

# Analyse sémantique d'un corpus exhaustif de décisions jurisprudentielles

Comité de suivi individuel – 15 juin 2018

---

Gildas Tagny Ngompé

Début de thèse: 15 Décembre 2015

## Direction de thèse:

- Jacky Montmain (École des mines d'Alès, LGI2P)
- Stéphane Mussard (Université de Nîmes, CHROME)

## Encadrement de proximité:

- Sébastien Harispe (Ecole des Mines d'Alès, LGI2P)
- Guillaume Zambrano (Université de Nîmes, CHROME)

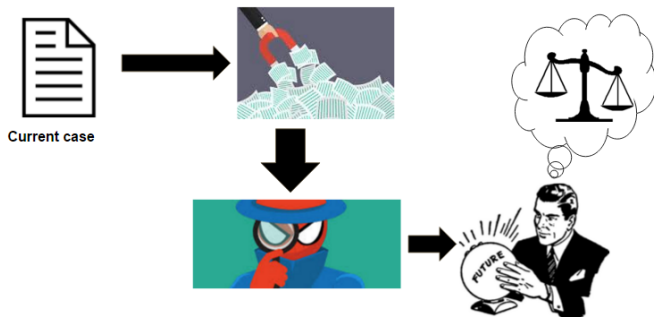


1. Motivations et objectifs
2. Extraction des demandes par pondération de termes clef et zonage
3. Classification binaire pour identifier le sens du résultat
4. Conclusion
5. Questions ?

## Motivations et objectifs

---

# Les juristes analysent les décisions



## Pquoi ?

- comprendre et comparer l'application loi (contentieux, ville, ...)
- estimer le risque judiciaire
- ...

# Motivation : documents non-structurés, langage complexe

ARRÊT N°

R.G : 11/03924

...

C 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 C 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 C :

...

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.

# Motivation : grand volume de décisions

Plus de 4 millions de décisions prononcées / an

	2010	2011	2012	2013	2014
<b>Justice civile</b>	2 673 131	2 654 179	2 647 813	2 761 554	2 618 374
Justice pénale	1 173 242	1 180 586	1 251 979	1 303 469	1 203 339
Justice administrative	224 787	225 608	228 680	221 882	230 477

Source : <http://www.justice.gouv.fr/budget-et-statistiques-10054/chiffres-cles-de-la-justice-10303/>

TABLE — Nombre de décisions prononcées en France par an

# Motivation : recherches et analyses sémantiques difficiles

## Moteurs de recherche juridique à mots-clés

### Pas d'analyse synthétique des décisions

☐ Recherche simple ☒ Recherche avancée

Mots ou expressions

Recherche

Ex : gérant **et** pouvoir, bail **s/5** résilt  
[Aide à la recherche](#)

Gestion automatique des :

☒ Singulier / Pluriel ☒ Masculin / Féminin

☐ Verbes conjugués **avoir** cherche **ayons**

Sources ☒ \*Toutes les sources

[Répertoire des sources](#)

ou

☐ Encyclopédies

☐ Codes et Lois

☐ JurisData

☐ Toute la jurisprudence

☐ Revues

☐ Bibliographies

☐ Actualités

☐ Bulletins Officiels

☐ Autorités administratives

☐ Parlement

☐ Europe

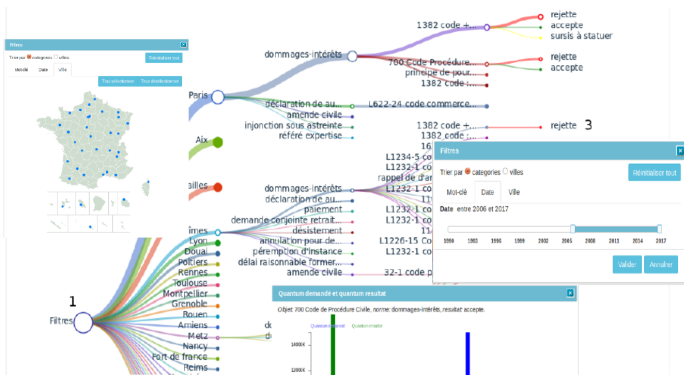
☐ Conventions Collectives

Période

See : LexisNexis.com

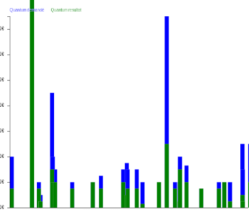
# Objectif : Moteur d'analyse sémantique

