

Analyse sémantique d'un corpus exhaustif de décisions jurisprudentielles pour l'élaboration d'un modèle prédictif du risque judiciaire

Comité de suivi individuel – 12 juillet 2017

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:

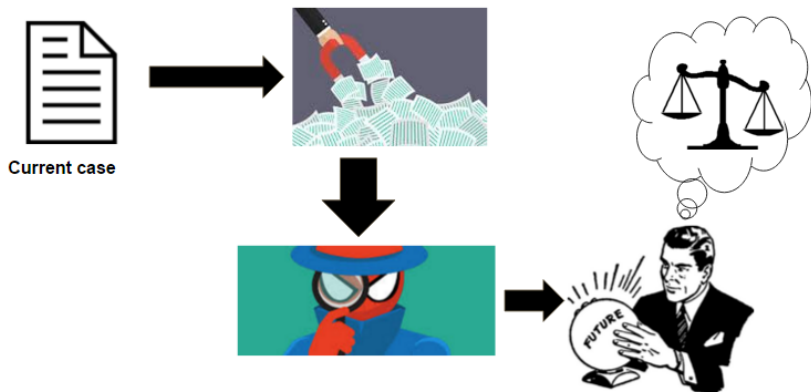
- Sébastien Harispe (Ecole des Mines d'Alès, LGI2P)
- Guillaume Zambrano (Université de Nîmes, CHROME)



1. Motivations et objectifs
2. Détection de sections et d'entités
3. Extraction d'informations sur les demandes
4. Activités complémentaires
5. Conclusion et plan de travail
6. Questions ?

Motivations et objectifs

Les juristes analysent les décisions afin d'anticiper



Plus de 4 millions de décisions prononcées / an

| | 2010 | 2011 | 2012 | 2013 | 2014 |
|-------------------------------|-----------|-----------|-----------|-----------|-----------|
| Justice civile | 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

Défis : 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

Ex : gérant **et** pouvoir, bail **s/5** résil!
[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

Source : LexisNexis.com

Défis : Documents non-structurés

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,

...

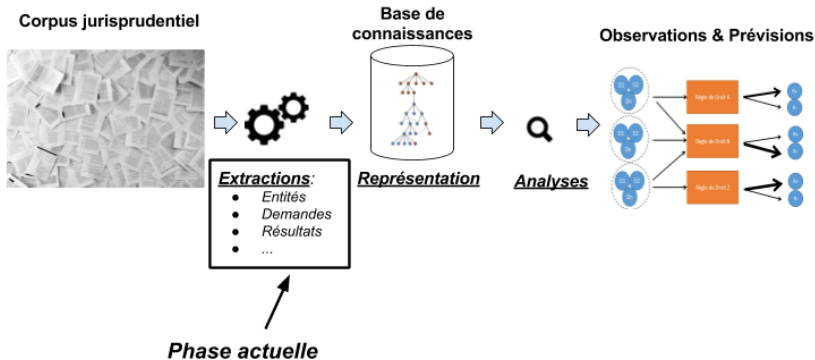
Déboute Madame A. de sa demande de
provision sur dommages-intérêts.

...

Vu l'article 700 du Code de Procédure
Civile,

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

Notre projet : Automatiser la structuration et l'analyse



Elaboration et mise en oeuvre de techniques de :

- Traitement du langage naturel
- Représentation des connaissances
- Recherche d'information

Détection de sections et d'entités

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.

