Présentation du jeu de données - Offres d'emploi et compétences du milieu de l'assurance

publié le 5 avril 2022

Institut intelligence et données

Ce jeu de données a été recueilli dans le cadre du projet multidisciplinaire *Femmes face aux défis de la transformation numérique : une étude de cas dans le secteur des assurances* de l'Université Laval, financé par le Centre des compétences futures. Il regroupe des offres d'emploi, en français, de compagnies d'assurance entre 2009 et 2020. Ces données sont dépersonnalisées afin d'éviter la reconnaissance directe des compagnies y ayant contribué. De plus, une partie des offres sont accompagnées d'annotations associées à l'une des quatre catégories de compétences identifiées pour ce projet.

1 Objectifs initiaux

Initialement, l'analyse des offres d'emploi visait les objectifs suivants :

1. Développer et tester une méthode de reconnaissance automatique des compétences dans des offres d'emploi.
2. Permettre, à partir de cette analyse automatique, de mesurer les tendances annuelles dans les compétences demandées dans le domaine de l'assurance.

2 Méthodologie de création du jeu de données

Le projet a été approuvé par les Comités d'éthique de la recherche avec des êtres humains de l'Université Laval. Cependant, pour la collecte des offres d'emplois, aucune approbation éthique n'était nécessaire étant donnée qu'aucune donnée personnelle et confidentielle n'est présente dans ces offres d'emploi.

Les données ont été recueillies auprès des partenaires sous plusieurs formats (e.g. Word et PDF). L'insertion dans le fichier des offres a été réalisé de manière semi-automatique : une partie des données brutes ont été insérées automatiquement l'autre partie a été insérée manuellement. Une fois le prétraitement terminé, un nettoyage des données a été réalisé manuellement au vu du faible nombre d'offres. Parmi ce nettoyage on peut citer la suppression des retours à la ligne en milieu d'une phrase, la suppression de puces et la suppression de saut de ligne multiple à la fin du texte.

L'étiquetage des compétences dans les offres a été réalisé par une étudiante en science de l'éducation sous la supervision d'une professeure membres du Centre de recherche et d'intervention sur l'éducation et la vie au travail (CRIEVAT). Cet étiquetage a seulement été réalisé sur un sous ensemble des offres (47).

La dépersonnalisation des données a été effectuée de manière semi-automatique. Une partie automatique a permis de retirer les éléments les plus simples à identifier [1]. Par la suite, les développeurs

1. Élements tels que les noms des entreprises partenaires, les adresses courriels et tout autre élément pouvant permettre l'identification des dites entreprises (logiciels utilisées, nom d'un service propriétaire, noms de direction/service, etc.)

et les entreprises partenaires ont effectué des vérifications manuelles pour s'assurer que la dépersonnalisation a bien été effectuée.

Le processus de collecte et de création du jeu de données a impliqué la participation d'employés des compagnies d'assurance (salaire horaire inconnu), de professionnels de recherche de l'Institut intelligence et données et d'étudiants (via contrats de recherche) de l'Université Laval (salaire horaire selon les normes syndicales de l'Université Laval en vigueur en 2020 et 2021). Le temps de la collecte des données peut être estimé à 10 jours (7h/jour) de travail sur ordinateur (consommation 120W), soit 8,4 kWh d'énergie, soit environ 4 Kg de CO_2 [2] [3]. Le temps de préparation des données (montage du jeu de données et dépersonnalisation) est évalué à environ 20 jours (7h/jour), soit 16,8 kWh, soit environ 8 Kg de CO_2. Le temps de travail d'étiquetage est estimé à 10 jours, soit 8,4 kWh d'énergie, soit également 4 Kg de CO_2. Ainsi, au total, les émissions de gaz à effet de serre directs pour la création du jeu de données sont estimées à 16 Kg de CO_2. Ce total inclut seulement la collecte et non la rédaction des offres elle-même.

3 Caractéristiques sommaires

Les caractéristiques principales du jeu de données sont les suivantes :

1. Le jeu de données contient 867 offres en français.
2. Le volume des fichiers est de 2 Mo environ.
3. Le jeu de données ne comporte aucune donnée personnelle ni sensible.
4. Le jeu de données contient certaines des offres d'emploi de postes techniques et administratifs de compagnies d'assurance Québécoise pour les années 2009 à 2020.
5. Les données ont été collectées de mai à décembre 2020. Elles ne contiennent donc pas toutes les offres de 2020.
6. Le jeu de données annoté comporte 499 phrases annotées comprenant 932 annotations sur 47 offres différentes.

Quelques biais connus Au vu de notre processus de création du fichier semi-automatique par plusieurs personnes, il est possible que certaines offres ne contiennent pas tout le texte initial de l'offre. De ce fait, toutes les offres fournies dans ce fichier ne sont pas uniformes dans leur contenu. Par exemple, certaines offres contiennent une présentation de l'entreprise dans une colonne distincte alors que d'autres ne contiennent pas cette information de manière distincte.

L'annotation ayant été faite sur des offres sélectionnées aléatoirement, la répartition des années et entreprise n'est pas similaire à celle du jeu de données complet. Certaines années et certaines entreprises sont surreprésentées dans le jeu de données annoté par rapport au jeu de données non annoté.

