Défi EGC 2020 : Visualisation et embeddings de mots pour détecter les thématiques récentes

Philippe Suignard*

*EDF R&D, 7 boulevard Gaspard Monge, 91120 Palaiseau
philippe.suignard@edf.fr

Résumé. Dans cet article, nous montrons comment des techniques de type word et document embeddings, appris sur le corpus des articles de la conférence EGC, associés à une technique de visualisation de graphe, permettent de détecter les thématiques en émergence de cette conférence.

1 Introduction

Pour sa 20ème édition qui aura lieu en janvier 2020, EGC organise un défi, dont l'objectif « est de faire le bilan de l'évolution de la communauté EGC ces 20 dernières années et tenter d'en prédire l'avenir. Le principe est d'appliquer des techniques d'extraction et de gestion de connaissances afin d'expliquer la structure et l'évolution de l'ensemble des données au fil des années ». Il se trouve qu'un précédent défi (en 2016) avait été organisé sur le même sujet. Nous avons pris le parti de regarder ce qu'il y avait de nouveau depuis ce moment.

Les techniques d'embeddings de mots ont connu beaucoup de succès ces dernières années (Mikolov et al., 2013), mais sont généralement entraînées sur des corpus de données très volumineux. Nous voulons montrer ici, que cette technique peut être utilisée et entraînée sur des corpus de faible taille comme c'est le cas pour ce défi constitué d'un corpus d'un millier de documents environ.

2 Les données

Les données fournies par EGC sont constituées des titres des articles issus de la conférence EGC, de leur résumé, de leurs auteurs, d'un lien vers la 1ère page de l'article et d'un lien vers l'article (en PDF). Une partie des données est rédigée en anglais : titre et/ou résumé soit au final 1269 documents répartis sur 15 ans (de 2004 à 2018). Pour ne pas perturber les traitements, les parties en anglais sont éliminées, comme l'ont fait (Guille et al., 2016). Pour nos traitements, un document sera constitué par la concaténation du titre et du résumé (en ne gardant que ce qui est rédigé en français). Au final, nous disposons de **1179** documents.