

Missions avec Bluecoders

Delong LI


Bluecoders


Business

- Chasse de tête
- Mission RPO
- Formation

Problèmes

- Données sous-exploitées (sur domain et technos)
- Mal à match les offres d'emploi et les utilisateurs (candidates)





Delong Li
Data Scientist @Ecole Polytechnique&HEC Paris
Paris, Ile-de-France, France · [Contact info](#)
201 connections
[Open to](#) [Add profile section](#) [More](#)

Show recruiters you're open to work — you control who sees this.
[Get started](#)

Find potential clients by showcasing the services you provide.
[Get started](#)

Experience

Data Scientist
Scopeo · Internship
Sep 2021 - Present · 4 mos
Paris

Data Scientist
Natixis · Internship
May 2020 - Aug 2020 · 4 mos
Paris
· Coded in Python to parse and process 10K+ documents into numeric embeddings through NLP tools including TF-IDF, Doc2vec, BERT, etc. ...see more

Education

HEC HEC Paris
Master's degree, Data Science for Business
2020 - 2021
Courses:
Tableau, Reinforcement Learning, Strategy and Data with BCG, Data Consulting Challenge wit ...see more



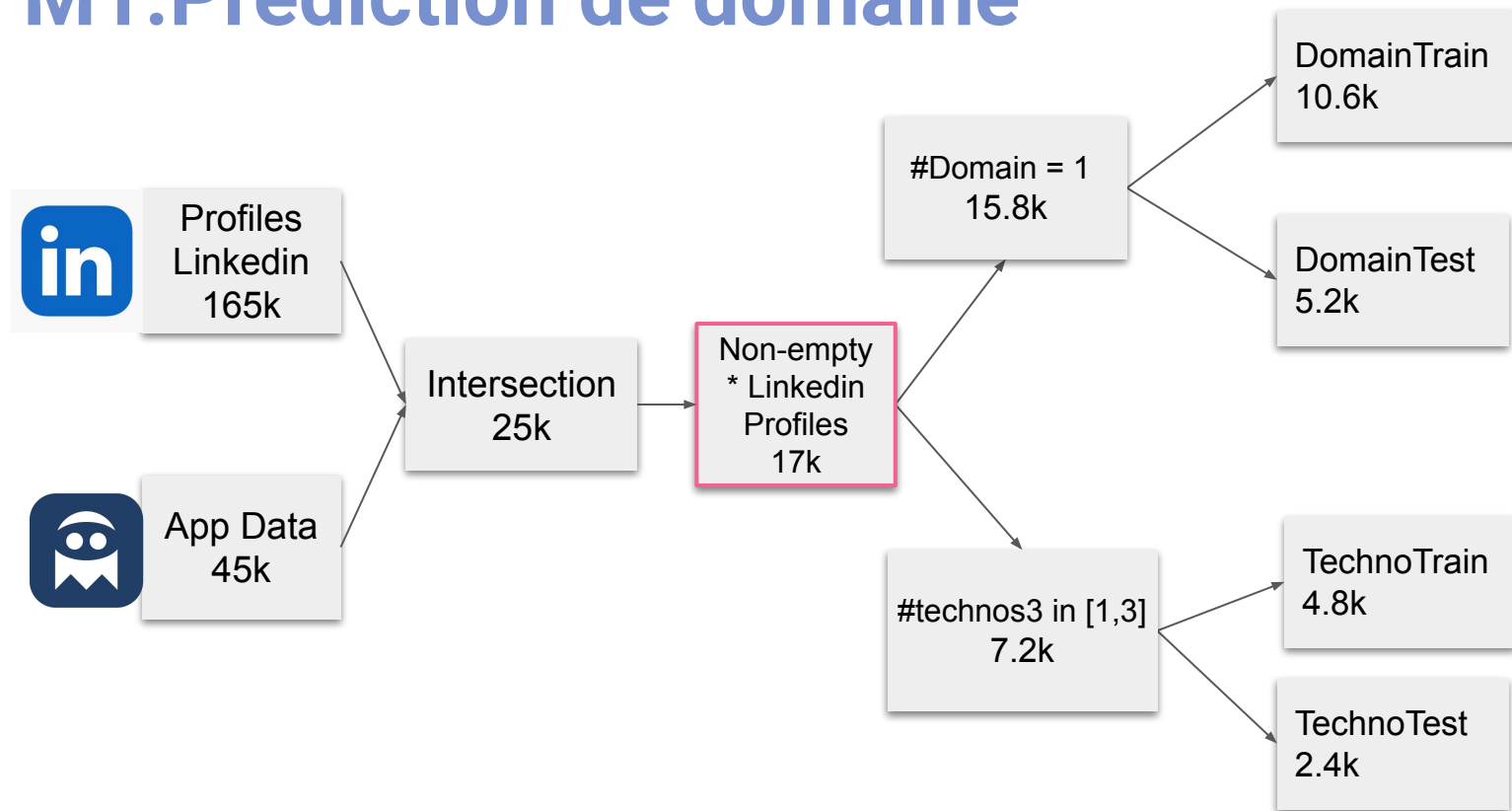
id	name	created_at	domain	technos
XXXX	XXXX	XXXX	}	}

