Sujet de projet tuteuré de M1 MIAGE UNC 2021-2022

Analyse de cartes cognitives sur la représentation sociale de la mine de nickel en Nouvelle-Calédonie

Valentine BOUDJEMA (IRD) Frédéric FLOUVAT (UNC/ISEA)
Catherine SABINOT (IRD) Romuald THION (UNC/ISEA)

— Tuteur : Romuald THION (UNC/ISEA) romuald.thion@unc.nc
— Experts : Valentine BOUDJEMA (IRD) valentine.boudjema@ird.fr, Frédéric FLOUVAT (UNC/ISEA) frederic.flouvat@unc.nc, Catherine SABINOT (IRD) catherine.sabinot@ird.fr
— Site Web du Projet : http://deployeur.univ-nc.nc/cnrt_MineTerritoires/
— Projet GitHub sur les cartes cognitives: https://romulusfr.github.io/cnrt_cartes_cog/
— Mot-clefs : sciences des données, visualisation graphique de données, traitement de la langue naturelle, cartes cognitives sociologie, anthropologie.

---

1 Contexte

Le projet CNRT *Mine et Territoires* a pour objectif de *recenser, consolider et cartographier les informations utiles des données socio-économiques et géographiques* afin d'établir une typologie des formes d'influence de la mine sur les territoires et les habitants pour ensuite *analyser et modéliser ces phénomènes*. La mine étant ici entendue largement comme l'industrie minière et métallurgique dans son ensemble.

Dans ce cadre, les partenaires de l'IRD en Nouvelle-Calédonie, ont enquêté les populations pour constituer des **cartes cognitives** sur la représentation sociale de la mine. Les cartes cognitives sont **des listes ordonnées de mots** librement énoncées par les répondants. Ces enquêtes ont permis de retenir, à ce jour, 404 réponses à chacune des deux questions suivantes :

— "quelle est votre perception actuelle de la mine",
— "quelle est votre perception de la mine dans les années à venir".

Les partenaires ont également produit *un thésaurus*, qui associe à chaque mot *énoncé* un mot *concept*, à chaque mot *concept* un mot *mère* et à chaque mot *mère* un mot *grand-mère*. Ce thésaurus munit ainsi le corpus des cartes d'une structure arborescente en quatre niveaux.

Par exemple, à la question "quelle est votre la perception actuelle de la mine", la répondante numéro 19 a répondu *dans cet ordre* les sept mots "économie", "destruction", "pollution", "évolution", "civilisation", "dépendance" et "santé". Dans le thésaurus, le mot énoncé "destruction" est associé au mot *concept* éponyme "démolition", qui est associé au mot *mère* "impact environnemental" lui-même associé au mot *grand-mère* "dégradation environnementale".

Une base logicielle a été développée pour automatiser le traitement des cartes cognitives (par exemple, les calculs d'occurrences, de positions ou le remplacement des mots par ceux des niveaux supérieurs du thésaurus) et pour développer les aspects exploratoires suivants :

— identifier et évaluer les **représentations graphiques** susceptibles de représenter les cartes et le thésaurus pour en faciliter l'analyse par les expert-e-s;
— calculer et représenter graphiquement les **cooccurrences** de mots, pour pouvoir analyser les mots qui apparaissent ensemble, en prenant en compte leur distance dans les énonciations;
— analyser les cartes cognitives en les **segmentant selon les profils des répondant-e-s** (âge, proximité professionnelle à la mine, commune de résidence, durée de présence sur le territoire).

Le développement est publiquement accessible à l'adresse https://romulusfr.github.io/cnrt_cartes_cog/. Sur cette page sont notamment proposées les différentes représentations graphiques des cartes cognitives et du thésaurus. Le lecteur est invité à les consulter pour se faire une idée plus précise du sujet.

Notons enfin que les partenaires continuent d'étoffer le fond de cartes cognitives et prévoient à court terme de considérablement réorganiser le thésaurus.

