

Groupe de travail CorText
14/06/2012 - Marne-la-Vallée

Nuages arborés pour extraire et visualiser des informations de corpus scientifiques

Philippe Gambette

LIGM

Université Paris-Est

Marne-la-Vallée



Une problématique liée à la terminologie

Construction automatique de l'index d'un rapport scientifique
→ liste des mots à sélectionner ?

Liste des mots par fréquence décroissante :

- enlever les mots outils
- comparer avec un index réel

Une problématique liée à la terminologie

Liste des mots par fréquence décroissante sans mots outils

Par rapport à l'index réel, **bruit** :

- **important** si mots **non lemmatisés** → lemmatiser
- **important** pour les **verbes, adverbes, noms d'auteurs**
- **augmente fortement** quand la **fréquence diminue**
(présence de singletons dans l'index réel)

Par rapport à l'index réel, **silence** :

- **expressions multi-mots** → segments répétés
- certains **mots outils** (face)
- **nominalisation** d'adjectifs, de certains verbes

Une problématique liée à la terminologie

Liste des mots par fréquence décroissante sans mots outils

Par rapport à l'index réel, **bruit** :

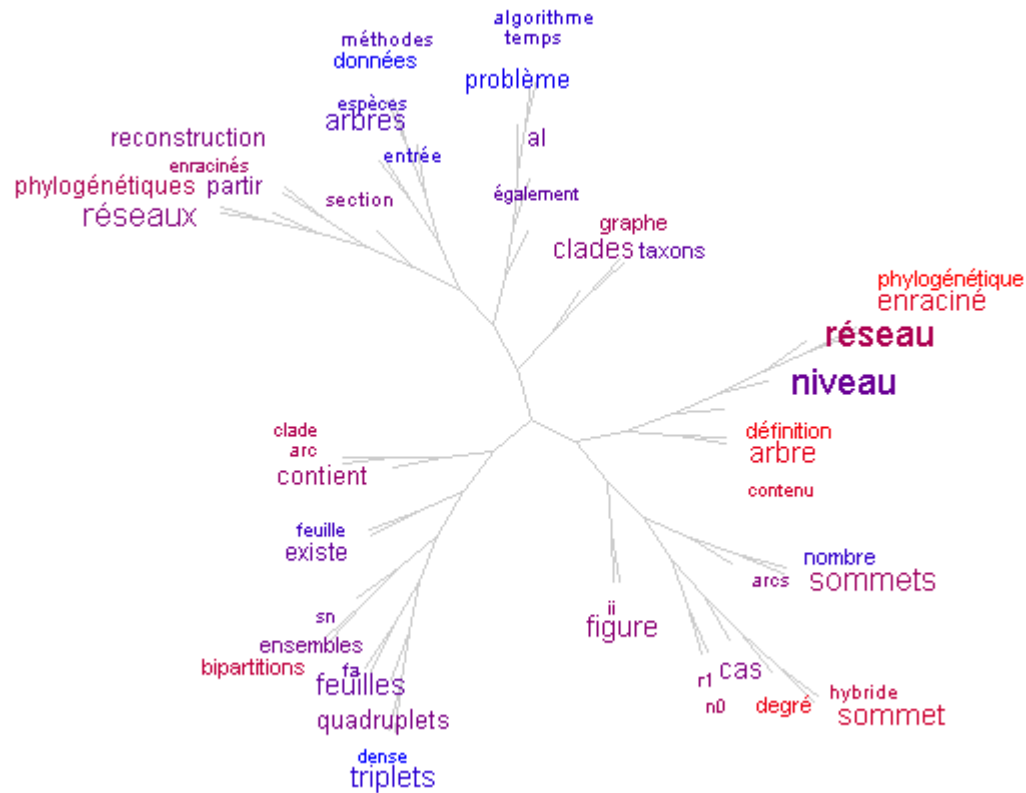
- **important** si mots **non lemmatisés** → lemmatiser
- **important** pour les **verbes, adverbes, noms d'auteurs**
- **augmente fortement** quand la **fréquence diminue**
(présence de singletons dans l'index réel)

Par rapport à l'index réel, **silence** :

