
code ... , ...

MOTIFS DE LA DÉCISION  
La cour constate au préalable que le jugement n' est pas remis  
en causes ...  
...  
La Cour estime par contre que ... application  
de l'article 700 du code de procédure civile en cause d' appel  
au profit des  
intimés et il convient de leur allouer à ce titre la somme de 1.000  
€ .

PAR CES MOTIFS  
La Cour , ...

▼ Original markups

- ☒ appelant
- ☒ avocat
- ☐ corps
- ☒ date
- ☐ decision
- ☐ dispositif
- ☐ entete
- ☒ fonction
- ☒ formation
- ☒ intime
- ☒ juge
- ☒ juridiction
- ☐ litige
- ☐ motifs
- ☒ norme
- ☒ rg
- ☒ ville

# 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, arguments 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.

...

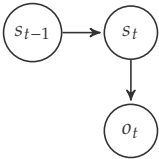
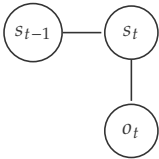
Vu l'article 700 du Code de Procédure Civile,  
Condamne Madame JONES-B. à verser à Ma-  
dame A. la somme de 2.500 euros.

**Dispositif** : résultats et normes

**Entêtes** : méta-données

# Approches probabilistes de détection d'entités

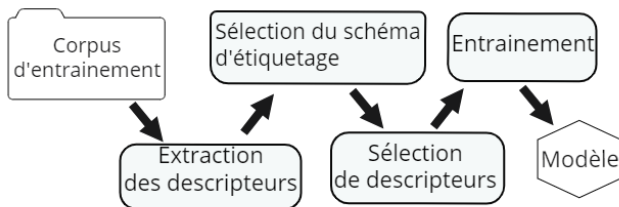
## Modèles probabilistes à états et observations

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>[Seymore et al., 1999]</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>[Peng and McCallum, 2006]</p>

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

Entraînement sur des séquences préalablement étiquetées

# Sélection de modèle



- sélection du schéma d'étiquetage
- sélection des descripteurs

# Confusions de classes

---

# Nombre nécessaire de données d'entraînement

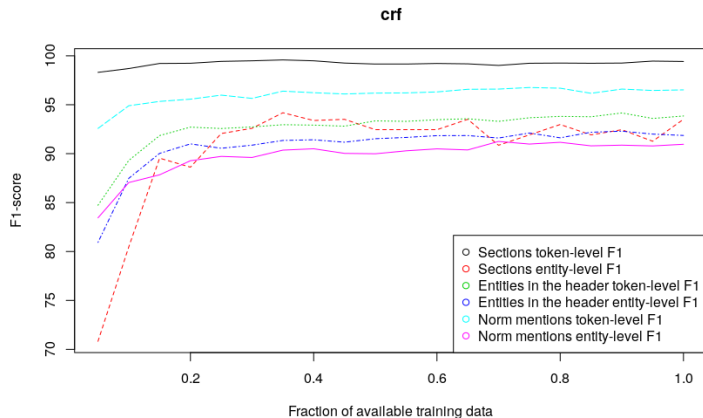


FIGURE – Résultats en fonction du nombre de données d'entraînement (fractions d'environ 380 décisions)



# Description manuelle vs. représentation apprise

---

1. Introduction

2. Annotation des sections et entités judiciaires

3. Identification des demandes

3.1 Objectif : identifier les informations sur les demandes

3.2 Méthode : identifier les passages, puis les informations

3.3 Expérimentations sur 6 catégories de demandes

4. Identification du sens du résultat

5. Découverte des circonstances factuelles

6. Conclusions

# Objectif : identifier les informations sur les demandes

Exemple : dommage-intérêts pour procédure abusive (danaïs)

Jennifer M. et Catherine M. ... demandent à la Cour de :

- **infirmen** le dit jugement en toutes ses dispositions ; ...

Statuant à nouveau ...