Stage été 2017 [ PRYSIAZHNIUK Anastasiia ]



2 Tableau de résultats

RG	Ville	Date	Jurisdiction	Quantum demandé	
15/02360	Metz	Thu, 24 Apr 2003 00:00:00 GMT	CA	-	2
15/00796	Metz	Thu, 21 Jul 2011 00:00:00 GMT	CA	5 000,00 €	
14/01658	Metz	Sun, 22 Jul 2012	CA	5 000,00 €	



Resultat	Description
accepte	Voir
rejette	Voir
rejette	Voir



# Problématiques d'extraction d'information



## Utilité

Recherche  
d'information

Prédiction

Exploration

Conseil



Corpus



Chaîne d'Analyse  
Sémantique



Base de Connaissances

## Objectif : Extraction d'Information

**Références :**  
Quand ? Où ? Qui ?

**Demandes :**  
Catégorie ?  
Montant demandé ?

**Résultats :**  
Accepte / rejette ?  
Montant obtenu ?  
Pourquoi ?

**Circonstances factuelles**

Extraction des demandes par pondération  
de termes clef et zonage

---

# Tâche : extraire les quanta et le sens du résultat

Expressions non structurées, par **référence**, par **agrégation**

## EXPRESSION DE DEMANDE ET RESULTAT

Jennifer M. , Catherine M. et Sandra M. ... demandent à la C 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 ; ...

La c, ... CONFIRME **le jugement entreprise** 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 €

TABLE — Informations à extraire (Dommages-intérêts pour procédure abusive)

# Décomposition du problème

Problème décomposé en 3 tâches :

1. Identification des catégories présentes dans le document
2. Détection des quanta demandés, quantas obtenus, et sens du résultat
3. Liaison des informations relatives à la même demande

# Problèmes similaires : objectif des tâches

- Extraction d'évènement :

Champs	[ACE, 2005]	Analogie chez les demandes
Type	Die	Catégorie="Dommages-intérêts pour procédure abusive"
Expression ( <i>extend</i> )	"Il est <b>mort</b> hier d'une insuffisance rénale."	(voir page précédente)
Déclencheur	"mort"	"procédure abusive"
Argument	Victim-Arg="il" Time-Arg="hier"	Quantum-demandé="3000€" Quantum-obtenu="0 €"
Attribut	Polarity=POSITIVE, Tense=PAST	Sens-résultat="Rejeté"

- Remplissage de champs des entités **Demande** (*slot-filling*) :  
**Catégorie, Quantum-demandé, Quantum-obtenu, Sens-résultat**
- Extraction d'entités et relations : par ex. (**quantum demandé, quantum obtenu**)

# Problèmes similaires : Approches

Type d'approches	Exemples
Chaîne de traitement	Chaîne de classifieurs [Ahn, 2006]
Modélisation probabiliste de la structure de l'évènement	Modèle joint d'inférence des entités, arguments, déclencheurs $p_{\theta}(t_i, r_i, a i, N_i, x)$ [Yang and Mitchell, 2016]
Réseau de neurones pour automatiser la génération des caractéristiques et la modélisation de la structure	(i) <b>Architecture multicouche de réseaux de neurones récurrents</b> : encodage de la phrase, encodage des contextes, prédiction du déclencheur, prédiction des rôles, mémoire matricielle d'interdépendance déclencheur-argument [Nguyen et al., 2016], (ii) <b>Réseau de pointeur</b> ( <i>pointer network</i> ) : un encodeur de la phrase et des contextes, plusieurs décodeurs (un pour chaque champ) [Palm et al., 2017]

