Collecte et analyse d'offres d'emploi

Claire DE MARICOURT

Soutenance de stage de fin d'études

5 septembre 2018

Plan

Présentation du projet

Problématique

Collecter automatiquement des offres d'emploi sur Internet et les classer par métier

Présentation du projet

- Environnement : la Dares (Direction de l'animation de la recherche, des études et des statistiques)

DARES
direction de l'animation de la recherche, des études et des statistiques

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DU TRAVAIL, DE L'EMPLOI, DE LA FORMATION PROFESSIONNELLE ET DU DIALOGUE SOCIAL

- Département d'acceuil : DMQ (Département Métiers et Qualifications)
- Objectif à terme : révision de l'indicateur de tension

Présentation du projet

- Calcul de l'indicateur de tension

$$tension = \frac{offres}{demandes}$$

- Source de données : Pôle Emploi
- Objectif : étendre le champ des données
- Nécessité : coder les codes ROME de ces offres pour pouvoir les utiliser à des fins statistiques

Nomenclature ROME

K			**SERVICES A LA PERSONNE ET A LA COLLECTIVITE**
K	**21**		**Formation initiale et continue**
K	**21**	**08**	**Enseignement supérieur**
K	21	08	Attaché / Attachée temporaire d'enseignement et de recherche -ATER-
K	21	08	Chargé / Chargée de cours
K	21	08	Chargé / Chargée d'enseignement
K	21	08	Chargé / Chargée d'enseignement du supérieur
K	21	08	Chef de département de l'enseignement supérieur
K	21	08	Enseignant-chercheur / Enseignante-chercheuse
K	21	08	Lecteur / Lectrice de langues dans l'enseignement supérieur
K	21	08	Maître / Maîtresse de conférences
K	21	08	Maître / Maîtresse de langues dans l'enseignement supérieur
K	21	08	Moniteur / Monitrice d'initiation à l'enseignement supérieur
K	21	08	Professeur / Professeure de l'enseignement supérieur
K	21	08	Professeur / Professeure des universités
K	21	08	Responsable filière enseignement supérieur
K	21	08	Responsable matière enseignement supérieur
K	21	08	Responsable programme enseignement supérieur