**M1.Prédiction de domaine à partir de
profile Linkedin**

**M2.Recommandation d'entreprise pour
les coders**



M1.Prédiction de domaine



* Au moins un champs parmi [Occupation, Titles, Description, Field of Study] est valid

Valid Intersection

			<div>in</div>						
id	name	Occupation	Title	Experience	Education	domain	technos
XXXX	XXXX	XXX	XXX	XXX	XXX	{“Full-stack”}	{“PHP”}
XXXX	XXXX	XXX	XXX	/	/	{“Software”}	{“Java”}
XXXX	XXXX	/	/	/	XXX	{“Data”}	{“Python”}

Linkedin Data

id	name	Occupation	Title	Experience	Education
XXXX	XXXX	XXX	/	/	/
XXXX	XXXX	/	XXX	XXX	/



domain	technos
?	?
?	?

Stage de 6 mois Développement de plusieurs applications en C++ avec le framework Qt dans un environnement Linux. Adresse email: David@abe.com

1. Text cleaning

Enlever urls, emails, etc.

Convertir les majuscules en minuscule

Ne garder que les lettres, les chiffres et certains caractères a-z0-9âàçéèëîïôùû-+&#

stage de 6 mois développement de plusieurs applications en c++ avec le framework qt dans un environnement linux

2. Tokenization & Stopwords

Tokeniser les phrases en tokens

Remplacer chiffre par token (#number)

Enlever stopwords (le la de du d'...)

[stage, #number ,mois, développement, plusieurs, applicati~~o~~n~~s~~, c++, framework, qt, environnement, linux, adresse, email]

3. Lemmatization

Détecter la langue (français, anglais)

Enlever pluriels, accords, conjugaisons

[stage, #number, mois, développement, plusieurs, applicati~~o~~n, c++, framework, qt, environnement, linux, adresse, email]

4. N-gram

Ajouter 2-grams, 3-grams à la liste

[stage, #number, mois, développement, plusieurs, applicati~~o~~n, c++, framework, qt, environnement, linux, adresse, email
stage #number, #number mois, mois développement, développement plusieurs,...
stage #number mois, #number mois développement, mois développement plusieurs,...]

TF-IDF

Fréquence du terme (TF)

La fréquence « brute » d'un terme est simplement le nombre d'occurrences de ce terme dans le document considéré

Fréquence inverse de document (IDF)

$$idf(t) = \log \left| \frac{n}{\{d \in D: t \in d\}} \right|$$

La fréquence inverse de document est une mesure de l'importance du terme dans l'ensemble du corpus. Elle vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants.

