Outils d’analyse de données textuelles non structurées

9 Décembre 2021

Mathilde de Saint Leger

Chef de projet à Modyco (www.modyco.fr)

mdesaintleger@parisnanterre.fr

Plan

- Pourquoi des outils ?
- Analyse de données textuelles guidée : comment ?
- Divers approches d’analyse, outils associés
- Conclusion

Pourquoi des outils ?

- Pour rechercher, extraire et lier de l'information utile et inconnue ou « cachée » dans de gros volumes de données stockées
- Pour valoriser l'information non structurée

Comment ?

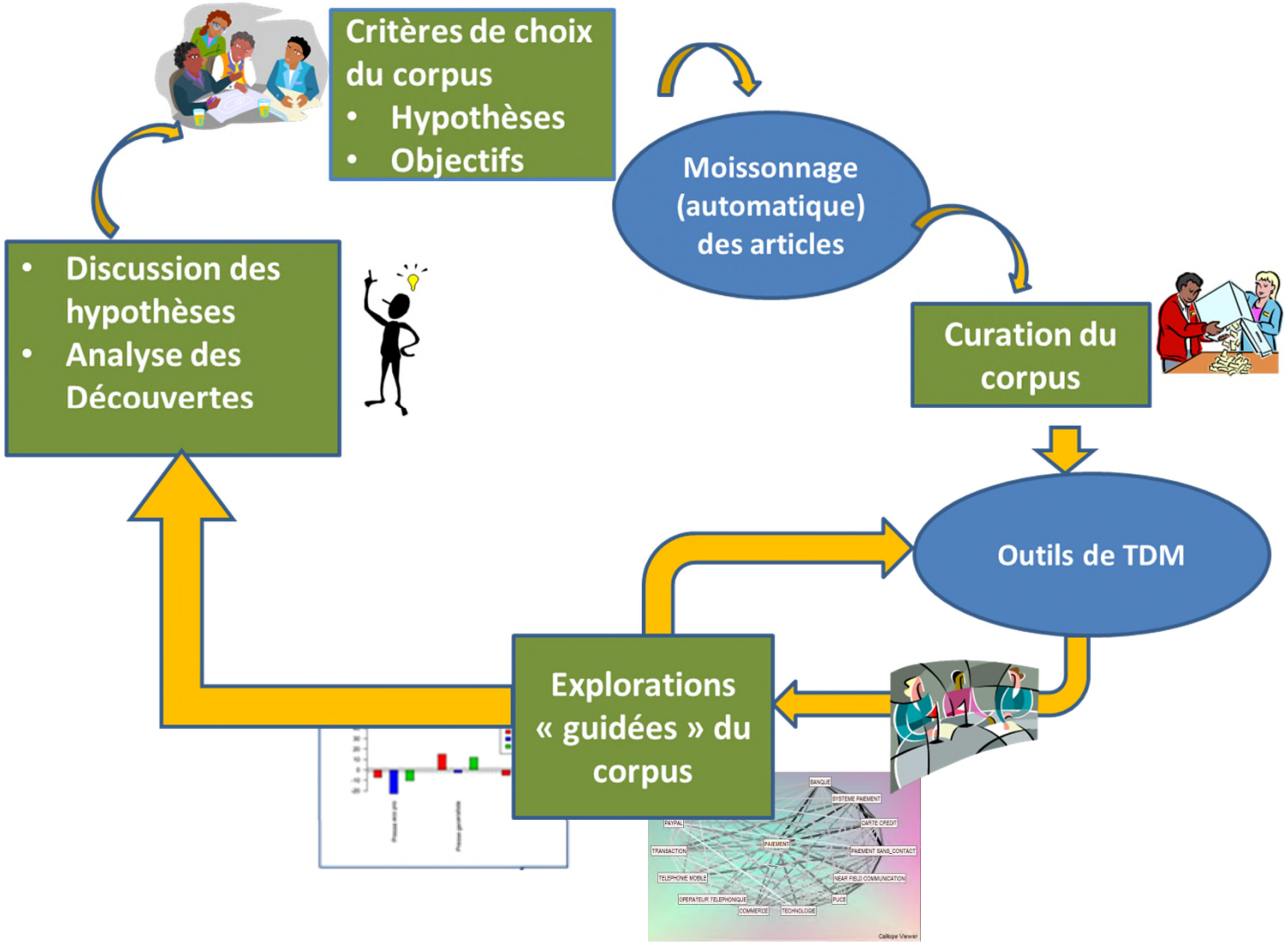
Il faut rendre des données textuelles "calculables" pour faire des opérations de tri, de classification, de regroupement, etc.

Pourquoi la fouille de données textuelles ?