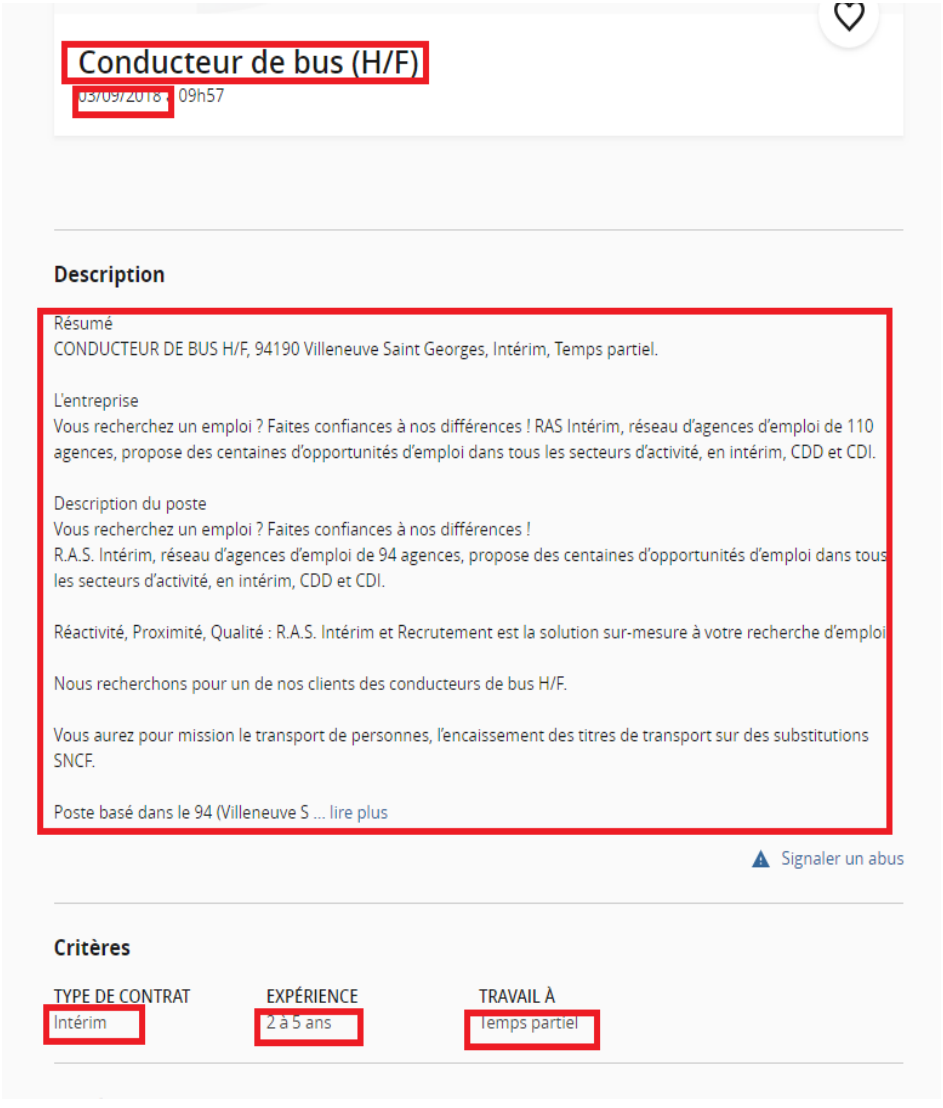
Les grands domaines

Lettre	Grand domaine de Pôle Emploi
A	Agriculture et pêche, espaces naturels et espaces verts, soins aux animaux
B	Art et façonnage d'ouvrages d'art
C	Banque, assurances et immobilier
D	Commerce, vente et grande distribution
E	Communication, media et multimédia
F	Construction, bâtiment et travaux publics
G	Hôtellerie-restauration, tourisme, loisirs et animation
H	Industrie
I	Installation et maintenance
J	Santé
K	Services à la personne et à la collectivité
L	Spectacle
M	Support à l'entreprise
N	Transport et logistique

