Représentations vectorielles de corpus collaboratifs sur la ville de Paris

Carmen Brando*, Catherine Dominguès**

*EHESS, 190-198 Avenue de France, Paris
carmen.brando@ehess.fr
**Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, Saint Mandé
catherine.domingues@ign.fr

Ce travail vise à analyser un corpus collaboratif produit dans le contexte d'appels à contributions menés par la Mairie de Paris. Ces contributions sont reçues sur la plateforme en ligne "Madame la maire, j'ai une idée" ; ces contributions sont sollicitées par la ville qui souhaite recueillir les opinions des Parisiens sur certaines thématiques. Ces derniers déposent des témoignages, des propositions sur leur environnement urbain ainsi que des demandes sur des projets d'aménagement qu'ils souhaitent voir entrepris par la mairie. Ce corpus que nous appelons dorénavant le *corpus mairie de Paris* est composé de quatre sous-corpus thématiques : 1) reconquête des berges de Seine, sous-corpus *berges*, 150 contributions, 18 244 mots, 2) construisons notre métropole, sous-corpus *métropole*, 342 contributions, 63 357 mots, 3) budget participatif, sous-corpus *budget*, 5116 contributions, 1 002 085 mots, et 4) réinventons nos places, sous-corpus places, 338 contributions, 69 953 mots. Ce dernier est lui-même divisé en sept sous-corpus dédiés à sept places parisiennes choisies par la mairie, dont la place de la Bastille, la place Gambetta, la place d'Italie, la place de la Nation.

Pour explorer ces corpus, nous employons des représentations vectorielles de mots, en particulier le modèle de réseau neuronal *word2vec* (Mikolov et al., 2013). ll s'agit de construire la représentation des mots d'un texte dans un espace vectoriel de grande dimension ; un mot est modélisé en tant que vecteur de nombres réels et des mots apparaissant dans des contextes similaires seront représentés par des vecteurs plus proches que d'autres mots apparaissant dans des contextes différents. Cette représentation permet de calculer la ressemblance entre mots par la similarité cosinus et de retrouver, par conséquent, des relations sémantiques entre les mots. Il est également possible d'effectuer des opérations arithmétiques sur les vecteurs (par exemple, la soustraction) pour retrouver d'autres types de relations (Levy et Goldberg, 2014).

Les représentations vectorielles de mots remportent des succès dans de nombreuses tâches de traitement du langage naturel et plus récemment, en recherche d'information. En effet, Despres et al. (2016) proposent de représenter le contexte d'un document par un vecteur de mots qui sera utilisé en tant que modèle de langues pour améliorer la recherche dans des documents quand on s'intéresse à des documents multilingues. En outre, Muchemi et Grefenstette (2016) utilisent ces modèles pour produire un vocabulaire spécifique à la thématique d'un corpus (ex : l'astronomie) sous forme de taxonomie, sans avoir à contraster le corpus avec un corpus généraliste déjà constitué, par exemple un texte encyclopédique. Dans ce travail, nous utilisons *word2vec* afin de retrouver, par similarité sémantique, des mots qui désignent des besoins ou des préoccupations inattendus des citoyens (autrement dit, des sujets qui n'ont pas été