Des caractéristiques plus détaillées de ce jeu de données sont présentées dans l'article suivant :

> "FIJO" : a French Insurance Soft Skill Detection Dataset, Beauchemin, David and Laumonier, Julien and Le Ster, Yvan and Yassine, Marouane, *Proceedings of the Canadian Conference on Artificial Intelligence, Canadian Artificial Intelligence Association (CAIAC)*, 2022

4 Dictionnaire de données

Offres d'emplois Le fichier des offres d'emplois est au format TSV, décrit par la Table 1. Puisque la date limite de réception candidatures n'était pas toujours dans le même format entre les données, nous avons normalisé celles-ci en ajoutant un jour et mois lorsque ces derniers étaient absents. Autrement dit, le jour et le mois des offres n'est pas une donnée fiable mais l'année est toujours valide. De plus, il est possible que certaines phrases ne soient pas coupées convenablement (retour à la ligne en milieu de phase, titre inclus dans la phrase d'après, etc.) puisque ce processus de nettoyage a été effectué automatiquement sur les données.

2. https://www.hydroquebec.com/data/developpement-durable/pdf/approvisionnements-energetiques-emissions-atmospheriques-2019.pdf
3. https://sustainability.tufts.edu/wp-content/uploads/Computer_calculations.pdf

Colonne	Description
id	Identifiant de l'offre dans le jeu de données pour faire le lien avec les annotations.
Titre du poste	Titre du poste tel que présenté dans l'annonce.
Date limite de réception candidatures	Format JJ-MM-AAAA
Description Interne	Description de l'offre dépersonnalisée en français.
Compagnie	Identifiant dépersonnalisé

TABLE 1 – Description du fichier des offres d'emploi

Annotations Le fichier des annotations est en format JSON. Un champ `_id` permet de faire le lien avec l'identifiant de l'offre complète. Le format du champ `_id` est le suivant : `id_offre.id_phrase`. Étant donnée la nature textuelle des données, les étiquettes correspondent à des morceaux de phrases appartenant à une classe.

Les quatre étiquettes utilisées dans les offres d'emploi sont « Pensée », « Résultat », « Relationnel » et « Personnel ». Le modèle de compétences utilisé, décrit dans la Table 2, est basé sur plusieurs référentiels de compétences publiques, tel que celui de l'association québécoise d'établissement de santé et de service sociaux (AQESSS) [4], mais aussi sur les différents référentiels utilisés par les compagnies d'assurance partenaires, basés eux-mêmes sur des référentiels développés par des entreprises tels que Korn Ferry [5] ou bien SPB [6]. Nous n'avons pas étiqueté les offres de manière plus précise que ces catégories principalement dû au fait que, en général, les algorithmes d'apprentissage sont connus pour ne pas être performants avec un faible ratio nombre d'exemples/nombre d'étiquettes.

Le processus d'étiquetage a démontré que certaines phrases ou morceaux de phrase appartiennent à plusieurs catégories de compétences en même temps. Par exemple, dans l'offre 393, la portion de texte « Répondre aux appels des clients » pourrait être classée comme une compétence de type « Relationnel » (Maîtriser les relations interpersonnelles) mais aussi comme un type « Pensée » (Orientation Client). Cependant, pour des raisons de simplification, il a été décidé que les morceaux étiquetés n'auraient qu'une seule étiquette.

Dépersonnalisation Les étiquettes de dépersonnalisation que l'on peut retrouver dans les offres d'emploi et dans les annotations sont présentées dans la Table 3.

5 Évolution du jeu de données

Une prochaine évolution de ce jeu de données est prévue pour fin 2022 ou début 2023.

6 Licence

Les données sont distribuées sous licence CC-BY-SA-NC 4.0 [7].

4. http://catalogue.iugm.qc.ca/GED_IUG/109311392759/Referentielcomp.pdf
5. https://www.kornferry.com/
6. https://www.spb.ca/
7. https://creativecommons.org/licenses/by-nc-sa/4.0/deed.fr

Type	Compétences
Pensée	Recherche d'information
	Sens de l'innovation et créativité
	Orientation client
	Prendre des décisions de qualité
	Maîtriser la technologie
	Gérer la complexité
	Avoir une bonne connaissance du secteur
Résultats	Orientation vers l'excellence
	Initiative
	Gérer les priorités
Relationnel	Impact et influence
	Écoute
	Collaboration
	Communication interactive
	Maîtriser les relations interpersonnelles
	Gérer les conflits
	Développement des partenariats
Personnel	Souplesse et ouverture
	Inspirer confiance
	Être résilient
	Rigueur et qualité
	Maîtrise de soi
	Imputabilité
	Ouverture à l'apprentissage

TABLE 2 – Modèle de compétences

Étiquettes	Description
`<anon_name>`	Nom d'une personne
`<anon_company>`	Nom d'une des entreprises partenaires
`<anon_misc>`	Élément lié à aux entreprises partenaires (e.g : nom d'un service propriétaire, logiciel...)
`<anon_location>`	Adresse ou emplacement géographique

TABLE 3 – Étiquettes de dépersonnalisation