- **Pondérer les titres en fonction de l'ancienneté**

Les expériences plus récentes sont plus pertinentes

Raw data

Document ID	Textual description
1	"Data science is fun"
2	"Artificial intelligence is the future"
3	"Business and artificial intelligence combination is the key"

TF

Word	"data"	"science"	"fun"	"artificial"	"intelligence"	"future"	"business"	"combination"	"key"
Document									
1	1	1	1	0	0	0	0	0	0
2	0	0	0	1	1	1	0	0	0
3	0	0	0	1	1	0	1	1	1

X

IDF

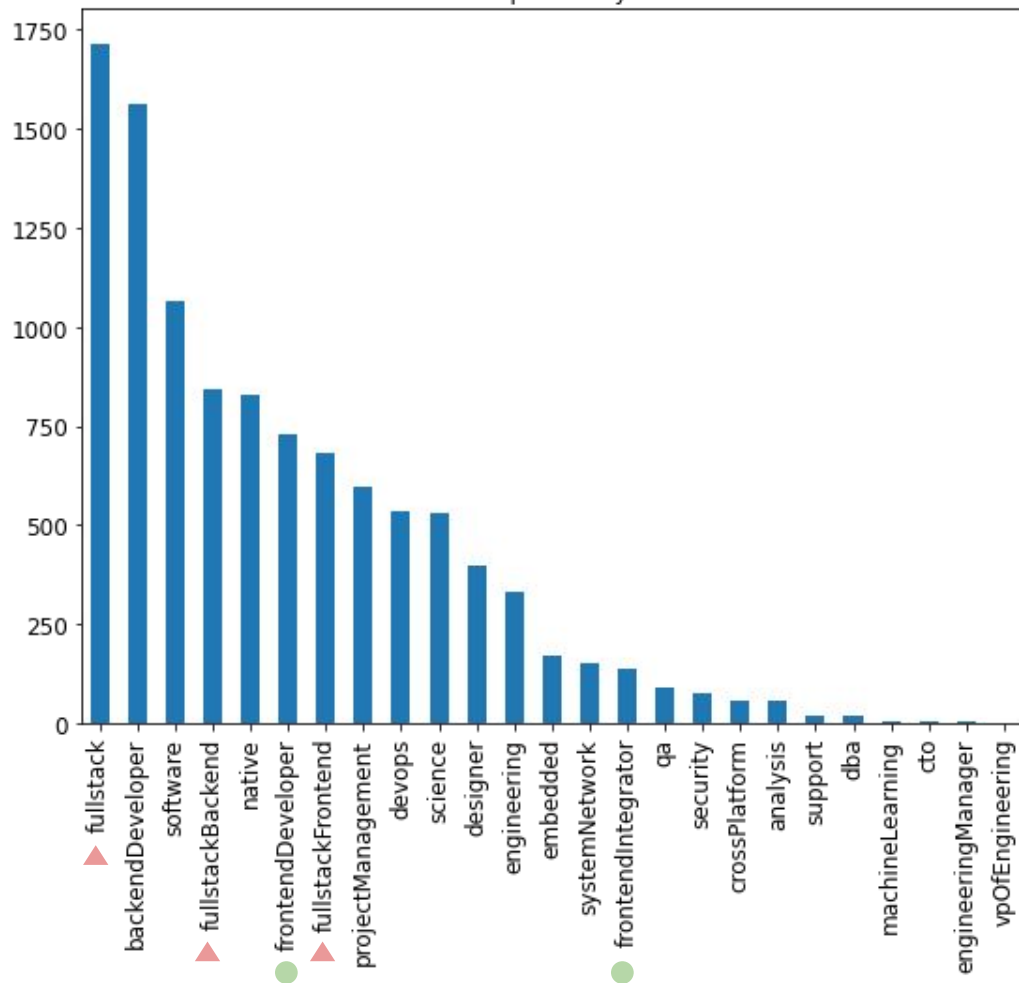
Word	"data"	"science"	"fun"	"artificial"	"intelligence"	"future"	"business"	"combination"	"key"
Document									
1	0,48	0,48	0,48	0,18	0,18	0,48	0,48	0,48	0,48
2	0,48	0,48	0,48	0,18	0,18	0,48	0,48	0,48	0,48
3	0,48	0,48	0,48	0,18	0,18	0,48	0,48	0,48	0,48