Collecte de données

- De nombreux sites d'offres d'emploi

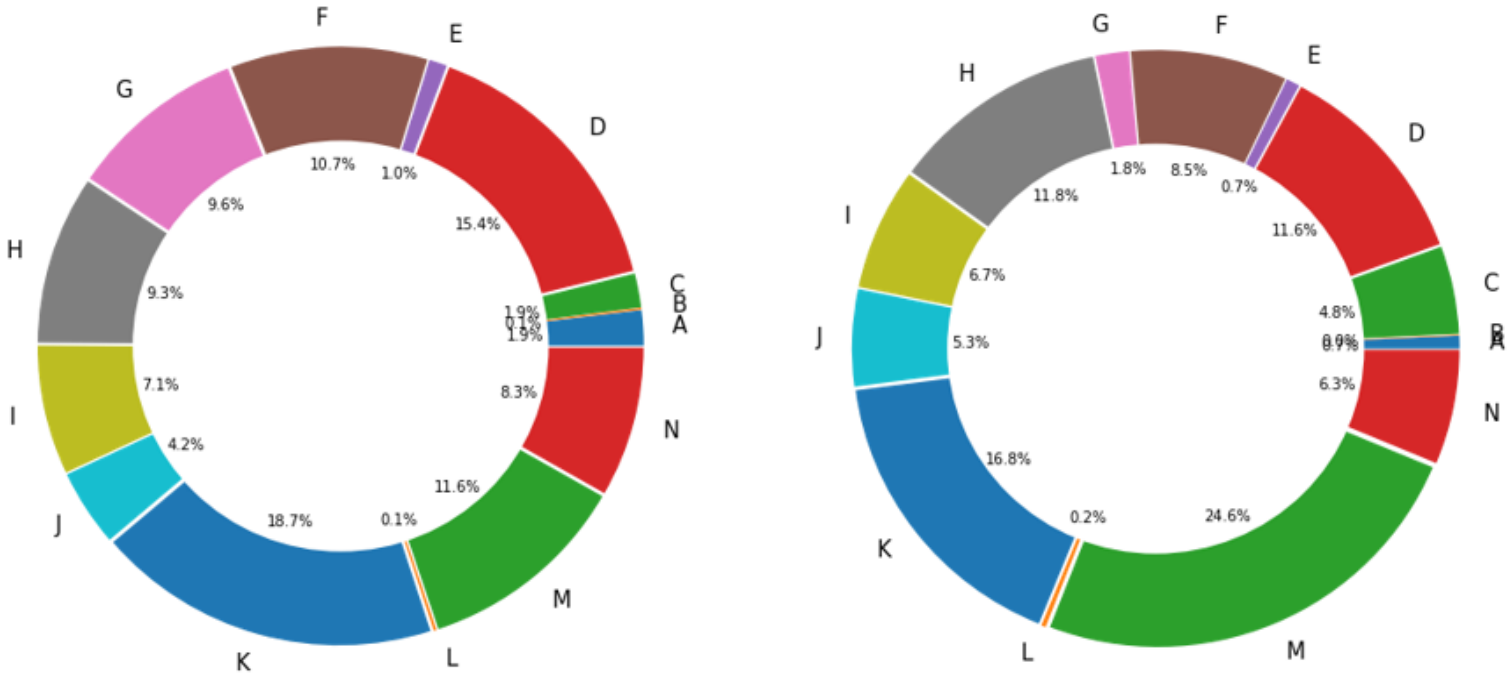
Collecte de données



- Mise à jour du scraping et du parsing
- Ajout de nouveaux sites
- Automatisation de la collecte et gestion du stockage
- Collecte de données étiquetées sur Pôle Emploi

Données Pôle Emploi

- Offres Pôle Emploi étiquetées manuellement
- Offres partenaires étiquetées automatiquement



Répartition des offres Pôle Emploi (G) et partenaires (D) selon les codes ROME

Méthodologie

- Utilisation du titre

Titre $\Longrightarrow$ Titre ROME $\Longrightarrow$ Code ROME $\Longrightarrow$ Grand domaine

employé.e commercial au rayon bazar à lyon (69) cdi h/f $\Longrightarrow$
Employé / Employée de rayon bazar $\Longrightarrow$ D1507 $\Longrightarrow$ D

- Calcul de similarité après nettoyage
  sim(employé commercial rayon bazar; employée de rayon bazar) = 0.87
  sim(employé commercial rayon bazar; boulanger) = 0

Méthodologie - Nettoyage

Étapes du nettoyage	Exemple
Tout mettre en minuscules	employé.e commercial au rayon bazar à lyon (69) cdi h/f
Enlever la ponctuation	employée commercial au rayon bazar à lyon 69 cdi hf
Enlever les chiffres	employée commercial au rayon bazar à lyon cdi hf
Enlever les *stopwords*	employée commercial rayon bazar lyon cdi hf
Enlever certains mots	employée commercial rayon bazar lyon
Enlever le nom de la ville	employée commercial rayon bazar
Lemmatiser	employé commercial rayon bazar

Méthodologie - Appariement par mesure de similarité

Trois mesures de similarité + une

- Jaro-Winkler avec comparaison mot à mot
- Damerau-Levenshtein
- Tf-Idf et similarité cosinus
- Produit de la similarité de Jaro-Winkler mot à mot et de la similarité cosinus