- les condamner au paiement d' une somme de 3 000,00 € pour procédure abusive et aux entiers dépens ; ...

La cour ... CONFIRME le jugement entrepris en toutes ses dispositions.

IDENTIFICATION DE LA DECISION			DESCRIPTION DE LA PRETENTION			DESCRIPTION DU RESULTAT	
Type	Ressort	RG	OBJET	NORME	QUANTUM	RESULTAT	QUANTUM RESULTAT (obtenu)
CA	Saint Denis	14/01082	dommages-intérêts	1382 code civil + 32-1 code de procédure civile : en procédure abusve	3.000.00 €	rejette	0.00 €

## Difficultés

- Présence de plusieurs demandes de catégories similaires et/ou différentes dans une même décision
- Toutes les catégories ne sont pas connues d'avance (+500 catégories)
- Difficile d'annoter une base d'évaluation pour toutes les couvrir
- Enoncés non structurés, avec des références, et des agrégations

# Retrouver les demandes à l'aide des termes clés

1. Détermination automatique de la terminologie (déclencheurs) de la catégorie

$$n_{gl}(t, c) = \frac{\sqrt{N}(N_{t,c}N_{\bar{t},\bar{c}}) - (N_{t,\bar{c}}N_{\bar{t},c})}{\sqrt{N_t N_{\bar{t}} |D_c| |D_{\bar{c}}|}}.$$

2. Détection de la présence de la catégorie par classification de la décision ( $c$  vs.  $\bar{c}$ )
3. Identification des passages de demandes et résultats
4. Exploiter la proximité entre les déclencheurs de la catégorie et sommes d'argent pour extraire les quanta :

Section Litige : identification de la demande

Jennifer M. et Catherine M. ... demandent à la Cour de :

- infirmer le dit jugement en toutes ses dispositions ; ...

Statuant à nouveau ...

- [ les condamner au paiement d' une somme de 3 000,00 € pour  
**procédure abusive** et aux entiers dépens ; l<sub>demande\_danais</sub>

...

Section Dispositif : identification du résultat

La cour ...

CONFIRME le jugement entrepris en toutes ses dispositions.

5. Mise en correspondance des informations relatives à la même demande

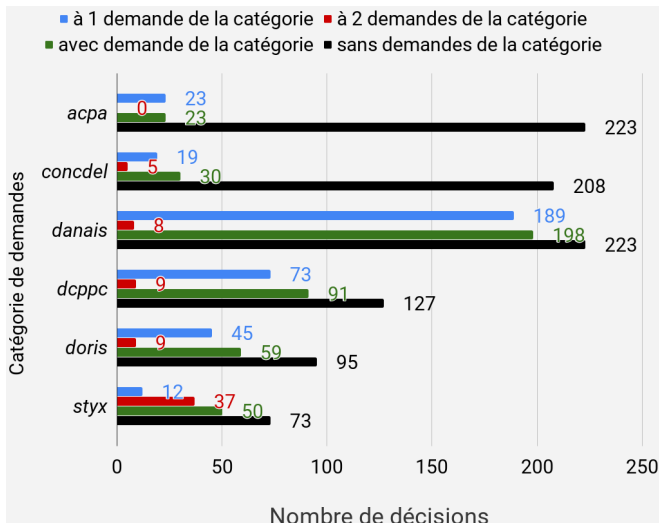
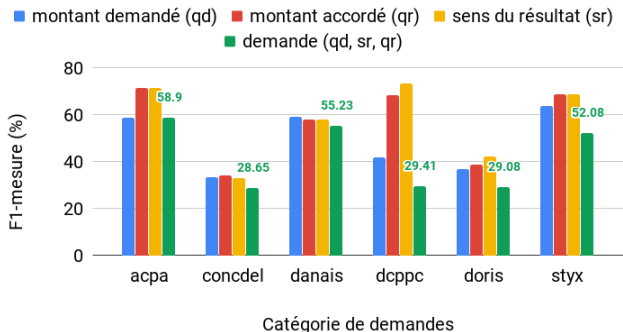


FIGURE — Répartitions des demandes dans les documents annotés.

