Temporal Top2vec

Evolution temporelle des thématiques scientifiques : clustering dynamique des documents

Contacts

- Hubert Naacke hubert.naacke@lip6.fr, Hamed Rahimi: hamed.rahimi@lip6.fr

Contexte

Ce projet s'inscrit dans la thèse d'Hamed Rahimi intitulée Sémantisation de corpus scientifiques à large échelle - Application à l'analyse interactive de l'évolution des sciences [1]. Le défi scientifique abordé dans ce projet est de mesurer le changement d'un thème (ou topic) à travers plusieurs périodes temporelles. Par exemple on veut observer l'évolution du topic le plus proche du mot climat en une dizaine d'année. En 2010 le topic contient principalement les termes « réchauffement » et « giec » tandis qu'en 2022 il contient principalement les termes « bilan carbone » et « migration ». Au-delà de cet exemple, le besoin est fort de mieux comprendre quelles interactions entre topics peuvent donner naissance à de futurs domaines scientifiques.

Le stage s'appuie sur Top2vec [2], une approche statistique récente pour traiter des documents textuels. Top2vec repose sur (a) un modèle de langage qui associe un vecteur à chaque document et chaque mot du vocabulaire ; (b) une réduction de la dimension des vecteurs de documents de 300 à 5. (c) un clustering des documents pour former des topics. Toutefois, Top2Vec ne tient pas compte de la date de publication des documents : un topic peut contenir des documents publiés en des années très différentes, ce qui ne permet pas de détecter si un topic change au cours du temps. L'objectif est donc d'améliorer Top2Vec pour « temporaliser » les topics sur plusieurs tranches de temps consécutives.

Objectif du projet : Clustering dynamique des documents

En 2022 (cf stage) nous avons adapté Top2vec pour tenir compte de l'année de publication des documents. On découpe l'ensemble des documents par **périodes** glissantes de 3 ans avec une année en commun entre 2 périodes consécutives. Puis on applique l'algorithme **Aligned**UMAP [3] pour réduire la dimension des documents dans toutes les périodes de manière cohérente : c'est-à-dire, les documents présents dans deux périodes consécutives ont les mêmes coordonnées. L'étape suivante est d'appliquer un clustering dans chaque tranche de temps pour former des topics. Pour cela, on veut garantir que chaque topic de la 1ère tranche pourra être « aligné » avec un topic de la 2ème tranche, et ainsi de suite.

L'objectif du projet PLDAC est de concevoir une solution de clustering « dynamique ». Le clustering doit se baser sur les clusters de la tranche précédente pour calculer, dans la tranche courante, des clusters qui soient alignés avec ceux calculés de la tranche précédente. On veut que la précision de l'alignement entre deux clusters soit maximisée.

Parmi les pistes possibles, vous pourrez définir la précision d'un lien entre deux topics, ajuster un algorithme de clustering existant tel que DBSCAN. Vous pourrez aussi proposer une méthode pour affecter un label (les mots les plus représentatifs) à un topic pour pouvoir expliquer l'évolution des topics.

La dernière phase du projet sera de valider la solution en mesurant la qualité des topics obtenus à chaque période (par exemple en termes de diversité/cohérence des topics). Vous pourrez aussi comparer votre solutions avec des approches connexes.

Selon le temps restant, une visualisation « temporelle » des topics et de leurs labels pourrait être réalisée (par exemple avec le framework python Dash).

Références:

[1] Thèse d'Hamed Rahimi démarée en 2021 : Sémantisation de corpus scientifiques à large échelle - Application à l'analyse interactive de l'évolution des sciences https://www.lip6.fr/actualite/personnes-fiche.php?ident=D2534
[2] top2vec : https://top2vec.readthedocs.io/en/latest/Top2Vec.html#usage
[3] AlignedUMAP https://umap-learn.readthedocs.io/en/latest/aligned_umap_basic_usage.html