=

TF-IDF

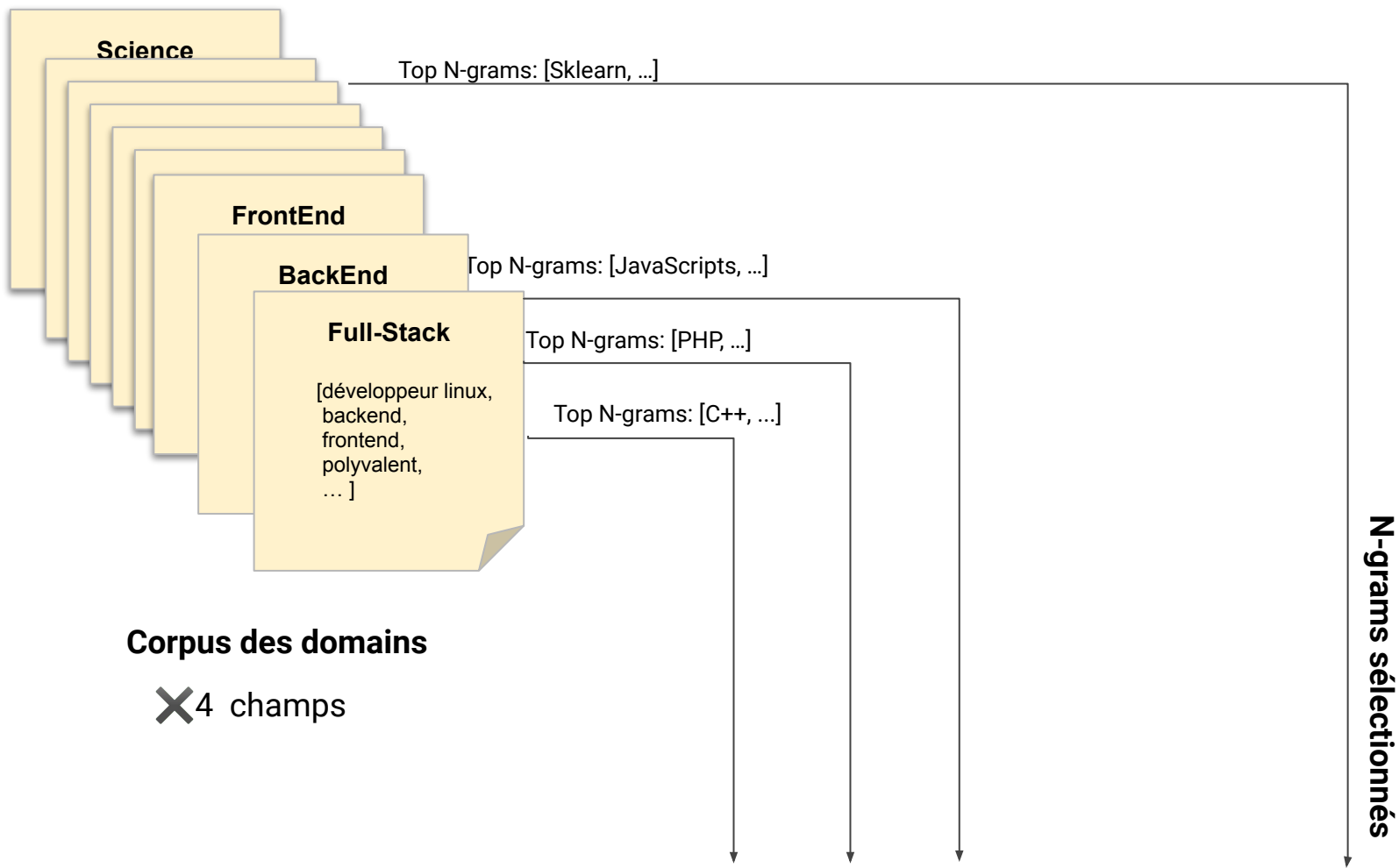
Word	"data"	"science"	"fun"	"artificial"	"intelligence"	"future"	"business"	"combination"	"key"
Document									
1	0,48	0,48	0,48	0,00	0,00	0,00	0,00	0,00	0,00
2	0,00	0,00	0,00	0,18	0,18	0,48	0,00	0,00	0,00
3	0,00	0,00	0,00	0,18	0,18	0,00	0,48	0,48	0,48