TABLE — Type d'approches

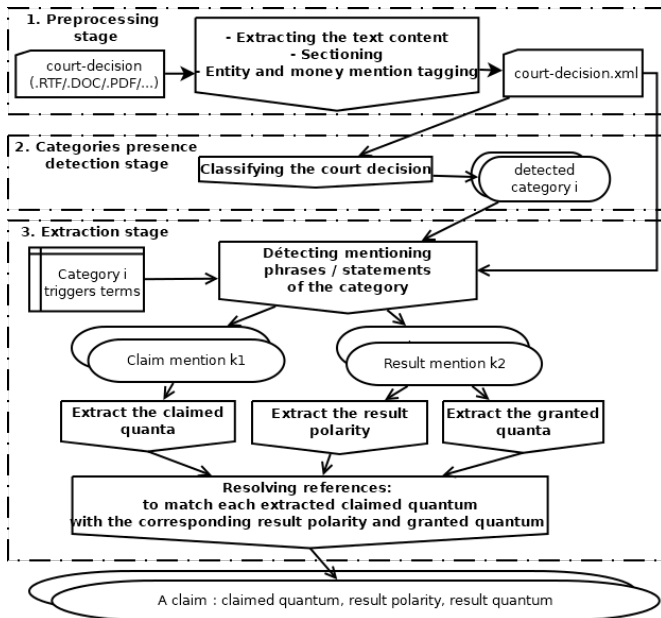
# Difficultés liées à l'extraction des demandes

- 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

Il faut une approche :

- qui s'adapte à la catégorie à extraire
- qui permette de rajouter de nouvelles catégories

# Architecture du pipeline d'extraction





# Identification des passages et des informations

Demande dans la section *Litige* (Faits, procédures, et moyens des parties)

Résultat dans la section *Dispositif*

Demande	Résultat (organisé par polarité)		
	accepte	sursis à statuer	rejette
<i>accorder, admettre, admission, allouer, condamnation, condamner, fixer, laisser, prononcer, ramener, surseoir</i>	<i>accorde, accordons, admet, admettons, alloue, allouons, condamne, condamnons, déclare, déclarons, fixe, fixons, laisse, laissons, prononce, prononçons</i>	<i>réserve, réservons, surseoit, sursoyons</i>	<i>déboute, déboutons, rejette, rejettons</i>

TABLE – Mots introduisant les énoncés de demandes et de résultats

- le <demande categorie="acpa">condamner à payer une <trigger categorie="acpa">**amende civile**</trigger> de <argent> **1.500 euros** </argent> pour procédure abusive ...
- le</demande> condamner à payer la somme ..."

# Identification des passages et des informations(2)

## ○ Identification des passages :

1. Soit par la seule **présence d'un trigger** : on zone aut des triggers
2. Soit par **pondération des zones à argent** :
  - 2.1 on zone aut des sommes d'argent
  - 2.2 on pondère les zones (par ex. somme des poids des triggers)
  - 2.3 on sélectionne une zone si elle a un poids  $\geq$  POIDS SEUIL

## ○ Identification des informations :

1. quantum : somme d'argent près d'un trigger
2. sens du résultat :
  - soit en fonction du verbe introductif de l'énoncé du résultat
  - soit "*rejette*" si pas d'énoncé du résultat

## ○ Résolution des références :

- matching des énoncés (similarité textuelle)
- matching des quanta (Hypothèse d'apparition dans le même ordre)

# Phase d'entraînement

Catégorie  $c_i$ , Corpus d'entraînement  $D = D_{c_i} \cup D_{\bar{c}_i} = \{D_j\}_{1 \leq j \leq |D|}$

## 1. Détecteur de catégorie :

- vectoriser les décisions de la base d'entraînement :  
 $w(t_k, D_j) = lw(t_k, D_j) \times gw(t_k) \times nf(D_j)$  [Salton and Buckley, 1988]
- entraîner un algorithme de classification (SVM, Naïf Bayésien, K plus proches voisins ...)

## 2. Extracteur de triplets de quanta et sens du résultat :

- Apprendre les triggers sur la base d'entraînement (passages à quanta vs. passages sans quanta)

2.1 pondération des termes  $t_k$  avec une métrique de RI par ex. :