- **expressions multi-mots** → segments répétés
 - certains **mots outils** (face)
 - **nominalisation** d'adjectifs, de certains verbes
- Méthode semi-automatique
-
- ```
graph TD; A[segments répétés] --> B[Méthode semi-automatique]; C[nominalisation] --> B;
```

# Méthode semi-automatique

## Nuage arboré des mots les plus fréquents





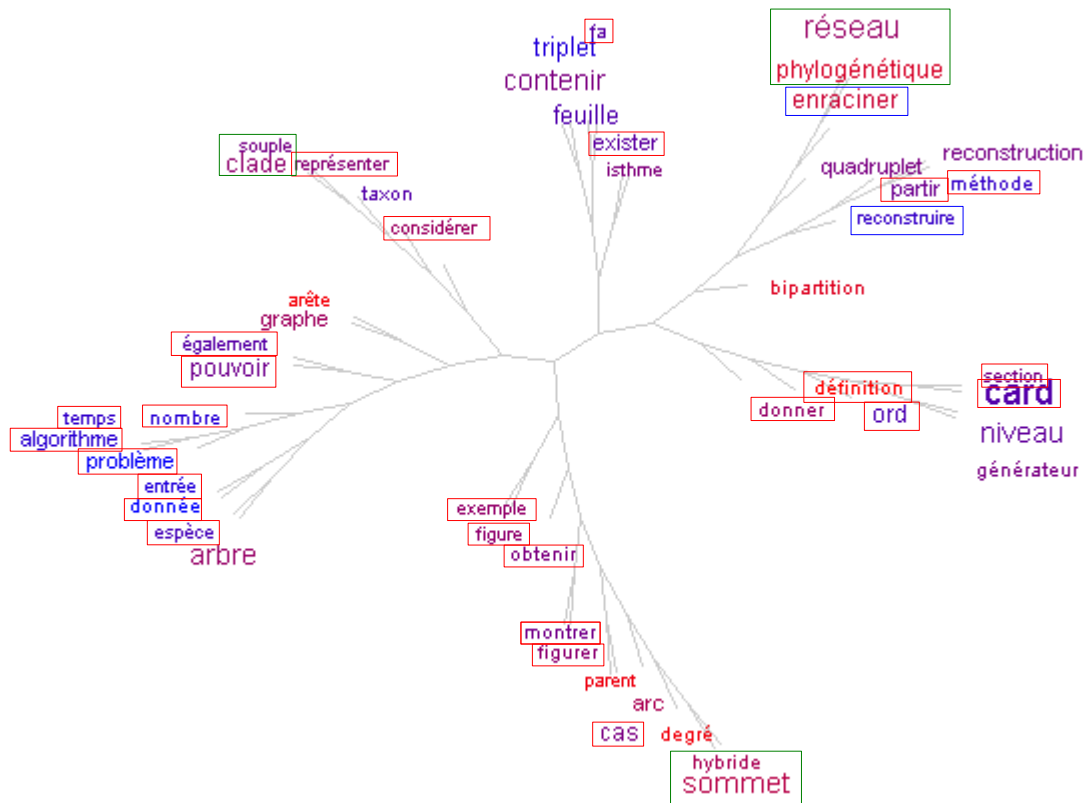