Nubmer of profiles by domain



Nettoyage de domain

- Fusionner des domaines similaires:
Fullstack
Frontend
.....
- Laisser de côté quelques domaines non-tech dans le but d'éviter over-fitting



Classification : Bag of Words

Corpus des documents

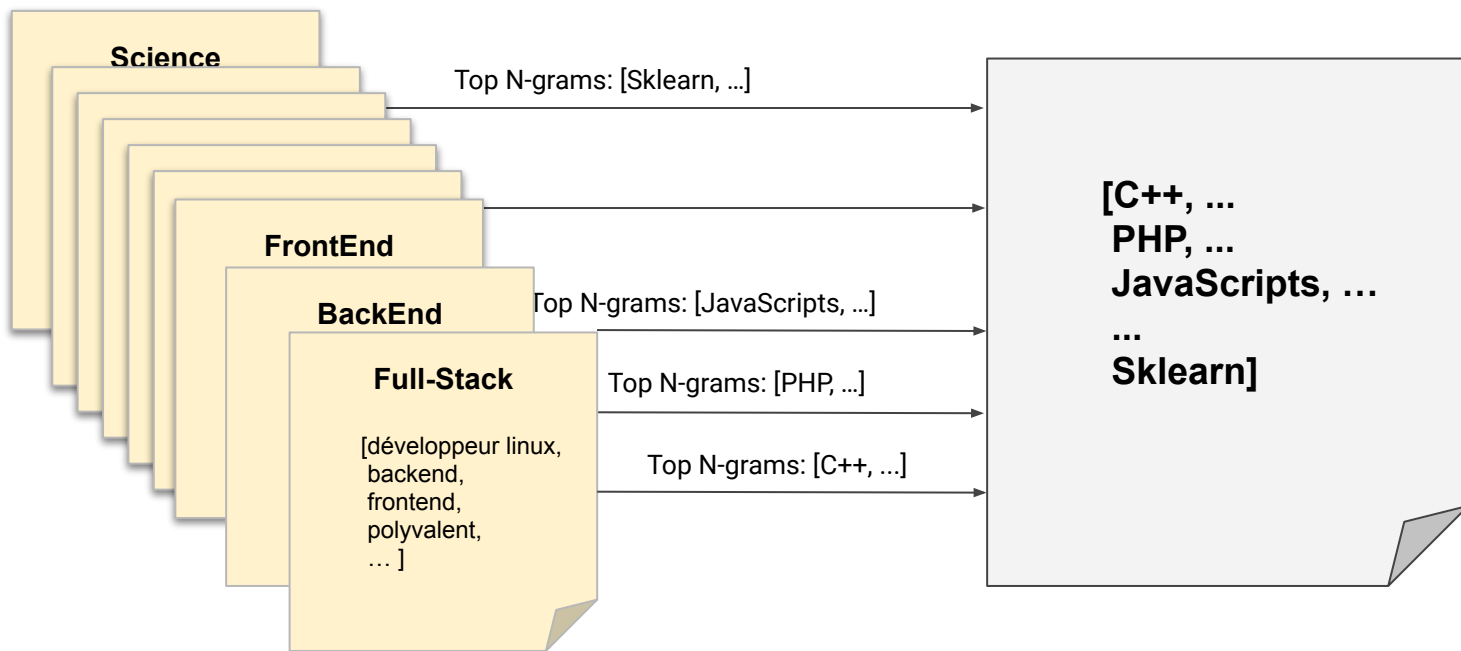
[PHP, Database, Node, ASP.NET...]