Métrique non supervisée :

$$idf(t_k) = \log_2\left(\frac{N}{N_{t_k}}\right) \text{ [Sparck Jones, 1972]}$$

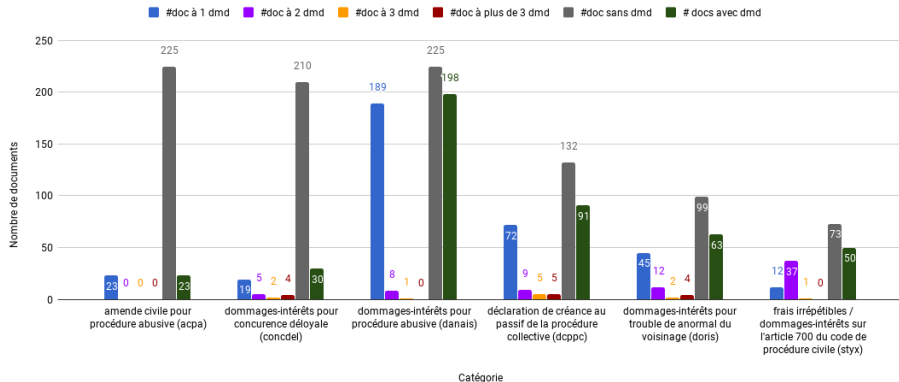
Métriques supervisées :

$$gss(t_k, c_i) = (N_{t_k, c_i} N_{\bar{t}_k, \bar{c}_i}) - (N_{t_k, \bar{c}_i} N_{\bar{t}_k, c_i}) \text{ [Galavotti et al., 2000]}$$

$$ngl(t_k, c_i) = \frac{\sqrt{N}((N_{t_k, c_i} N_{\bar{t}_k, \bar{c}_i}) - (N_{t_k, \bar{c}_i} N_{\bar{t}_k, c_i}))}{\sqrt{N_{t_k} N_{\bar{t}_k} N_{c_i} N_{\bar{c}_i}}} \text{ [Ng et al., 1997]}$$

2.2 sélection des termes aux poids

## Répartition des demandes dans les documents annotées pour chaque catégorie



P chaque catégorie, plus de 50% des documents annotées ont une seule demande. (sauf pour l'article 700)

# Métriques d'évaluation

Catégorie  $c_i$ , tuple d'information  $I \subseteq \{Q_{DMD}, S_{RST}, Q_{RST}\}$

Corpus d'évaluation  $D = D_{c_i} \cup D_{\bar{c}_i} = \{D_j\}_{1 \leq j \leq |D|}$ , où  $D_j$  est un document

$$\text{Nombre de vrais positifs (bons)} : TP_{c_i, I, D} = \sum_{j=1}^{|D|} TP_{c_i, I, D_j}$$

$$\text{Nombre de faux positifs (en trop)} : FP_{c_i, I, D} = \sum_{j=1}^{|D|} FP_{c_i, I, D_j}$$

$$\text{Nombre de faux négatifs (manqués)} : FN_{c_i, I, D} = \sum_{j=1}^{|D|} FN_{c_i, I, D_j}$$

$$Precision_{c_i, I, D} = \frac{TP_{c_i, I, D}}{TP_{c_i, I, D} + FP_{c_i, I, D}}$$

$$Rappel_{c_i, I, D} = \frac{TP_{c_i, I, D}}{TP_{c_i, I, D} + FN_{c_i, I, D}}$$

$$F1_{c_i, I, D} = 2 \times \frac{Precision_{c_i, I, D} \times Rappel_{c_i, I, D}}{Precision_{c_i, I, D} + Rappel_{c_i, I, D}}$$

# Evaluation de la détection des catégories

TABLE – Resultats d’une 5-fold cross-validation sur  $D$  pour la detection categorie (P= Precision, R=Rappel, F1 = F1-mesure)

	Naïf Bayésien			Arbre de décision			KNN			SVM		
Category	P	R	F1	P	R	F1	P	R	F1	P	R	F1
acpa	1.0	1.0	1.0	0.996	0.955	0.972	1.0	1.0	1.0	0.996	0.955	0.972
concdel	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.995	0.967	0.979
danais	0.988	0.989	0.988	0.996	0.995	0.995	0.995	0.995	0.995	0.993	0.993	0.993
dcppc	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
doris	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
styx	1.0	1.0	1.0	0.984	0.983	0.983	1.0	1.0	1.0	1.0	1.0	1.0

# Quelle métrique de pondération et quel zonage ?

TABLE — Comparaison des métriques de pondération et des stratégies de zonage sur le corpus  $D$  (F1-mesure sur l'extraction du tuple  $(Q_{DMD}, S_{RST}, Q_{RST})$ )