# Méthode semi-automatique

Nuage des mots les plus fréquents du **texte lemmatisé**



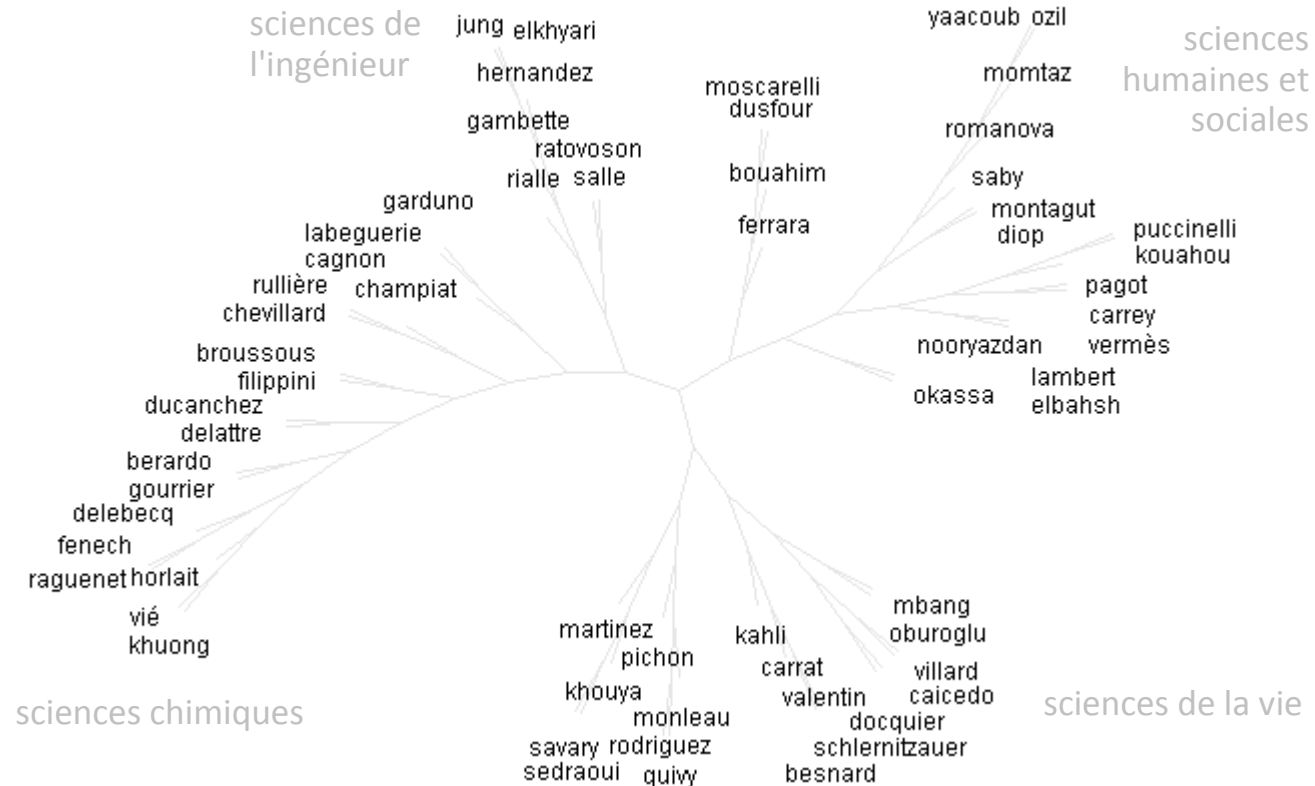
**Bruit** : mots  
à supprimer

**Nominalisations**  
à effectuer

**Expressions  
multimots** à créer

# Autres utilisations du nuage arboré

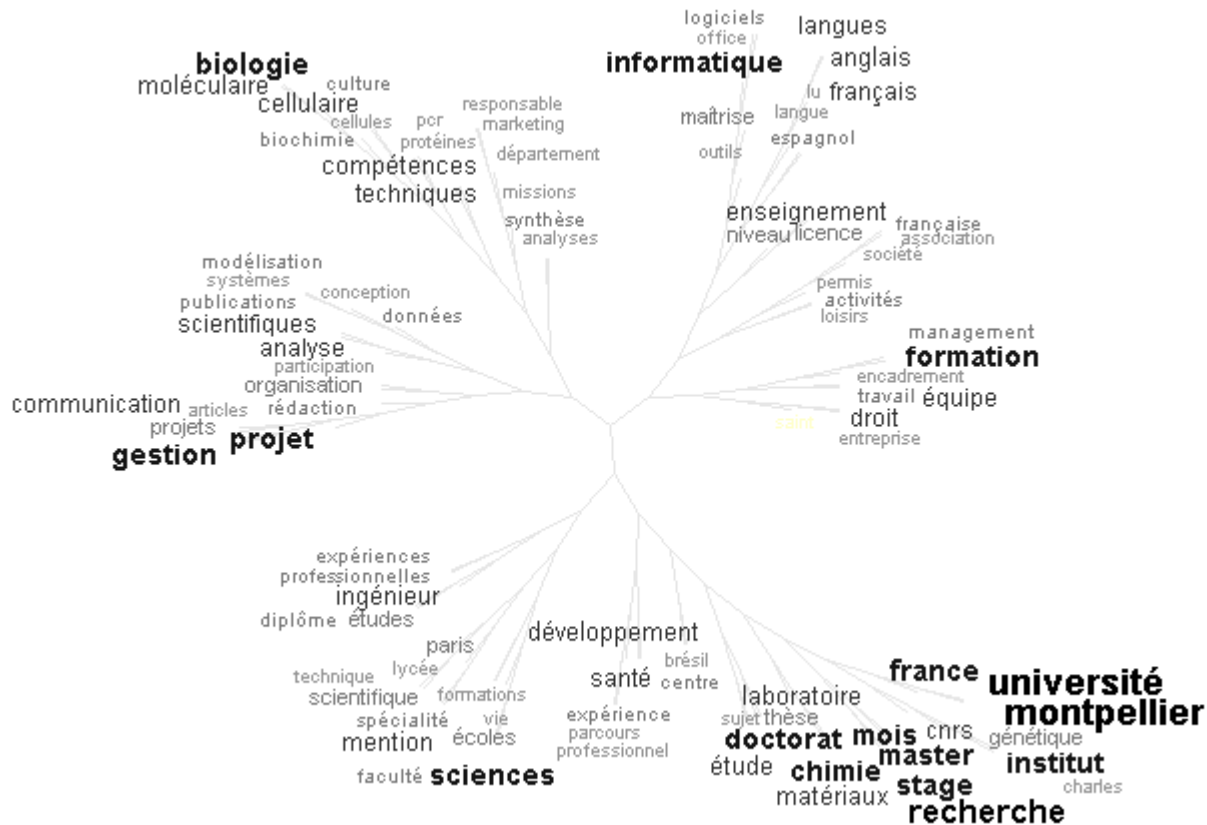
## Catégorisation de textes



Catégorisation de CV de docteurs pour la rencontre Docteurs & Entreprises à Montpellier en 2011

