Etude comparative de modèles d'extraction non supervisée de mots-clés pour la recommandation d'emploi

Bissan Audeh*, Maia Sutter**, Christine Largeron**

* Inasoft, 2507 avenue de l'Europe, 69140, Rillieux-La-Pape, France
bissan.audeh@inasoft.fr,
** Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France
maia.sutter@etu.univ-st-etienne.fr, chistine.largeron@@univ-st-etienne.fr

1 Introduction

La recherche d'emploi peut être facilitée en augmentant la visibilité des postes qui correspondent le mieux à un candidat. Dans un tel contexte, il est essentiel d'attribuer des mots-clés aux offres d'emploi et aux CV pour réaliser l'indexation et la mise en correspondance mais aussi effectuer des analyses intéressantes pour la prise de décision quant au marché de l'emploi. Dans cet article, nous cherchons à évaluer dans quelle mesure les mots-clés extraits avec des approches non supervisées peuvent représenter du contenu textuel même sans apprentissage et sans connaissance externe dans un contexte de recommandation d'emploi. Nous avons sélectionné six méthodes non supervisées d'extraction de mots-clés que nous avons évaluées sur de vrais offres d'emploi et CV anonymisés. Pour cela nous avons élaboré un protocole d'évaluation qui inclut la construction d'un gold standard.

2 Cadre méthodologique

Nous avons évalué quatre modèles parmi les plus courants dans l'état de l'art : deux modèles statistiques TF-IDF et RAKE (Rose et al. (2010)), TextRank (Mihalcea et Tarau (2004)) un modèle basé sur le score PageRank et KeyBERT (Grootendorst (2020)) qui utilise les plongements de texte. En plus, nous avons évalué les extensions KeyBERT+ et TF-IDF+ des modèles KeyBERT et TF-IDF respectivement, où nous remplaçons la sélection de mots-clés candidats des méthodes par une sélection basée sur le modèle de langue français de Spacy [1] qui permet d'identifier des termes composés. Nous avons adapté toutes ces méthodes à notre corpus par des pré-traitements et post-traitements. Les données disponibles pour ce projet étaient constituées de 818 CV et 858 offres d'emploi mais seul un sous-ensemble a été annoté en vérifiant manuellement 12316 mots-clés ; ce qui a permis de construire un jeu d'évaluation contenant 57 CV et 29 offres d'emploi. Le texte des CV a déjà été extrait et anonymisé en supprimant les données personnelles, et les offres d'emploi étaient au format HTML. Bien que la taille de ce jeu soit limitée, l'expérimentation a néanmoins permis de comparer les performances des algorithmes.

1. https ://spacy.io/models/fr#fr_core_news_lg

3 Résultats et Discussion

Pour chaque modèle évalué, la précision, le rappel et le F-score sont calculés pour chaque document en comparant des mots-clés validés manuellement aux dix meilleurs mots clés retrouvés par le modèle. Bien que les scores présentés dans le tableau 1 soient relativement faibles, ils confirment les résultats fréquemment rapportés dans la littérature consacrée à l'extraction des mots-clés. Parmi les approches testées, RAKE et TF-IDF+ surpassent les autres en précision pour les CV et les offres d'emplois. Alors que TF-IDF+ a besoin de statistiques au niveau du corpus pour calculer les scores des mots-clés, RAKE pourrait être un choix plus approprié pour ce contexte applicatif. Il faut noter que KeyBERT et KeyBERT+ ont été utilisées avec des plongements de texte pré-entraînés sur des corpus génériques qui ne sont pas composés de CV et offres d'emploi. Des expériences supplémentaires, notamment l'ajustement de ces modèles à notre contexte, pourraient apporter des améliorations. Une difficulté majeure dans ce travail était la subjectivité de l'évaluation, qui rend difficile l'interprétation des résultats. Une solution possible consiste à avoir plusieurs annotateurs travaillant sur le même document pour pouvoir établir un consensus. Par ailleurs, notons que comme les approches sont à l'origine développées et testées avec du texte en anglais, elles ne tiennent pas toujours compte des spécificités de la langue française, comme l'utilisation de parenthèses en français pour ajouter le «e» pour la forme féminine. Enfin, si cette étude permet d'éclairer les options d'extraction de mots-clés de CV et d'offres d'emploi, elle ne constitue qu'une étape dans l'élaboration d'un système de recommandation d'emploi qui nécessite d'aller plus loin.

		CV			Offres d'emploi		
	Modèle	Avg.Préc	Avg.Rappel	F-Score	Avg.Préc	Avg.Rappel	F-Score
@10	TF-IDF	0.193	0.096	0.128	0.186	0.127	0.151
	TF-IDF+	0.234	0.126	0.164	**0.193**	**0.130**	**0.155**
	TextRank	0.165	0.071	0.099	0.160	0.093	0.118
	RAKE	**0.299**	**0.141**	**0.192**	0.148	0.101	0.120
	KeyBERT	0.146	0.078	0.102	0.041	0.028	0.034
	KeyBERT+	0.217	0.122	0.156	0.128	0.095	0.109

TAB. 1 – *Résultats de l'évaluation à 10 mots-clés extraits. Meilleurs résultats en gras.*

4 Remerciements

Nous remercions Servan Cazenave, directeur d'Inasoft, pour son soutien à ce projet, son apport sur la connaissance métier, et son aide pour l'annotation.

Références

Grootendorst, M. (2020). Keybert : Minimal keyword extraction with bert. *Internet]. Available : https ://maartengr.github. io/KeyBERT/index.html*.

Mihalcea, R. et P. Tarau (2004). Textrank : Bringing order into text. In *EMNLP*.

Rose, S., D. Engel, N. Cramer, et W. Cowley (2010). Automatic keyword extraction from individual documents. Technical report.