# Résultats de l'extraction

- Détection de catégorie facile par des classifieurs traditionnels (K-plus-proches-voisins, SVM, Bayésien naïf, Arbre) :  $98.8\% \leq F_1\text{-mesure} \leq 100\%$
- Résultat plus facile à extraire que le montant



- Source d'erreurs :
  - Sélection difficile de déclencheurs rares
  - Non exploitation des références aux jugements antérieurs
  - Certains quanta sont absents des sections Litige et Dispositif
  - Mauvaise méthode de mise en correspondance

1. Introduction

2. Annotation des sections et entités judiciaires

3. Identification des demandes

4. Identification du sens du résultat

4.1 Restriction et motivations

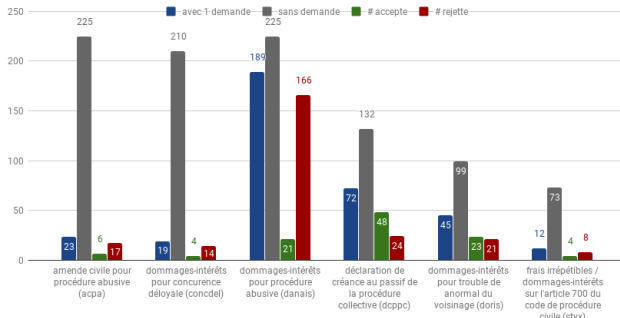
4.2 Synthèse bibliographique

5. Découverte des circonstances factuelles

6. Conclusions

# Restriction et motivation

- Uniquement les décisions à une demande de la catégorie considérée
  - Raison : plus de 50% des documents dans la majorité des catégories
- Classification binaire
  - Raison : le sens d'un résultat est pratiquement toujs une de ces deux valeurs : **accepte** ou **rejette**





$x^{(k)} = f^{(k)}$  vecteur initial des caractéristiques du texte  $k$

$r = \log \left( \frac{p/\|p\|_1}{q/\|q\|_1} \right)$ , vecteur poids du classifieur bayésien multinomial

avec  $p = \alpha + \sum_{i:y^{(i)}=1} f^{(i)}$ ,  $q = \alpha + \sum_{i:y^{(i)}=-1} f^{(i)}$

L'idée : transformer les caractéristiques réduites à leur simple

présence  $\hat{f}^{(k)}$  avec  $r$  ( $\tilde{f}^{(k)} = \hat{r} \circ \hat{f}^{(k)}$ )

$\hat{r}$  est calculé avec  $\hat{f}^{(k)}$

les nouveaux  $x^{(k)} = \tilde{f}^{(k)}$  sont utilisés dans un SVM.

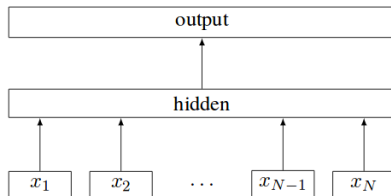


FIGURE — Architecture similaire au model CBOW : le label remplace le mot au milieu.

$$\text{Entraînement : } \min \left( -\frac{1}{N} y_n \cdot \sum_{n=1}^N y_n \cdot \log f(B \cdot A \cdot x_n) \right)$$

$$\text{où } f \text{ est la fonction softmax } f(z) = \left[ \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \right]_{\forall j \in \{1, \dots, K\}}$$

