# Vectorisation paramétrée des données textuelles

Célia da Costa Pereira*, Mathieu Lafourcade**,
Patrick Lloret***, Cédric Lopez****, Mathieu Roche**,♯

* I3S, UMR 7271, Sophia Antipolis - France – celia.pereira@unice.fr
** LIRMM, UMR 5506, Montpellier - France – {prénom.nom}@lirmm.fr
*** Succeed Together, Paris - France – plloret@succeed-together.eu
**** Objet Direct, Grenoble - France – clopez@objetdirect.com
♯ UMR TETIS, Montpellier - France – mathieu.roche@cirad.fr

## 1 Introduction

L'expression en langage naturel recèle des informations riches que les analystes souhaitent souvent explorer. Dans le cadre de l'activité de la Société *Succeed Together* qui consiste, entre autres, à recueillir et analyser des informations produites lors de séminaires interactifs, les animateurs développent et structurent les discussions établies avec les participants. Les réponses ou remarques apportées par les participants peuvent alors être consignées puis traitées, une phase de regroupent est au préalable nécessaire. Le but est ainsi de mettre en exergue des sentiments partagés par les participants selon une thématique donnée. Dans ce cadre, les travaux menés par le LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) liés au traitement automatique des données textuelles, permettent aux experts de Succeed Together d'analyser semi-automatiquement et à plus grande échelle les données. Ainsi, nous avons focalisé notre étude sur la représentation des données textuelles par des méthodes de TAL (Traitement Automatique du Langage Naturel). Ceci permet, en particulier, d'améliorer les méthodes de classification et/ou regroupement effectuées par le deuxième collaborateur académique du projet (I3S / Université de Nice).

Dans un premier temps, en section 2, nous décrivons les méthodes de représentation des descripteurs textuels. Une application spécifiquement dédiée au projet a été développée. Cette application est décrite en section 3. Les résultats issus des données fournies par la société sont décrits et analysés en section 4. Enfin, quelques perspectives sont données en sections 5.

## 2 Descripteurs textuels pour les tâches de Clustering

La sélection de descripteurs pertinents à partir de textes est une étape indispensable pour une tâche de clustering (regroupement) qui consiste à regrouper les documents ayant des contenus sémantiques proches. Pour appliquer les algorithmes de regroupement, il est dans un premier temps nécessaire d'établir une représentation pertinente des documents (Béchet (2009)). Dans cet article, nous nous concentrons sur la représentation vectorielle de Salton et al. (1975).