# Autres utilisations du nuage arboré

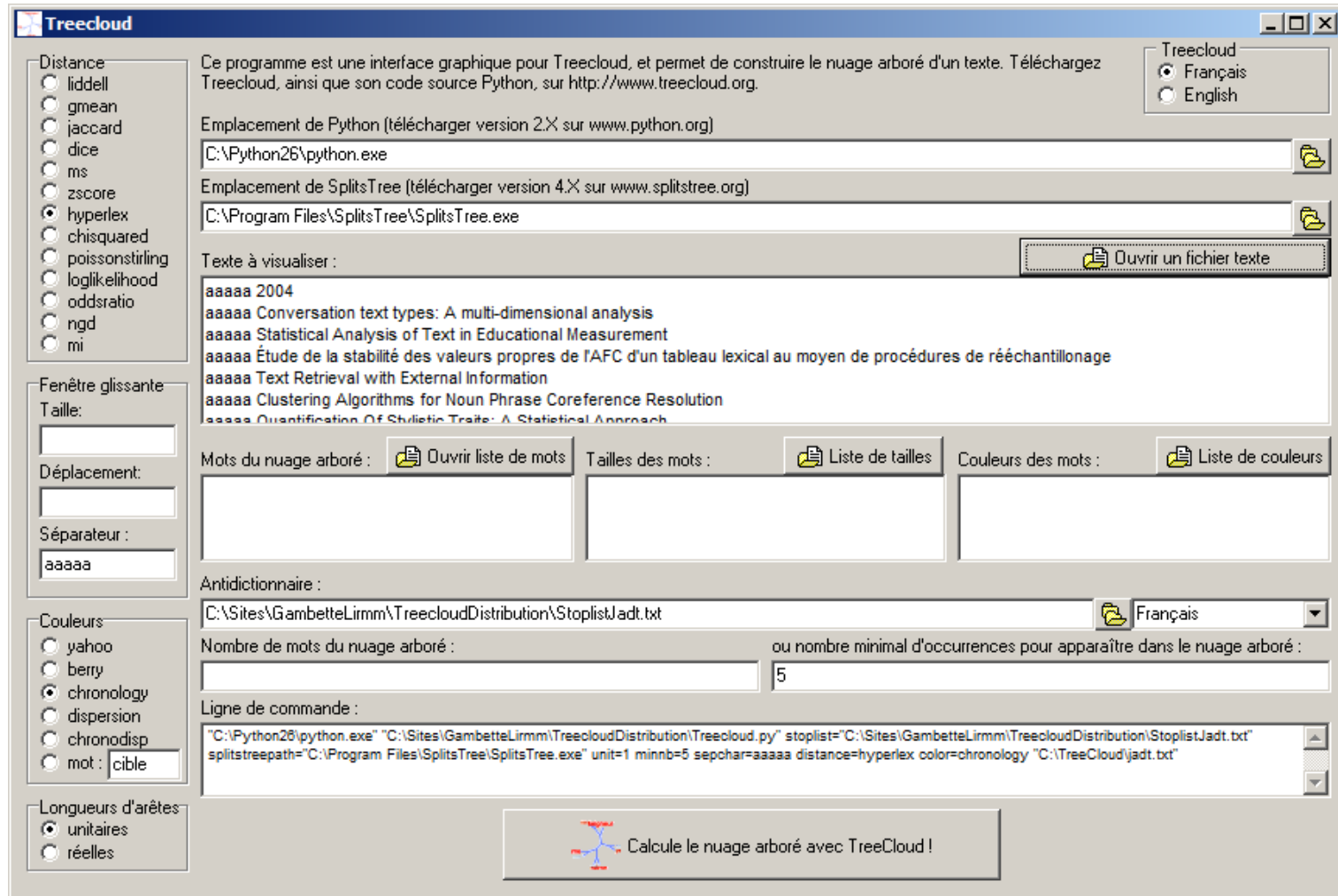
## Thématiques de textes



Catégorisation de mots-clés issus de CV de docteurs pour la rencontre Docteurs & Entreprises à Montpellier en 2011

# Autres utilisations du nuage arboré

## Thématiques de conférences



The screenshot shows the Treecloud application window. The title bar reads "Treecloud". The interface is divided into several sections:

- Distance:** A list of radio buttons for distance metrics: liddell, gmean, jaccard, dice, ms, zscore, hyperlex (selected), chisquared, poissonstirling, loglik.elikhood, oddsratio, ngd, and mi.
- Fenêtre glissante:** A section for window settings including "Taille:" (empty), "Déplacement:" (empty), "Séparateur:" (aaaaa), and "Couleurs:" with radio buttons for yahoo, berry, chronology (selected), dispersion, chronodisp, and mot: cible.
- Longueurs d'arêtes:** Radio buttons for unitaires (selected) and réelles.
- Main Content Area:**
  - Text: "Ce programme est une interface graphique pour Treecloud, et permet de construire le nuage arboré d'un texte. Téléchargez Treecloud, ainsi que son code source Python, sur <http://www.treecloud.org>."
  - Emplacement de Python: C:\Python26\python.exe
  - Emplacement de SplitsTree: C:\Program Files\SplitsTree\SplitsTree.exe