# Résultats obtenus avec fastText et NBSVM

## Influence du déséquilibre et de la (très) faible taille des données

Cat. Dmd.	Algo.	Préc.	Préc. équi.	err-0	err-1	f1-0	f1-1	f1-macro-a
dcppc	nbsvm	0.875	0.812	0.375	0	0.752	0.916	<b>0.834</b>
danais	fasttext	<b>0.888</b>	0.5	1	0	0	0.941	0.47
danais	nbsvm	0.888	0.5	0	1	0.941	0	0.47
concdel	fasttext	0.775	0.5	1	0	0	0.873	0.437
concdel	nbsvm	0.775	0.5	0	1	0.873	0	0.437
acpa	fasttext	0.745	0.5	1	0	0	0.853	0.426
acpa	nbsvm	0.745	0.5	0	1	0.853	0	0.426
doris	nbsvm	0.5	0.492	0.85	0.167	0.174	0.63	0.402
dcppc	fasttext	0.667	0.5	0	1	0.8	0	0.4
styx	fasttext	0.667	0.5	1	0	0	0.8	0.4
styx	nbsvm	0.667	0.5	0	1	0.8	0	0.4
doris	fasttext	0.523	0.5	0	1	0.686	0	0.343

0 == accepte

1 == rejette

# Application des extensions de la Régression PLS (1)

PLS standard (Régression partielle des moindres carrés)

Réduction supervisée des dimensions  $x_1, x_2, \dots, x_p$  en  
composantes orthogonales  $t_1, \dots, t_h$

$$t_h = w_{h1}x_1 + \dots + w_{hj}x_j + \dots + w_{hp}x_p$$

$$\text{avec } w_{hj} = \frac{\text{cov}(u_{(h-1)j}, \epsilon_h)}{\sqrt{\sum_{p=1}^j \text{cov}^2(u_{(h-1)j}, \epsilon_h)}}, \quad y = c_1t_1 + \dots + c_h t_h + \epsilon_h,$$

$$\text{et } x_j = \beta_{1j}t_1 + \dots + \beta_{hj}t_h + u_{(h-1)j}$$

# Application des extensions de la Régression PLS (2)

1. Gini-PLS : élimination de la sensibilité au *outliers* en remplaçant la covariance  $cov(x_j, y)$  par la covariance de Gini  $cog(y; x_j) := cov(y; R(x_j))$  pour l'estimation des résidus  $u_{(h)j}$  et des poids  $w_{hj}$   
[Mussard and Souissi-Benrejab, 2018]
2. Logit-PLS :  $\forall j > 1$ , les  $w_{hj}$  sont les coefficients de la régression logistique de  $y$  sur les composantes  $t_1, \dots, t_{h-1}, u_{(h-1)j}$  [Tenenhaus, 2005]
3. Gini-Logit-PLS : covariance Gini pour  $u_{(h)j}$  et coefficient Logit pour les  $w_{hj}$

# Résultats : meilleures configurations

Vecteur	classifieur	F1	min	Cat. min	max	Cat. max	$F1 - 1^{er} F1$	max - min	ra
GSS*TF	Tree	<b>0.668</b>	0.5	doris	0.92	dcppc	<b>0</b>	<b>0.42</b>	1
AVG-G*TF	LogitPLS	<b>0.648</b>	0.518	danais	0.781	dcppc	<b>0.02</b>	<b>0.263</b>	1
AVG-G*TF	StandardPLS	<b>0.636</b>	0.49	danais	0.836	dcppc	<b>0.032</b>	<b>0.346</b>	2
DELTADF*TF	GiniPLS	<b>0.586</b>	0.411	danais	0.837	dcppc	<b>0.082</b>	<b>0.426</b>	10
DELTADF*TF	GiniLogitPLS	<b>0.578</b>	0.225	styx	0.772	dcppc	<b>0.09</b>	<b>0.547</b>	2

AVG-G == Moyenne des métriques globales de pondération

En moyenne, la meilleure zone est la partie principale (litige\_motifs\_dispositif)

Les extensions du PLS ne sont pas très éloignées (si on choisi le bon schéma de vectorisation)