Métrique	acpa		concdel		danais		dcppc		doris		styx	
	tp	zw	tp	zw	tp	zw	tp	zw	tp	zw	tp	zw
CHI2	0.683	0.698	0.061	0.061	0.443	0.411	0.259	0.264	0.187	0.071	0.321	0.366
DBIDF	0.683	0.698	0.076	0.033	0.461	0.416	0.254	0.264	0.084	0	0.331	0.358
DELTADF	0.683	0.698	0.144	0.082	0.443	0.41	0.259	0.264	0.143	0.142	0.334	0.281
DSIDF	0.678	0.698	0.076	0.052	0.399	0.152	0.014	0	0.019	0	0.343	0.33
GSS	0.683	0.698	0.144	0.082	0.443	0.41	0.259	0.264	0.143	0.142	0.334	0.281
IDF	0.067	0	0.033	0	0.04	0	0	0	0	0	0	0
IG	0.011	0.049	0.05	0.034	0.304	0.073	0	0	0.019	0	0.058	0
KLD	0.432	0.398	0.146	0.124	0.459	0.409	0.252	0.254	0.158	0.154	0.243	0.42
MAR	0.683	0.698	0.144	0.091	0.443	0.42	0.259	0.264	0.156	0.146	0.334	0.281
NGL	0.683	0.698	0.061	0.034	0.443	0.411	0.259	0.264	0.122	0.02	0.321	0.347
RF	0.683	0.698	0.202	0.043	0.491	0.367	0.242	0.21	0.101	0.058	0.387	0.351
Max	0.683	<b>0.698</b>	<b>0.202</b>	0.124	<b>0.491</b>	0.42	0.259	<b>0.264</b>	<b>0.187</b>	0.154	0.387	<b>0.42</b>

tp = zonage par la seule présence d'un trigger

zw = zonage par pondération des passages à somme d'argent

La métrique et la stratégie de zonage dépendent de la catégorie

# Exemple de termes sélectionnés

concdel		danais	
NGL	DSIDF	NGL	DSIDF
déloyale	concurrence déloyale	procédure abusive	procédure abusive et injustifiée
perte	déloyale	32-1	fondement de l' article 32-1
actes		abusive	dommages-intérêts pour procédure abusive
50.000	agissements	intérêts pour procédure	titre de dommages-intérêts pour procédure abusive

$$n_{gl}(t_k, c_i) = \frac{\sqrt{N}((N_{t_k, c_i} N_{\overline{t_k}, \overline{c_i}}) - (N_{t_k, \overline{c_i}} N_{\overline{t_k}, c_i}))}{\sqrt{N_{t_k} N_{\overline{t_k}} N_{c_i} N_{\overline{c_i}}}}$$

$$dsidf(t_k, c_i) = \log\left(\frac{(N_{\overline{c_i}} N_{t_k, c_i}) + 0.5}{(N_{c_i} N_{\overline{t_k}, \overline{c_i}}) + 0.5}\right)$$



# Entrainement avec sélection de la meilleure métrique (1)

$c_i$	Tuple d'info ( $I$ )	$P_{c_i, I, D_{c_i}}$	$R_{c_i, I, D_{c_i}}$	$F1_{c_i, I, D_{c_i}}$	Docs. Parfaits	#extraits/#attendus/ $ D_{c_i} $
acpa	( $Q_{DMD}$ )	0.709	0.73	0.705	0.47	5.2/4.6/4.6
	( $S_{RST}$ )	0.691	0.7	0.683	0.48	
	( $Q_{RST}$ )	0.72	0.74	0.716	0.48	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.651	0.65	0.638	0.43	
concdel	( $Q_{DMD}$ )	0.461	0.393	0.376	0.233	11.6/11.6/6.0
	( $S_{RST}$ )	0.544	0.442	0.427	0.2	
	( $Q_{RST}$ )	0.595	0.482	0.465	0.2	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.337	0.299	0.28	0.167	
danais	( $Q_{DMD}$ )	0.548	0.516	0.527	0.346	36.6/38.8/37.0
	( $S_{RST}$ )	0.69	0.646	0.661	0.454	
	( $Q_{RST}$ )	0.714	0.666	0.682	0.465	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.482	0.46	0.466	0.314	
dcppc	( $Q_{DMD}$ )	0.334	0.392	0.358	0.217	26.8/22.2/16.6
	( $S_{RST}$ )	0.665	0.798	0.721	0.544	
	( $Q_{RST}$ )	0.62	0.744	0.672	0.509	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.22	0.26	0.237	0.181	
doris	( $Q_{DMD}$ )	0.279	0.373	0.314	0.033	26.8/20.0/12.4
	( $S_{RST}$ )	0.391	0.524	0.439	0.146	
	( $Q_{RST}$ )	0.329	0.414	0.361	0.131	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.177	0.229	0.197	0.017	
styx	( $Q_{DMD}$ )	0.762	0.642	0.695	0.46	15.0/17.8/10.0
	( $S_{RST}$ )	0.701	0.593	0.64	0.34	
	( $Q_{RST}$ )	0.824	0.696	0.752	0.46	
	( $Q_{DMD}, Q_{RST}, S_{RST}$ )	0.44	0.372	0.402	0.28	