  - Texte à visualiser: A list of text sources including "aaaaa 2004", "aaaaa Conversation text types: A multi-dimensional analysis", "aaaaa Statistical Analysis of Text in Educational Measurement", "aaaaa Étude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de rééchantillonnage", "aaaaa Text Retrieval with External Information", "aaaaa Clustering Algorithms for Noun Phrase Coreference Resolution", and "aaaaa Quantification Of Stylistic Traits: A Statistical Approach".
  - Buttons for "Mots du nuage arboré:", "Tailles des mots:", and "Couleurs des mots:" each with an "Ouvrir" (Open) button.
  - Antidictionnaire: C:\Sites\GambetteLirmm\TreecloudDistribution\StoplistJadt.txt
  - Langue: Français (dropdown menu)
  - Nombre de mots du nuage arboré: [empty] ou nombre minimal d'occurrences pour apparaître dans le nuage arboré: 5
  - Ligne de commande: "C:\Python26\python.exe" "C:\Sites\GambetteLirmm\TreecloudDistribution\Treecloud.py" stoplist="C:\Sites\GambetteLirmm\TreecloudDistribution\StoplistJadt.txt" splittreepath="C:\Program Files\SplitsTree\SplitsTree.exe" unit=1 minnb=5 sephar=aaaaa distance=hyperlex color=chronology "C:\TreeCloud\jadt.txt"
  - Bottom button: "Calcule le nuage arboré avec TreeCloud!" with a small tree icon.