Entêtes : méta-données

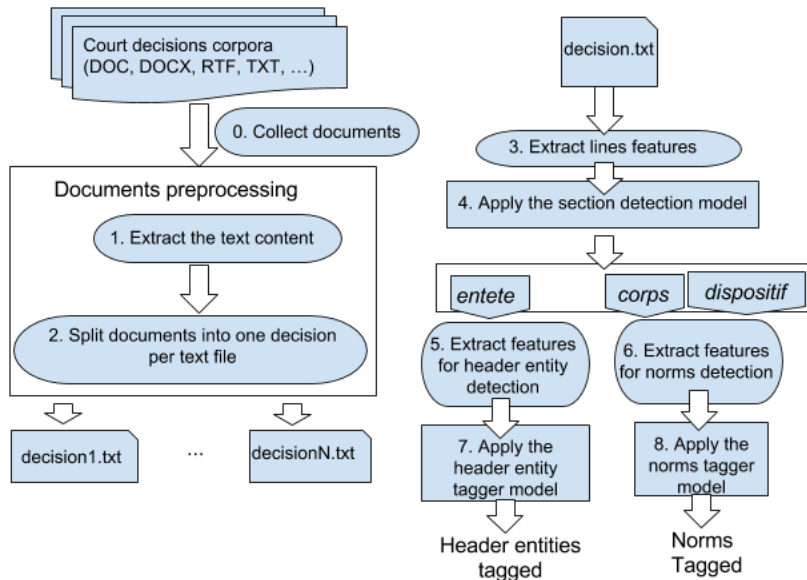
Dispositif : résultats et normes

Entités et sections à détecter

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

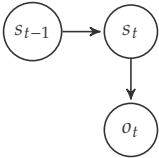
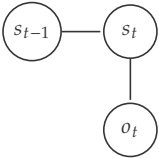
TABLE — Entités et leurs labels par section.

Architecture proposée



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>[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

Premiers résultats [Tagny Ngompé et al., 2017]

| | 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 & 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 (%).

- Utilité de la prise en compte des particularités des textes
 - forme : le mot est-il en majuscule, lemmes, longueur de la ligne, ...
 - contexte : mots voisins, position par rapport à un mot-clé, ...
- Certaines entités restent difficiles à détecter

Comment améliorer les résultats ?

Définir plus de caractéristiques :

- 14 pour les sections
- 35 pour les entêtes
- 28 pour les normes

Résultats avec plus de caractéristiques

| | Precision | Recall | F1 |
|--------------|-----------|--------|-------|
| I-corps | 99.57% | 99.69% | 99.63 |
| I-dispositif | 98.63% | 97.59% | 98.11 |
| I-entete | 99.51% | 99.55% | 99.53 |
| Overall | 99.48% | 99.48% | 99.48 |

| | | | |
|---------------|--------|--------|-------|
| I-appelant | 84.34% | 76.27% | 80.10 |
| I-avocat | 98.02% | 98.15% | 98.09 |
| I-date | 98.00% | 96.60% | 97.30 |
| I-fonction | 95.23% | 95.13% | 95.18 |
| I-formation | 98.80% | 99.45% | 99.12 |
| I-intervenant | 83.38% | 68.26% | 75.07 |
| I-intime | 82.54% | 83.33% | 82.93 |
| I-juge | 97.55% | 97.23% | 97.39 |
| I-juridiction | 98.91% | 99.69% | 99.30 |
| I-rg | 97.81% | 97.44% | 97.62 |
| I-ville | 98.94% | 99.15% | 99.04 |
| Overall | 95.13% | 94.51% | 94.82 |

| | | | |
|---------|--------|--------|-------|
| I-norme | 97.14% | 96.09% | 96.62 |
|---------|--------|--------|-------|

TABLE — Résultats du CRF avec l'ajout de caractéristiques.

Sélection des caractéristiques

| Detection Task | Tagger | Token-level F1 | Entity-level F1 | Features subset |
|-----------------|--------|----------------|-----------------|------------------|
| Sections | CRF | 99.31 | 90.48 | BDS |
| | | 99.55 | 85.76 | SFFS |
| | | 99.46 | 90.03 | ALL |
| | | 91.75 | 60.26 | token |
| | HMM | 90.99 | 3.89 | absLength |
| | | 86.97 | 3.65 | relLength |
| | | 37.59 | 18.81 | token |
| Header entities | CRF | 92.69 | 90.47 | BDS |
| | | 93.00 | 90.76 | SFFS |
| | | 92.74 | 90.81 | ALL |
| | | 82.73 | 72.17 | token |
| | HMM | 78.61 | 56.93 | token |
| | | 68.04 | 32.96 | lemma_Wo |
| | | 38.54 | 7.95 | POS |
| Norms | CRF | 96.31 | 90.80 | BDS |
| | | 95.57 | 89.29 | SFFS |
| | | 95.87 | 90.76 | ALL |
| | | 94.26 | 85.72 | token |
| | HMM | 91.66 | 74.9 | token |
| | | 91.54 | 69.35 | lemma_Wo |