[DevOps, C++, CSS,...]

[Scikit-learn, algorithms, developer, supervised...]

Full-Stack	BackEnd	FrontEnd	Science
2	6	1			0
5	1	1			3
2	0	2			8

- Compter la fréquence de N-grams de chaque domaine pour chaque document
- Attribuer le domaine dont la somme de fréquence est la plus élevée



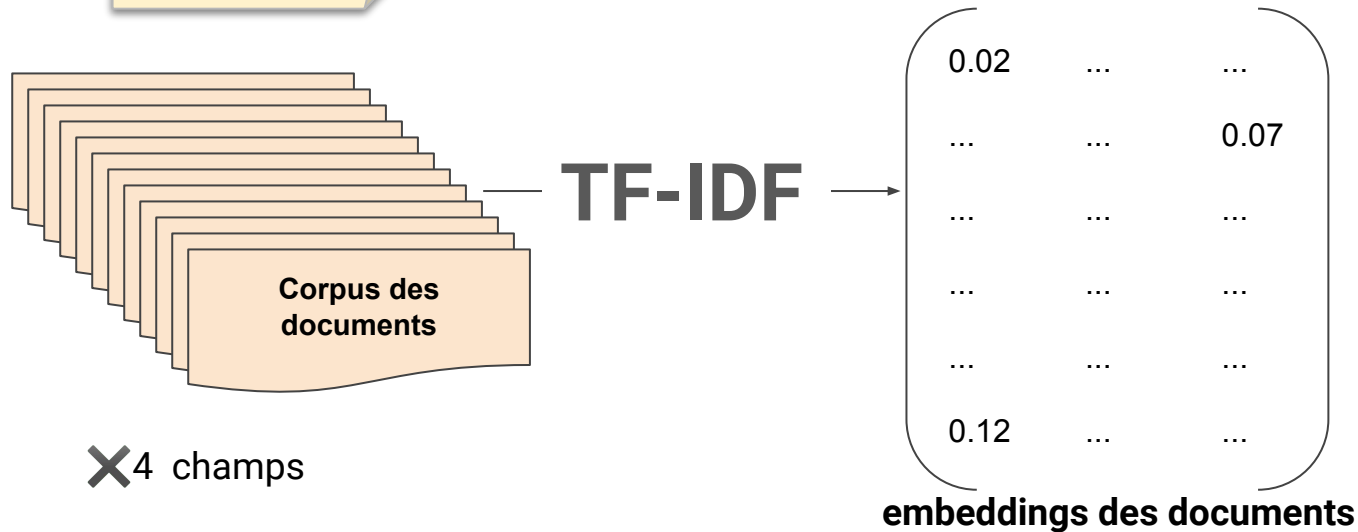
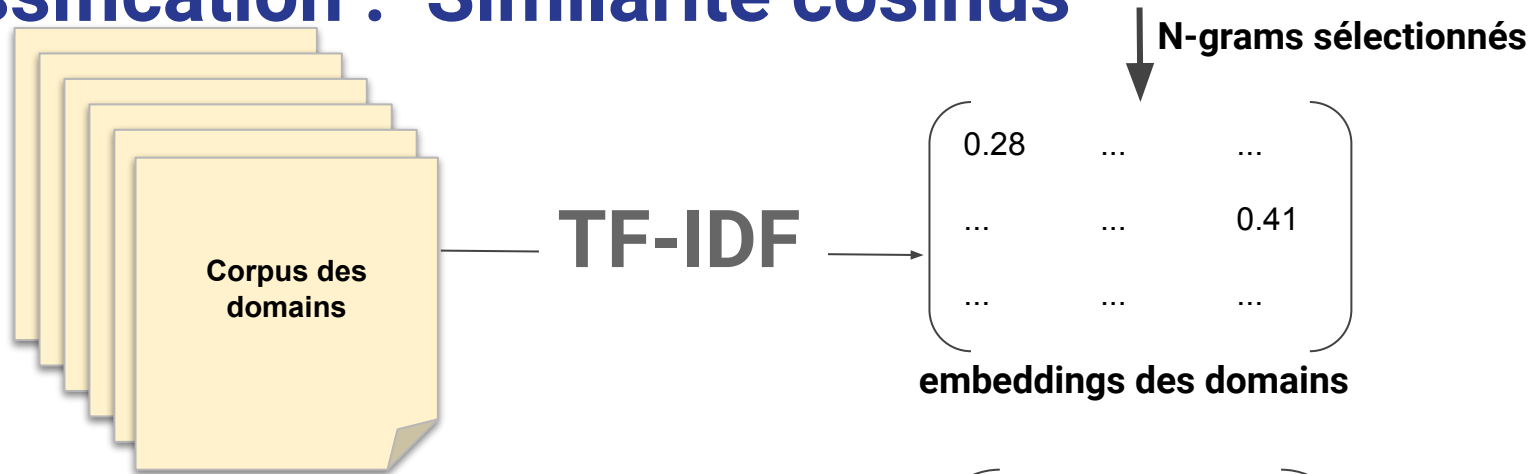
Corpus des domains