# Résultats pour chaque classe

Cat. Dmd	zone	Vecteur	classifieur	F1
acpa	demande_resultat_a_resultat_context	DBIDF*TF	Tree	0.846
acpa	litige_motifs_dispositif	DELTADF*TF	StandardPLS	0.697
acpa	litige_motifs_dispositif	AVERAGEGlobals*TF	LogitPLS	0.683
concdel	litige_motifs_dispositif	GSS*TF	Tree	0.798
concdel	motifs	IDF*TF	GiniLogitPLS	0.703
concdel	context	DBIDF*LOGAVE	StandardPLS	0.657
danais	demande_resultat_a_resultat_context	CHI2*AVERAGELocals	Tree	0.813
danais	demande_resultat_a_resultat_context	AVERAGEGlobals*ATF	LogitPLS	0.721
danais	demande_resultat_a_resultat_context	AVERAGEGlobals*ATF	StandardPLS	0.695
dcppc	demande_resultat_a_resultat_context	CHI2*TF	Tree	0.985
dcppc	demande_resultat_a_resultat_context	CHI2*TF	LogitPLS	0.94
dcppc	litige_motifs_dispositif	MARASCUILO*TP	StandardPLS	0.934
doris	litige_motifs_dispositif	DSIDF*TP	GiniPLS	0.806
doris	litige_motifs_dispositif	DSIDF*TP	GiniLogitPLS	0.806
doris	litige_motifs_dispositif	IG*ATF	StandardPLS	0.772
styx	motifs	DSIDF*TF	Tree	1
styx	demande_resultat_a_resultat_context	DSIDF*LOGAVE	GiniLogitPLS	0.917
styx	litige_motifs_dispositif	RF*TF	GiniPLS	0.833

De bonnes performances si on varie les métaparamètres en fonction de la catégorie de demande

1. Introduction
2. Annotation des sections et entités judiciaires
3. Identification des demandes
4. Identification du sens du résultat
5. Découverte des circonstances factuelles
  - 5.1 Objectifs
  - 5.2 Méthode : apprentissage d'une distance et utilisation pour du regroupement
  - 5.3 Sélection de la représentation des décisions
  - 5.4 Efficacité du regroupement



- Déterminer les situations distinctes où sont formulées les demandes d'une catégorie données.

## Catégorie : action en responsabilité civile professionnelle contre les avocats (arcpa)

- cas  $a$  : un avocat négligent qui envoie son assignation de manière tardive ;
  - cas  $b$  : un avocat qui n'a pas donné un conseil opportun, qui n'a pas soulevé le bon argument ;
  - cas  $c$  : un avocat qui n'a pas rédigé un acte valide ou réussi à obtenir un avantage fiscal ;
  - cas  $d$  : un avocat attaqué par son adversaire et non par son propre client.
- Formulation comme regroupement non supervisé des décisions

- Apprentissage d'une distance basé sur la transformation
  - Formulation de la distance pour un ensemble de modifications connues

$$Dis_M(d, d') = f(M_{(d, d')}) = \frac{\sum_{(d[k], d'[k]) \in M_{(d, d')}} Dis_{cos}(\overrightarrow{d[k]}, \overrightarrow{d'[k]})}{|d|}$$

- Génération d'un corpus d'entraînement
$$B_M = \{((d_1, d_2), Dis(d_1, d_2))_i\}_{1 \leq i \leq |B_M|}$$
- Entraînement d'un modèle de régression pour prédire la distance entre deux documents

$$Dis_M(d_i, d_j) = Reg_M(\vec{d}_i - \vec{d}_j)$$

- Utilisation de la distance dans un algorithmes de regroupement (K-moyennes et K-medoides)