- Efficacité,
- Développement de nouvelles connaissances,
- Analyse thématique d'un domaine,
- Vérification d'hypothèses de travail,
- Détection de thèmes émergents, de controverses,

Transformer les données en renseignements ou connaissances utiles

Démarche dans une analyse de données textuelles



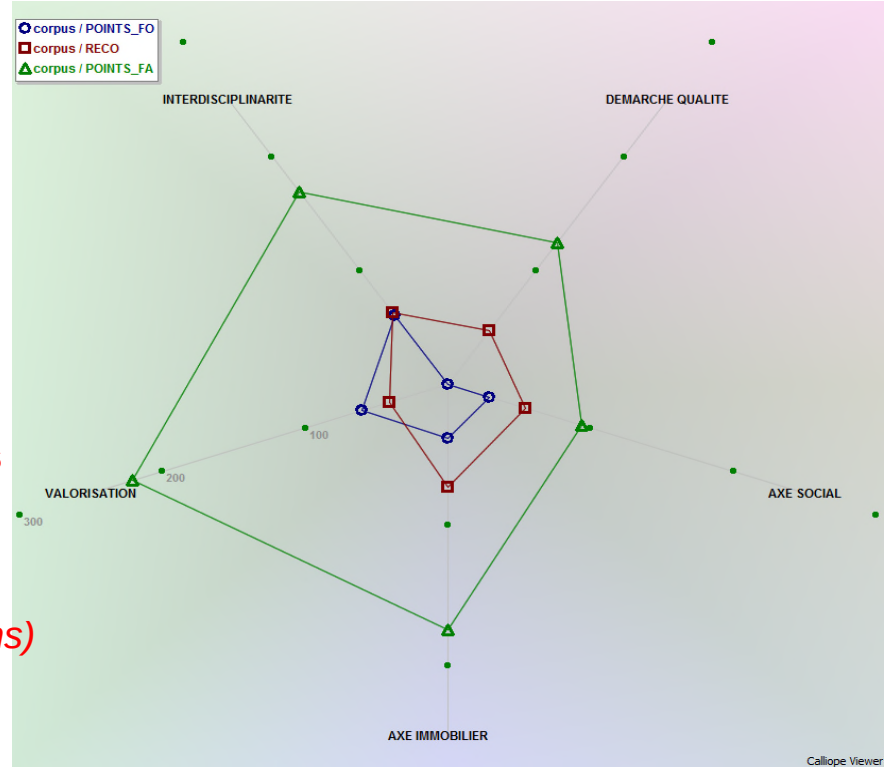
Analyse de données textuelles, comment ?

- Permettre de définir facilement des relations entre différents contenus et exploiter à « à la volée » ces relations
- Prétraitements appliqués aux textes pour les structurer (segmentation, lemmatisation, etc.)
- Représenter les textes (cartographies thématiques interactives, réseaux d'acteurs, etc.)
- Analyser

Analyse de données textuelles, comment ?

Retour en contexte

« Favoriser les partenariats publics et privés pour être en mesure de développer la recherche »
(extrait des Recommandations)



« Une opacité des procédures d'attribution de moyens sans réel pilotage, à l'exception des ATM et des affaires immobilières »
(extrait des Points faibles)

Panorama non exhaustif des méthodes

Approche sémantique

- S’appuyer sur des référentiels spécifiques au domaine du corpus (thésaurus, ontologie, etc.)
- Traitement morpho-syntaxique
- L’outil enrichit les documents du corpus
- Les éléments présents dans chaque texte sont normalisés pour faciliter la tâche d’analyse

Panorama non exhaustif des méthodes

Approche sémantique

- Coller à la spécificité du corpus documentaire
- Modéliser des connaissances métiers
- Coûts d'exploitation et de maintenance élevés

Résultats pertinents obtenus rapidement une fois le système devenu opérationnel

Palteforme Luxid (?) *(Temis)* ***Tropes*** *(Acetic ?)*

Panorama non exhaustif des méthodes

Approche textométrique

- Techniques statistiques pour l'étude des occurrences de motifs linguistiques dans un corpus
- Proposer des parcours de lecture sur un corpus par l’observation de la fréquence et de la disposition des motifs linguistiques
- Contextualiser ces motifs