✕ 4 champs

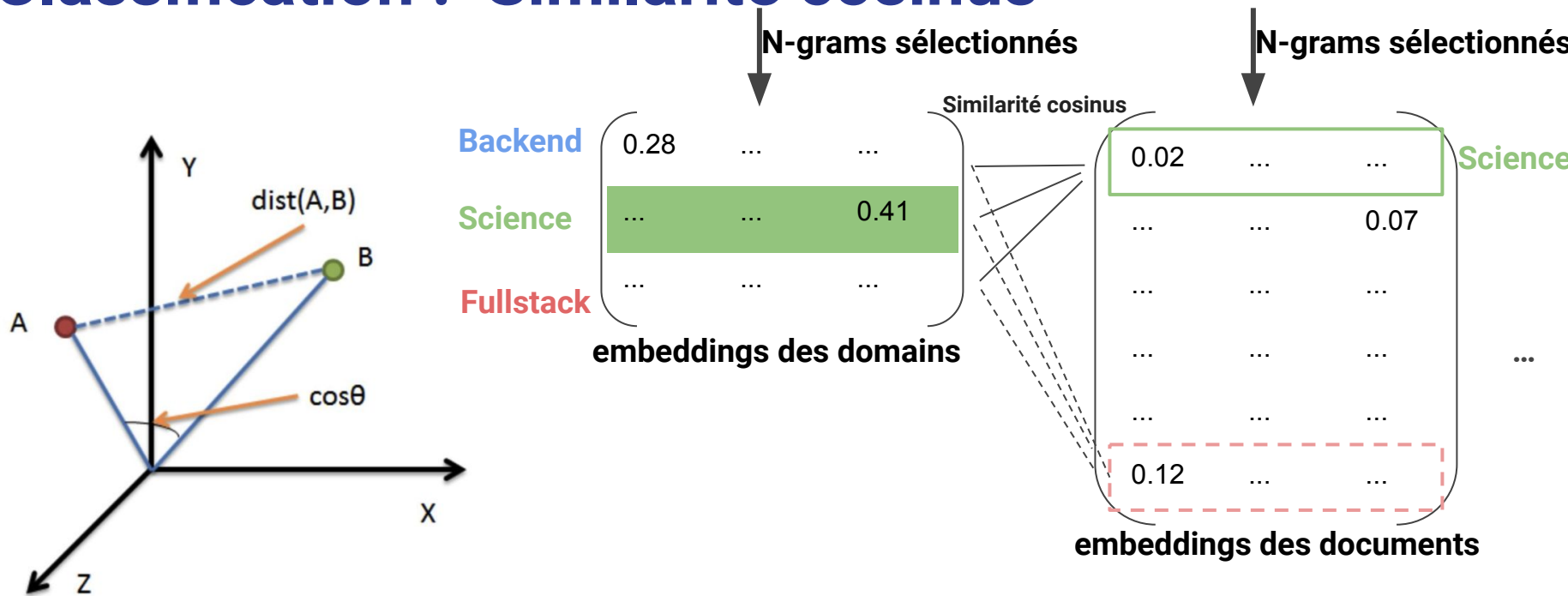
N-grams sélectionnés