# Sélection de la représentation : objectif

Trouver la représentation qui discrimine les cas sur leur champ sémantique

Corpus	Terminologie
<i>arcpa</i>	chance, perte chance, avocat, perte, diligence, chance obtenir, perdre, client, devoir conseil, manquement
<i>cas a</i>	chance, perte chance, chance succès, perte, client, préjudice indemnisable, article code commerce, indemnisable, condamnation emporter, emporter nécessairement rejet
<i>cas b</i>	défense intérêt, intérêt client, avocat, contractuel égard, responsabilité contractuel droit, responsabilité professionnel avocat, contractuel droit commun, assurer défense intérêt, civil avocat, grief articuler
<i>cas c</i>	rédacteur acte, rédacteur, avocat rédacteur acte, avocat rédacteur, qualité rédacteur acte, rédaction acte, qualité rédacteur, projet acte, prendre initiative conseiller, initiative conseiller
<i>cas d</i>	revêtir aucun, revêtir aucun caractère, article code, article code procédure, faire référence aucun, fautif madame, civil profit autre, civil depuis, mention expresse, moyen dont

TABLE – Terminologies de la catégorie *arcpa* et de ses cas

# Sélection de la représentation : résultats

Distance	Base <sup>a</sup>	Silhouette optimale (pondération, réduction, dim.)
$Dis_{jaccard}$	0.001	0.212 (TP-NGL, FNM, 4)
$Dis_{cos}$	0.002	0.202 (TP-NGL, FNM, 4)
$Dis_M$	-0.049	0.195 (TP-NGL, FNM, 4)
$Dis_{braycurtis}$	0.002	0.182 (TP-NGL, FNM, 4)
$Dis_{euclidienne}$	0.001	0.168 (TP-NGL, FNM, 4)
$Dis_{manhattan}$	-0.019	0.17 (TP-NGL, FNM, 4)
$Dis_{pearson}$	0.014	0.057 (TP-CHI2, aucune, 19763)
$Dis_{wmd}$	-0.096	-

<sup>a</sup> occurrence de mots pour  $Dis_{wmd}$ , et TF-IDF pour les autres distances.

TABLE – Meilleures représentations sur la catégorisation manuelle.

# Regroupement pour la catégorie annotée

Distance	Algorithme	K	Silhouette	ARI	NMI	R	P	$F_1$
$Dis_M$	K-moyennes	3	0.438	<b>0.407</b>	<b>0.423</b>	0.552	0.654	<b>0.599</b>
$Dis_M$	K-medoïdes	6	0.453	0.359	0.395	0.298	0.669	0.413
$Dis_{braycurtis}$	K-moyennes	4	0.473	0.383	0.407	0.446	0.658	0.532
$Dis_{braycurtis}$	K-medoïdes	5	0.448	0.344	0.375	0.331	0.645	0.437
$Dis_{cosine}$	K-moyennes	4	0.528	0.383	0.407	0.446	0.658	0.532
$Dis_{cosine}$	K-medoïdes	4	0.526	<b>0.398</b>	<b>0.421</b>	0.464	0.680	<b>0.551</b>
$Dis_{euclidean}$	K-moyennes	5	0.478	0.365	0.395	0.341	0.670	0.452
$Dis_{euclidean}$	K-medoïdes	5	0.456	0.313	0.346	0.335	0.619	0.434
$Dis_{jaccard}$	K-moyennes	4	0.570	0.367	0.391	0.439	0.643	0.522
$Dis_{jaccard}$	K-medoïdes	4	<b>0.560</b>	0.389	0.412	0.451	0.666	0.538
$Dis_{manhattan}$	K-moyennes	4	0.482	0.376	0.400	0.452	0.657	0.535
$Dis_{manhattan}$	K-medoïdes	5	0.452	0.368	0.397	0.345	0.675	0.456
$Dis_{pearson}$	K-moyennes	2	<b>0.611</b>	0.054	0.072	0.746	0.453	0.564
$Dis_{pearson}$	K-medoïdes	2	0.171	0.152	0.166	0.598	0.482	0.534
$Dis_{wmd}$	K-medoïdes	2	0.332	-0.016	0.002	0.545	0.397	0.459

TABLE — Evaluation de la catégorisation par K-moyennes et K-medoïdes sur  $D_{arcpa}$  avec détermination du nombre de clusters basée sur la silhouette.