Panorama non exhaustif des méthodes

Approche textométrique

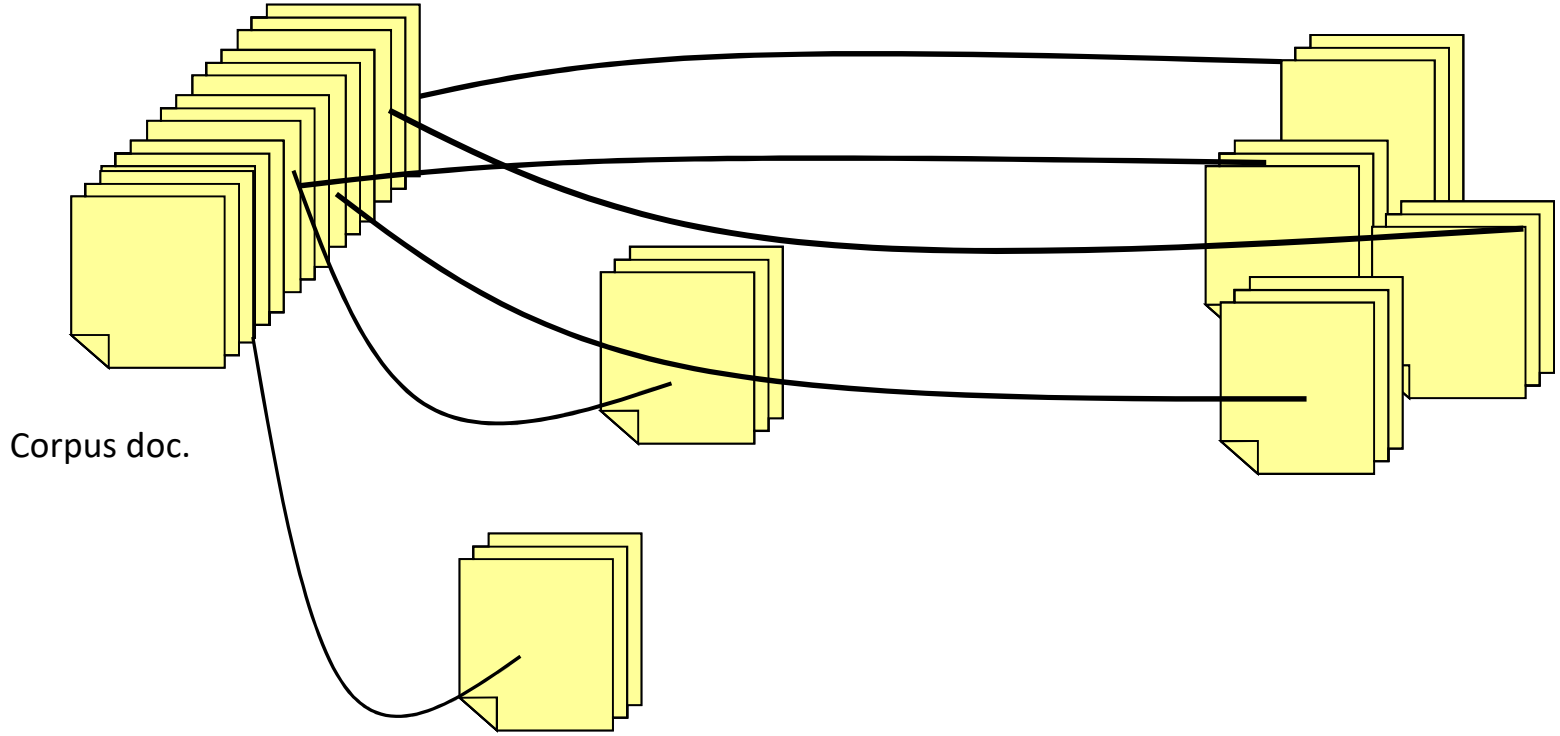
- Pas de synthèse du contenu du corpus
- Pas de calcul automatique du sens

Réorganisations synthétiques, sélectives et suggestives, des textes soumis à l'analyse

TXM, HyperBase, Lexico, Weblex

Panorama non exhaustif des méthodes

Approche classification automatique



Application de méthodes statistiques multidimensionnelles (AFC, ACP, clustering)

Panorama non exhaustif des méthodes

Approche classification automatique

- Représentation des classes obtenues par une cartographie, un dendogramme…
- Positionnement des classes/clusters les uns par rapports aux autres
- Méthode statistique, il est donc indispensable d'avoir un corpus de taille significative (au moins une centaine d'articles)
- Pas d'analyse de tonalités

Faire une synthèse du contenu du corpus en regroupant les documents thématiquement homogènes ensemble

Calliope, Iramuteq, …

Panorama non exhaustif des méthodes

Approche « classifieur »

Attribuer automatiquement des textes du corpus à des catégories prédéfinies

Le travail d'analyse est fait en amont pour construire le classifieur

Un classifieur est adapté à des corpus qui « se ressemblent »

Analyseur de spam, Analyseur d'opinion, Analyseur de l'impact de communication…

Panorama non exhaustif des méthodes

Approche de données textuelles structurées « datavisualisation »

Représenter des données de façon visuelle avec des algorithme s de spatialisation

Corpus d'objets / Relations

Donner un sens aux interactions

Détecter des communautés, détecter des complémentarités, détecter des classements…

Wordle, Gephi, VoisViewer, GraphViz…

http://en.wikipedia.org/wiki/Social_network_analysis_software