TABLE — Impact de la réduction des caractéristiques

BDS et SFFS très lents (plus de 10 h lors de nos tests)

Nombre nécessaire de données d'entraînement

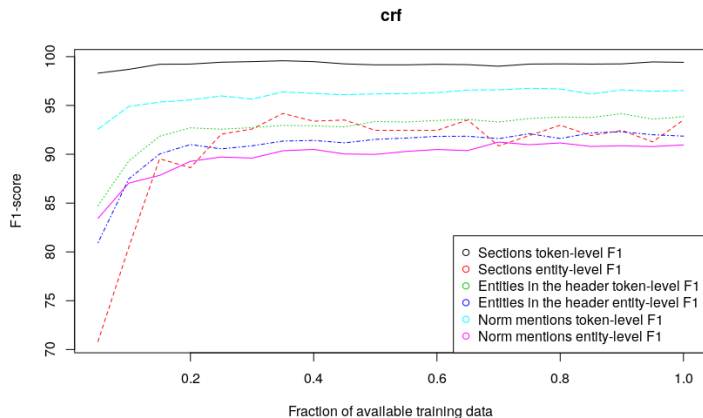


FIGURE — Résultats en fonction du nombre de données d'entraînement (fractions d'environ 380 décisions)

Extraction d'informations sur les demandes

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

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

EXPRESSION DE DEMANDE

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 DE RESULTAT

La cour, ...

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

Simplification du problème

- On suppose qu'une décision ne comprend qu'au plus une demande d'une catégorie donnée
- Méthode générique qui s'adapte aux spécificités de la catégorie traitée
- Définition incrémentale des catégories

Approche supervisée d'extraction des demandes

(1) Sélection de termes caractéristiques

DOMMAGES-INTERETS POUR ABUS DE PROCEDURE

| Terme (n-gram) | Poids global (NGL) |
|---------------------------------|--------------------|
| procédure abusive | 15.710 |
| pour procédure abusive | 15.007 |
| pour procédure | 14.890 |
| abusive | 13.721 |
| intérêts pour procédure | 10.306 |
| abus | 10.288 |
| intérêts pour procédure abusive | 9.984 |
| 32-1 | 9.534 |
| ... | ... |

$$n gl(w, c) = \frac{\sqrt{N}((N_{w,c}N_{\bar{w},\bar{c}})-(N_{w,\bar{c}}N_{\bar{w},c}))}{\sqrt{N_w N_{\bar{w}} N_c N_{\bar{c}}}} \text{ [Ng et al., 1997]}$$

Détection d'une catégorie par classification binaire

Conditions d'expérimentation

- Représentation vectorielle :

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

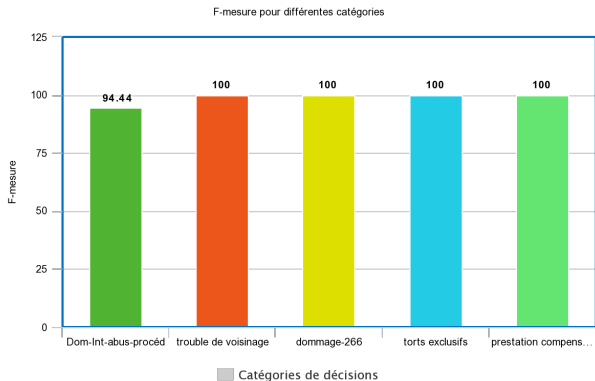
- Évaluation de différentes configurations :

- dimensions des vecteurs : 10, ..., 250, ...
- méthodes de sélection de termes discriminants :
 $\chi^2, \Delta_{DF}, Marascuilo, NGL, GSS$...
- méthodes de classification : SVM, arbre de décision, KNN, naïf bayésien (avec Weka[Frank et al., 2016])
- méthodes de pondération locale : 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