Méthodologie - Appariement par mesure de similarité

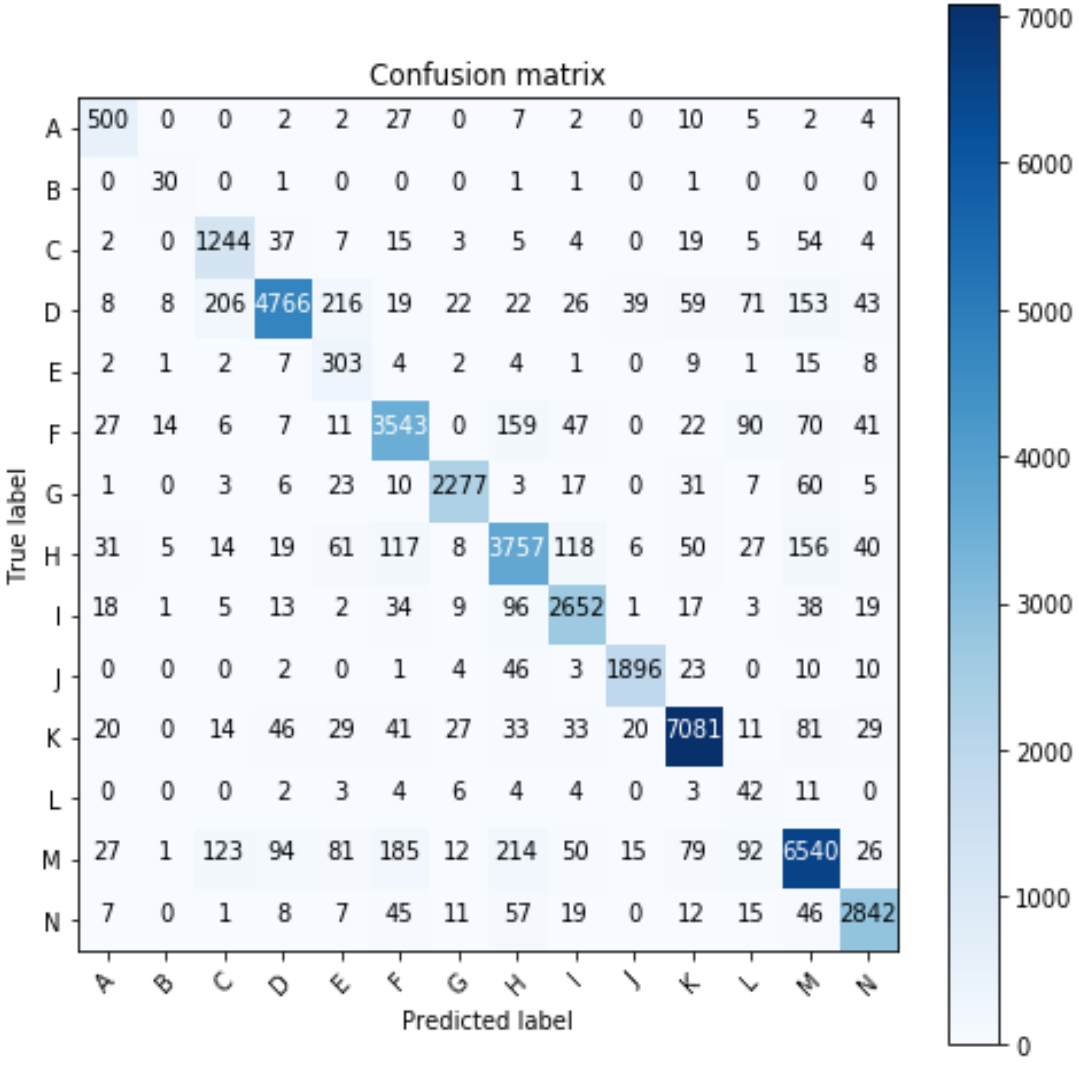
$$tok\_jw\_sim(T_1, T_2) = 0.5 \times partial\_tokens\_jw\_sim(T_1, T_2) + 0.5 \times partial\_tokens\_jw\_sim(T_2, T_1)$$

$$partial\_tokens\_jw\_sim(T_1, T_2) = \frac{\sum_{t^1 \in T_1} (\exists t^2 \in T_2, sim_{jaro-winkler}(t^1, t^2) > 0.9)}{Card(T_1)}$$

Tableau des résultats

	Accuracy	Précision	Rappel	Score F1
Jaro-Winkler mot à mot	89.3%	77.4%	86.8%	80.1%
Pôle Emploi	92.5%	81.2%	88.5%	84.1%
Partenaires	86.0%	72.1%	84.4%	74.8%
Damerau-Levenshtein	88.6%	70.6%	81.7%	73.7%
Pôle Emploi	90.4%	78.2%	85.0%	80.7%
Partenaires	78.2%	62.3%	78.2%	65.6%
Tf-idf et similarité cos	88.6%	81.3%	84.9%	82.8%
Pôle Emploi	93.0%	85.5%	88.1%	86.6%
Partenaires	84.1%	75.5%	81.0%	77.6%
Produit JW/Cosinus	91.1%	84.8%	87.6%	85.6%
Pôle Emploi	93.8%	86.3%	90.0%	87.6%
Partenaires	88.2%	82.7%	84.4%	82.7%

Matrice de confusion



Perspectives et Limites

- Possibilité de définir un seuil de validation de l'étiquetage
- Nouveaux sites à scraper
- Déduplication des offres
- Codage basé sur le corps de l'offre
- Réutilisation de ce codage pour d'autres projets
- Utilisation d'autres informations pour d'autres projets (localisation, compétences, ...)