

Stage M2 Recherche en Informatique :
**Classification automatique de documents
hétérogènes à faible contenu textuel**

Responsables : Mathieu Roche*, Nicolas Béchet*,
Vincent Poulain d'Andecy**

*Équipe TAL, LIRMM,
mroche@lirmm.fr

**Itesoft

Vincent.PoulaindAndecy@itesoft.com

1 Description

Le contexte du stage concerne l'étude et la mise œuvre de méthodes automatiques de classification de documents hétérogènes (notes, documents administratifs numérisés, etc) issus de la société *Itesoft*. La tâche de classification dans ce contexte peut se révéler extrêmement complexe. La complexité est due à la diversité des documents traités mais surtout au faible contenu de ces derniers. À titre d'exemple, les documents sont syntaxiquement pauvres et, de manière générale, ces derniers possèdent très peu de phrases bien formulées en langage naturel.

Au cours du stage, un état de l'art complet sur la classification de documents textuels à contenu faible et bruité devra être étudié. Par ailleurs, des techniques de classification adaptées à ce contexte devront être proposées et évaluées rigoureusement à partir des données d'*Itesoft*.