Détection d'une catégorie par classification binaire

Premiers résultats :



meta-chart.com

FIGURE — Résultats des meilleures configurations (taille des vecteurs, poids global, poids local, modèle de classifieur)

Extraction du sens du résultat (avec la même approche)

Classification des décisions d'une catégorie prédéfinie

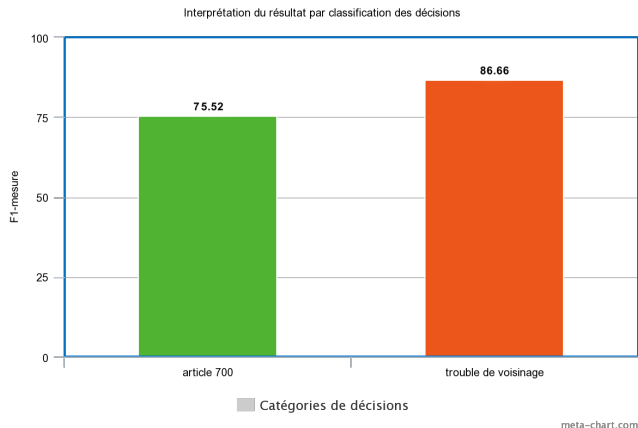


FIGURE – Résultats des meilleures configurations (taille des vecteurs, poids global, poids local, modèle de classifieur)

Extraction du sens du résultat (méthodes Gini-PLS)

Combinaison de 2 méthodes de régression :

1. PLS : réduction supervisée des dimensions x_1, x_2, \dots, x_p en composantes orthogonales t_1, \dots, t_h

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

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

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

2. Gini : élimination de la sensibilité au *outliers* en remplaçant la covariance $\text{cov}(x_j, y)$ par la covariance de Gini $\text{cog}(y; x_j) := \text{cov}(y; R(x_j))$

[Souissi and Mussard, 2013]

Activités complémentaires

- Formations complémentaires : 11 modules (132h)
- Enseignement : travaux pratiques (Big Data avec Hadoop)
- Valorisation des travaux :
 - Conférence EGC, Grenoble, janvier 2017
 - 1 article en relecture (AKDM8)
 - Démo des 1er résultats : SAT AXLR (Montpellier)
 - Séminaire e-juris (Lyon)
- Participation au challenge COLIEE : 4e place / 12

Conclusion et plan de travail

- Détection d'entités et de sections basée HMM / CRF
 - Bons résultats même avec un peu de données annotées
 - Difficultés :
 - Annotation manuelle d'un jeu suffisant d'exemples
 - Identification de bons descripteurs
 - Lenteur de la sélection de caractéristiques
 - Limite de l'approche :
 - Descripteurs définis manuellement
 - Etiquetage en plusieurs passes
- Détection de termes propres aux catégories de demandes
- Détection des catégories par classification
- Détection moins triviale du sens du résultat

1. Extraction des demandes et résultats par affinement de la segmentation des textes
2. Standardisation et représentation des informations extraites sous forme de base de connaissances
3. Détermination des facteurs associables aux décisions des juges (faits ou arguments)

Questions ?

References I



Frank, E., Hall, M. A., and Witten, I. H. (2016).
The WEKA Workbench, chapter Online Appendix for "Data Mining : Practical Machine Learning Tools and Techniques".
Morgan Kaufmann.



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.



Peng, F. and McCallum, A. (2006).
Information extraction from research papers using conditional random fields.
Information processing & management, 42(4) :963–979.



Seymore, K., McCallum, A., and Rosenfeld, R. (1999).
Learning hidden Markov model structure for information extraction.
AAAI-99 Workshop on Machine . . .



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



Tagny Ngompé, G., Harispe, S., Zambrano, G., Montmain, J., and Mussard, S. (January 2017).
Reconnaissance de sections et d'entités dans les décisions de justice : application des modèles probabilistes HMM et CRF.
In *In Extraction et Gestion des Connaissances - EGC 2017, Revue des Nouvelles Technologies de l'Information, Grenoble, France*.