# Regroupement des catégories non annotées

$D_{doris}$ (59)	$Dis_M$	K-medoïdes	2	0.509
	$Dis_M$	K-moyennes	3	0.527
	$Dis_{cosine}$	K-medoïdes	5	0.549
	$Dis_{cosine}$	K-moyennes	4	0.586
	$Dis_{jaccard}$	K-medoïdes	3	0.600
	$Dis_{jaccard}$	K-moyennes	4	0.645

TABLE — Evaluation non-supervisée des K-moyennes et K-medoïdes sur  $D_{doris}$ .

Cluster	Terminologie ( <i>ngl</i> )
0	excéder inconvenient, inconvenient normal, excéder inconvenient normal, normal voisinage, inconvenient normal voisinage, inconvenient, trouble excéder inconvenient, trouble excéder, excéder, normal
1	copropriétaire, syndicat copropriétaire, syndicat, condamner in, anormal voisinage, trouble anormal voisinage, in, trouble anormal, syndic, jouissance subir
2	deux fond   fonds, séparatif deux fond   fonds, limite séparatif deux, ordonner démolition, séparatif deux, implanter, condamner démolir, devoir établir toit, devoir établir, toit manière
3	manière plus, chose manière plus, chose manière, usage prohiber loi, prohiber loi règlement, prohiber loi, absolu, usage prohiber, manière plus absolu, plus absolu
4	situer zone, hauteur @card@ mètre, hauteur dépasser, appel contester, vitrer, dont hauteur dépasser, urbaniser, recevabilité <unknown> appel, cahier charge lotissement, charge lotissement

TABLE — Terminologies des circonstances factuelles découvertes en combinant les K-medoïdes et la distance cosinus sur  $D_{doris}$ .

1. Formulation comme problème de regroupement non supervisé de décisions de la catégorie
2. Méthode d'apprentissage d'une distance de dis-similarité au sein d'une catégorie
3. Sélection de la représentation des textes qui reflète la notion subjective de similarité de l'expert
4. Expérimentation des propositions sur 7 catégories de demandes dont 1 annotées

1. Introduction
2. Annotation des sections et entités judiciaires
3. Identification des demandes
4. Identification du sens du résultat
5. Découverte des circonstances factuelles
6. Conclusions
  - 6.1 Bilan
  - 6.2 Perspectives



## Contributions

- Etude de l'application du HMM et CRF pour détecter les sections et les entités juridiques
- Approche d'identification des demandes basée sur la proximité entre les termes-clés et les sommes d'argent
- Extensions du Gini-PLS pour identifier le sens du résultat
- Approche d'apprentissage d'une distance de similarité pour regrouper les décisions suivant les circonstances factuelles.

## Limites

- Évaluation sur de faibles quantité de données annotées ;
- Non expérimentation de méthodes récentes (réseaux de neurones)

# Conclusions : perspectives

## Amélioration des propositions

- **Désambiguïser les entités** détectées pour indexer les décisions
- Expérimentation des approches récentes pour l'identification des **demandes formalisées comme relation entre montant demandé et montant accordé**
- Découverte des circonstances factuelles vue comme **modélisation thématique**

## Applications

- **Anonymisation des décisions** : confidentialité des informations
- **Analyse prédictive** : identifier les raisons qui poussent les juges à accepter une demande

## Questions

# References I



Aletras, N., Tsarapatsanis, D., Preoțiuc-Pietro, D., and Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights : A Natural Language Processing perspective.  
*PeerJ Computer Science*, 2 :e93.



Andrew, J. J. and Tannier, X. (2018). Automatic Extraction of Entities and Relation from Legal Documents.  
*In Proceedings of the Seventh Named Entities Workshop*, pages 1–8.



Ashley, K. D. and Brüninghaus, S. (2009). Automatically classifying case texts and predicting outcomes.  
*Artificial Intelligence and Law*, 17(2) :125–165.