TABLE – Zonage par la seule présence d'un trigger (sur le corpus  $D_{c_i}$ )

# Entrainement avec sélection de la meilleure métrique (2)

$c_i$	Tuple d'info ( $I$ )	$P_{c_i, I, D_{c_i}}$	$R_{c_i, I, D_{c_i}}$	$F1_{c_i, I, D_{c_i}}$	Docs. Parfaits	#extraits/#attendus/ $ D_{c_i} $
acpa	( $Q_{DMD}$ )	0.753	0.61	0.672	0.57	3.8/4.6/4.6
	( $S_{RST}$ )	0.92	0.74	0.818	0.7	
	( $Q_{RST}$ )	0.92	0.74	0.818	0.7	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.753	0.61	0.672	0.57	
concdel	( $Q_{DMD}$ )	0.343	0.128	0.11	0.067	5.6/11.6/6.0
	( $S_{RST}$ )	0.535	0.15	0.17	0.067	
	( $Q_{RST}$ )	0.543	0.17	0.182	0.067	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.135	0.098	0.079	0.033	
danais	( $Q_{DMD}$ )	0.66	0.296	0.395	0.227	17.8/38.8/37.0
	( $S_{RST}$ )	0.732	0.328	0.438	0.27	
	( $Q_{RST}$ )	0.77	0.348	0.464	0.276	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.61	0.276	0.367	0.216	
dcppc	( $Q_{DMD}$ )	0.391	0.363	0.372	0.252	21.4/22.2/16.6
	( $S_{RST}$ )	0.732	0.688	0.703	0.532	
	( $Q_{RST}$ )	0.665	0.624	0.638	0.471	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.275	0.248	0.259	0.204	
doris	( $Q_{DMD}$ )	0.211	0.146	0.171	0.064	9.8/20.0/12.4
	( $S_{RST}$ )	0.418	0.217	0.268	0.114	
	( $Q_{RST}$ )	0.342	0.166	0.211	0.096	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.095	0.067	0.078	0.017	
styx	( $Q_{DMD}$ )	0.838	0.632	0.718	0.52	13.2/17.8/10.0
	( $S_{RST}$ )	0.772	0.571	0.654	0.36	
	( $Q_{RST}$ )	0.786	0.583	0.666	0.38	
	( $Q_{DMD}, S_{RST}, Q_{RST}$ )	0.573	0.44	0.496	0.32	

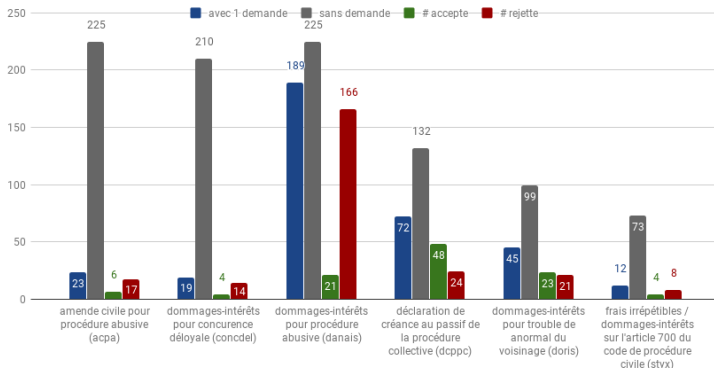
TABLE – Zonage par pondération des passages à somme d'argent (sur le corpus  $D_{c_i}$ )

