Méthodes d'analyse sémantique de corpus de décisions jurisprudentielles

Soutenance de thèse

Gildas TAGNY NGOMPÉ

24 janvier 2020

Jury:

- 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 (Examinateur)
- Fabrice MUHLENBACH, Maître de Conférences, Université Jean Monnet de Saint-Étienne (Examinateur)
- Guillaume ZAMBRANO, Maître de Conférences, Université de Nîmes (Encadrant de proximité)
- Sébastien HARISPE, Maître Assistant, IMT Mines Alès (Encadrant de proximité)

IMT Mines Alès
École Mines-Télécom

LGI²P

risques chroniques émergents
CHROME
EA 7352

ED 583
RISQUES & SOCIETE

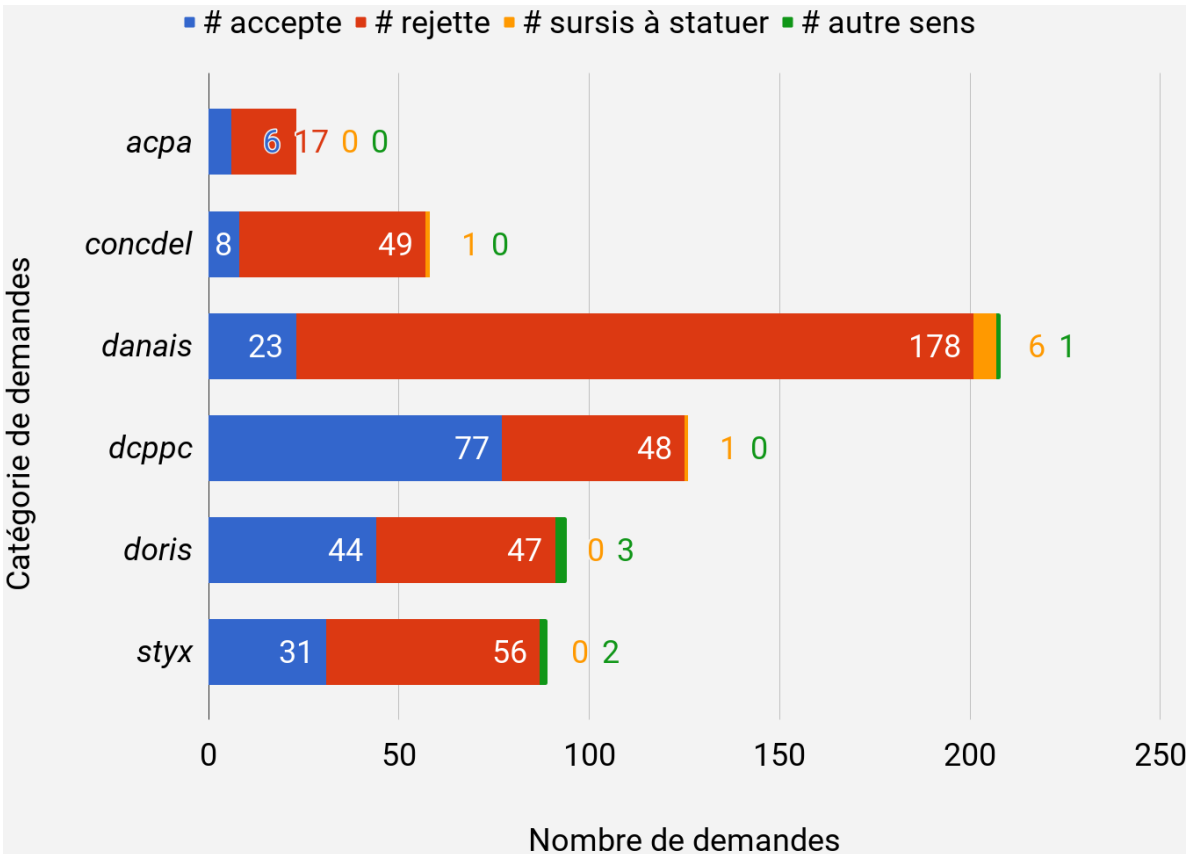
Collège Doctoral
Languedoc-Roussillon

Plan

1. Identification du sens du résultat

Restriction du problème d'identification des demandes

- Uniquement les décisions à une demande de la catégorie
  - Raison : plus de 50% des documents dans la majorité des catégories
- Classification binaire (éviter les subtilités de rédaction)
  - Raison : les demandes sont en majorité **acceptées** ou **rejetées**



Plusieurs algorithmes classiques existent

- Classifieur bayésien naïf [Duda et al., 1973]
- K-plus-proches-voisins [Cover and Hart, 1967]
- SVM [Vapnik, 1995]
- Arbre de décision
- Analyse discriminante linéaire [Fisher, 1936] et quadratique [McLachlan, 1992]
- NBSVM [Wang and Manning, 2012]
- fastText [Grave et al., 2017]
- etc.