Grave, E., Mikolov, T., Joulin, A., and Bojanowski, P. (2017). Bag of tricks for efficient text classification.  
*In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 427–431, Valencia, Spain.



Kumar, S., Reddy, P. K., Reddy, V. B., and Singh, A. (2011). Similarity analysis of legal judgments.  
*In Proceedings of Compute 2011 - Fourth Annual ACM Bangalore Conference*, page 17. ACM.



Ma, Y., Zhang, P., and Ma, J. (2018). An Efficient Approach to Learning Chinese Judgment Document Similarity Based on Knowledge Summarization.  
arXiv preprint arXiv :1808.01843 [cs.AI].

# References II



Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007).

Automatic detection of arguments in legal texts.

In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM.



Mussard, S. and Souissi-Benrejab, F. (2018).

Gini-PLS Regressions.

*Journal of Quantitative Economics*, pages 1–36.



Nair, A. M. and Wagh, R. S. (2018).

Similarity Analysis of Court Judgements Using Association Rule Mining on Case Citation Data - A Case Study.

*International Journal of Engineering Research and Technology*, 11(3) :373–381.



Peng, F. and McCallum, A. (2006).

Information extraction from research papers using conditional random fields.

*Information processing & management*, 42(4) :963–979.



Ravi Kumar, V. and Raghuvver, K. (2012).

Legal documents clustering using latent dirichlet allocation.

*International Journal of Applied Information Systems (IJAIS)*, 2(6) :34–37.



Seymore, K., McCallum, A., and Rosenfeld, R. (1999).

Learning hidden Markov model structure for information extraction.

*AAAI-99 workshop on machine learning for information extraction*.



Shulayeva, O., Siddharthan, A., and Wyner, A. (2017).

Recognizing cited facts and principles in legal judgements.

*Artificial Intelligence and Law*, 25(1) :107–126.

# References III



Șulea, O.-M., Zampieri, M., Malmasi, S., Vela, M., P. Dinu, L., and van Genabith, J. (2017a). Exploring the Use of Text Classification in the Legal Domain.

In *Proceedings of 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts*, page 5, London, United Kingdom. ASAIL'2017.



Șulea, O.-M., Zampieri, M., Vela, M., and van Genabith, J. (2017b). Predicting the Law Area and Decisions of French Supreme Court Cases.

In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722.



Tenenhaus, M. (2005).

La regression logistique PLS.

In Driesbeke, Jean-Jacques and Lejeune, Michel and Saporta, Gilbert, editor, *Modèles statistiques pour données qualitatives*, chapter 12, pages 263–276. Editions Technip.



Thenmozhi, D., Kannan, K., and Aravindan, C. (2017).

A Text Similarity Approach for Precedence Retrieval from Legal Documents.

In *Proceedings of Forum for Information Retrieval Evaluation - FIRE (Working Notes)*, pages 90–91.



Waltl, B., Landthaler, J., Scepankova, E., Matthes, F., Geiger, T., Stocker, C., and Schneider, C. (2017).

Automated extraction of semantic information from German legal documents.

In *IRIS : Internationales Rechtsinformatik Symposium. Association for Computational Linguistics*.



Waltl, B., Matthes, F., Waltl, T., and Grass, T. (2016).

LEXIA - A Data Science Environment for Semantic Analysis of German Legal Texts.

In *IRIS : Internationales Rechtsinformatik Symposium*.

Salzburg, Austria.

# References IV



Wang, S. and Manning, C. D. (2012).

Baselines and bigrams : Simple, good sentiment and topic classification.

In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.



Wyner, A. and Peters, W. (2010).

Lexical Semantics and Expert Legal Knowledge towards the Identification of Legal Case Factors.

In *JURIX*, volume 10, pages 127–136.



Wyner, A. Z. (2010).

Towards annotating and extracting textual legal case elements.

*Informatica e Diritto : special issue on legal ontologies and artificial intelligent techniques*, 19(1-2) :9–18.