Classification binaire pour identifier le sens  
du résultat

---

# 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 toujours 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  $r$  ( $f^{\sim(k)} = \hat{r} \circ \hat{f}^{(k)}$ )

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

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

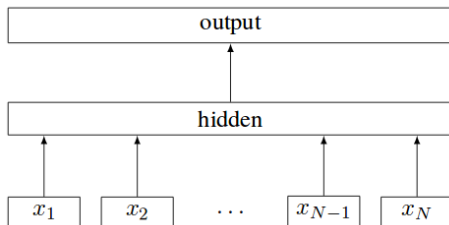


FIGURE – Architecture similaire au model CBOW : le label remplace le mot au milieu.

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

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

# 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-avg
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  $t_1, \dots, t_h$

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

$$\text{avec } w_{hj} = \frac{\text{cov}(u_{(h-1)j}, \epsilon_h)}{\sqrt{\sum_{p=1}^j \text{cov}^2(u_{(h-1)p}, \epsilon_h)}}, \quad y = c_1 t_1 + \dots + c_h t_h + \epsilon_h,$$

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



# Application des extensions de la Régression PLS (2)

1. Gini-PLS : élimination de la sensibilité aux *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  $w_{hj}$  [Souissi and Mussard, 2013]
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	rang
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>	13
AVG-G*TF	StandardPLS	<b>0.636</b>	0.49	danais	0.836	dcppc	<b>0.032</b>	<b>0.346</b>	24
DELTADF*TF	GiniPLS	<b>0.586</b>	0.411	danais	0.837	dcppc	<b>0.082</b>	<b>0.426</b>	169
DELTADF*TF	GiniLogitPLS	<b>0.578</b>	0.225	styx	0.772	dcppc	<b>0.09</b>	<b>0.547</b>	220

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

## Conclusion

---

- Extraction des informations relatives aux demandes :
  - La présence des catégories de demande fonctionne bien avec une simple classification : le vocabulaire de chaque catégorie est très discriminant ;
  - L'identification des informations est une tâche très difficile
  - l'expression du résultat est plus accessible que celle de la demande
- Détermination du sens du résultat par classification binaire :
  - Difficulté de fastText et le NBSVM : déséquilibre des données, faibles nombre d'échantillons,  $N_{accepte} \leq N_{rejette}$  ?

# Objectifs pour la suite

- Rédaction du mémoire
- Valorisation des résultats
- Travail en parallèle sur l'identification non-supervisée des circonstances factuelles

Questions ?

---

# References I



ACE (2005).

*ACE (Automatic Content Extraction) English Annotation Guidelines for Events.*

Linguistic Data Consortium, 5.4.3 edition.

<https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf>.



Ahn, D. (2006).

The stages of event extraction.

In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8. Association for Computational Linguistics.



Galavotti, L., Sebastiani, F., and Simi, M. (2000).

Experiments on the use of feature selection and negative evidence in automated text categorization.

In *International Conference on Theory and Practice of Digital Libraries*, pages 59–68. Springer.



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*, pages 427–431.



Ng, H. T., Goh, W. B., and Low, K. L. (1997).

Feature selection, perceptron learning, and a usability case study for text categorization.

In *ACM SIGIR Forum*, volume 31, pages 67–73. ACM.



Nguyen, T. H., Cho, K., and Grishman, R. (2016).

Joint event extraction via recurrent neural networks.

In *HLT-NAACL*, pages 300–309.



# References II



Palm, R. B., Hovy, D., Laws, F., and Winther, O. (2017).  
End-to-end information extraction without token-level supervision.  
*arXiv preprint arXiv :1707.04913*.



Salton, G. and Buckley, C. (1988).  
Term-weighting approaches in automatic text retrieval.  
*Information processing & management*, 24(5) :513–523.



Souissi, F. and Mussard, S. (2013).  
Gini-pls regressions.  
In *AFSE Meeting 2013*.



Sparck Jones, K. (1972).  
A statistical interpretation of term specificity and its application in retrieval.  
*Journal of documentation*, 28(1) :11–21.



Tenenhaus, M. (2005).  
La regression logistique PLS.



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.



Yang, B. and Mitchell, T. (2016).  
Joint extraction of events and entities within a document context.  
In *Proceedings of NAACL-HLT*, pages 289–299.