2 Travail à réaliser

Le présent projet tuteuré est **à vocation exploratoire** car il consiste en la recherche et la réalisation de preuves de concepts pour essentiellement :

1. assister les experts dans l'élaboration de nouvelles versions du thésaurus avec en particulier l'apport du Traitement Automatique des Langues Naturelles (TALN);
2. intégrer les visualisations graphiques existantes et en proposer de nouvelles, notamment en prenant en compte les profils des enquêté-e-s;
3. extraire des connaissances des cartes avec des méthodes de statistiques exploratoires, de fouilles voire d'apprentissage.

Le projet tuteuré sera jalonné par la nouvelle version du thésaurus. Ainsi, une première phase sera plutôt consacrée au point 1. La seconde phase sera quant à elle plutôt consacrée aux points 2. et 3.

Les deux axes de travail sur la première phase sont :

— l'utilisation d'outils d'analyse des sentiments dans la langue naturelle pour déterminer automatiquement si le champ lexical des cartes a une connotation méliorative ou péjorative. Les outils comme NLTK proposent de tels algorithmes https://www.nltk.org/howto/sentiment.html
— l'exploitation de bases linguistiques comme https://wordnet.princeton.edu/, notamment les relations d'hyperonymie et d'hyponymie pour regrouper automatiquement les mots et constituer un thésaurus sans intervention humaine auquel les experts pourront se comparer.

Quant à la seconde phase, concernant l'intégration des différentes visualisations graphiques et des données des enquêté-e-s, le développement d'une interface Web ou Python utilisant, en plus de ceux déjà utilisés, des outils d'exploration et de visualisation de jeux de données multidimensionnels comme https://github.com/crossfilter/crossfilter sont envisagés.

Enfin, l'extraction de connaissances est très ouverte : les étudiant-e-s seront forces de propositions pour sélectionner les outils et les utiliser. Des travaux sur l'utilisation de techniques comme l'analyse des correspondances multiples déjà réalisés pourront être repris et développés, mais des méthodes de *clustering* ou d'apprentissage pourront être proposées.

3 Outils et méthodes

Au surplus des éléments techniques cités précédemment, les étudiant-e-s utiliseront tout ou partie des outils suivants, la majorité déjà employée sur https://romulusfr.github.io/cnrt_cartes_cog/ :

- GitHub https://github.com/ pour partager le code, gérer ses différentes versions, gérer la documentation et également de générer automatiquement des pages Web à partir des documents versionnés.
- l'écosystème libre du langage de programmation Python https://www.python.org/, dont en particulier les bibliothèques :
  - Pandas https://pandas.pydata.org/ pour la gestion de données tabulaires;
  - Seaborn https://seaborn.pydata.org/ pour la visualisation graphique;
  - NetworkX https://networkx.org/ pour gestion de graphes;
  - NumPy https://numpy.org/ et SciPy https://scipy.org/ pour les calculs performants;
  - Scikit-learn https://scikit-learn.org/stable/ pour la fouille de données et l'apprentissage automatique;
  - Statsmodels https://www.statsmodels.org/stable/index.html pour les statistiques;
  - NLTK https://www.nltk.org/ ou Spacy https://spacy.io/ pour le TALN.
- l'écosystème de visualisation graphique https://vega.github.io/vega/ qui permet de produire des représentations graphiques Web riches et interactives à partir d'une spécification déclarative et éventuellement son interface Python nommée Altair https://altair-viz.github.io/.

4 Profil des candidat-e-s

Ce stage motivera les candidat-e-s intéressé-e-s par la science des données, le traitement automatique de la langue naturelle et la visualisation graphique de données. Il motivera aussi celles et ceux qui souhaitent découvrir le travail dans une équipe pluridisciplinaire, puisque nous serons amenés à faire des réunions avec les autres partenaires du projet, aussi bien informaticien-ne-s que géographes ou anthropologues.