Régression Gini-PLS

- PLS [Wold, 1966] (régression partielle des moindres carrés)
  - Réduction supervisée de dimensions $x_1, x_2, ..., x_m$ en composantes orthogonales $t_1, ..., t_h$

$$t_h = \sum_{k=1}^{m} w_{hk} \cdot \hat{U}_{(h-1)k}$$

    avec $\hat{U}_{0k} = x_k, \forall h > 0, \hat{U}_{hk}$ est le résidu de la régression de x_k sur $t_1, ..., t_{h-1}$ ()
    et $w_{hj} = \frac{cov(\varepsilon_h, \hat{U}_{(h-1)j})}{\sqrt{\sum_{j=1}^{m} cov^2(\varepsilon_h, \hat{U}_{(h-1)j})}}$
  - Régression de y dans l'espace réduit

$$y = c_1 t_1 + ... + c_h t_h + \varepsilon_h$$

- Gini-PLS [Mussard and Souissi-Benrejab, 2018]
  - Remplacement de la covariance $cov(x_j, y)$ par la covariance de Gini $cog(y; x_j) := cov(y; R(x_j))$
  -

1.2 Méthode proposée : adaptations de la régression Gini-PLS

1. Gini-PLS généralisé
   - Utilisation de l'opérateur co-Gini généralisé :
     $cog_\nu(x_\ell, x_k) := -\nu cov(x_\ell, r_k^{\nu-1});\ \nu > 1$

   pous disposer d'un curseur ν permettant de régler le compromis entre l'atténuation de la variabilité des variables et l'influence des queues de distributions de ces variables

2. Logit-PLS : $\forall j > 1$, les w_{hj} sont les coefficients de la régression logistique de y sur les composantes $t_1, ..., t_{h-1}, u_{(h-1)j}$ [Tenenhaus, 2005]

3. Gini-Logit-PLS : covariance Gini pour $u_{(h)j}$ et coefficient Logit pour les w_{hj}

Comaparaison des classifieurs PLS aux classifieurs classiques

Représentation	**Algorithme**	F_1	$F_{1_{\text{arbre}}} - F_1$	$F_{1_{\textbf{max}}} - F_{1_{\textbf{min}}}$
$tf - gss$	Arbre	0.668	0	0.42
$tf - avg_{global}$	LogitPLS	0.648	0.02	0.263
$tf - avg_{global}$	StandardPLS	0.636	0.032	0.346
$tf - \Delta_{DF}$	GiniPLS	0.586	0.082	0.426
$tf - \Delta_{DF}$	GiniLogitPLS	0.578	0.09	0.547
-	NBSVM	0.494	0.174	0.434
-	fastText	0.412	0.256	0.127

Amélioration de la classification par restriction du document

Catégorie	Zone	Représentation	Algorithme	F_1
acpa	**demande_resultat_a_resultat_context**	$tf - dbidf$	**Arbre**	**0.846**
	litige_motifs_dispositif	$tf - dbidf$	StandardPLS	0.697
	litige_motifs_dispositif	$tf - avg_{global}$	LogitPLS	0.683
concdel	**litige_motifs_dispositif**	$tf - gss$	**Arbre**	**0.798**
	motifs	$tf - idf$	GiniLogitPLS	0.703
	context	$logave - dbidf$	StandardPLS	0.657
danais	**demande_resultat_a_resultat_context**	$avg_{local} - \chi^2$	**Arbre**	**0.813**
	demande_resultat_a_resultat_context	$atf - avg_{global}$	LogitPLS	0.721
	demande_resultat_a_resultat_context	$atf - avg_{global}$	StandardPLS	0.695
dcppc	**demande_resultat_a_resultat_context**	$tf - \chi^2$	**Arbre**	**0.985**
	demande_resultat_a_resultat_context	$tf - \chi^2$	LogitPLS	0.94
	litige_motifs_dispositif	$tp - mar$	StandardPLS	0.934
doris	**litige_motifs_dispositif**	$tp - dsidf$	**GiniPLS**	**0.806**
	litige_motifs_dispositif	$tp - dsidf$	GiniLogitPLS	0.806
	litige_motifs_dispositif	$atf - ig$	StandardPLS	0.772
styx	**motifs**	$tf - dsidf$	**Arbre**	**1**
	demande_resultat_a_resultat_context	$logave - dsidf$	GiniLogitPLS	0.917
	litige_motifs_dispositif	$tf - rf$	GiniPLS	0.833

Questions

References I

Cover, T. and Hart, P. (1967).
Nearest Neighbor Pattern Classification.
IEEE Transactions on Information Theory, 13(1) :21–27.

Duda, R. O., Hart, P. E., et al. (1973).
Pattern Classification And Scene Analysis, volume 3.
John Wiley & Sons, New York.

Fisher, R. A. (1936).
The use of multiple measurements in taxonomic problems.
Annals of Eugenics, 7(2) :179–188.

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.

McLachlan, G. J. (1992).
Discriminant Analysis and Statistical Pattern Recognition.
John Wiley & Sons.

Mussard, S. and Souissi-Benrejab, F. (2018).
Gini-PLS Regressions.
Journal of Quantitative Economics, pages 1–36.

References II

Tenenhaus, M. (2005).
La regression logistique PLS.
In Droesbeke, Jean-Jacques and Lejeune, Michel and Saporta, Gilbert, editor, *Modèles statistiques pour données qualitatives*, chapter 12, pages 263–276. Editions Technip.

Vapnik, V. N. (1995).
The Nature of Statistical Learning Theory.
Springer.

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.

Wold, H. (1966).
Estimation of principal components and related models by iterative least squares.
Multivariate Analysis, pages 391–420.