Classification : Similarité cosinus

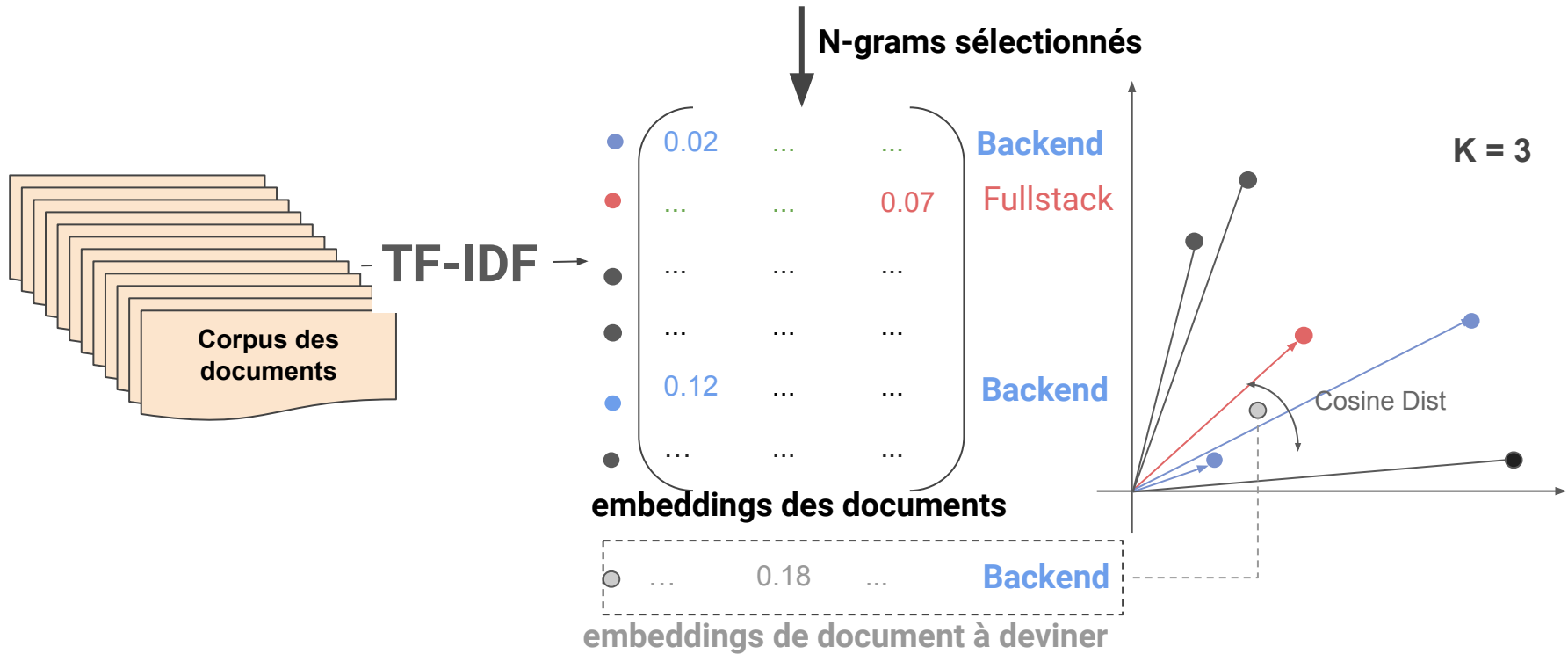


Classification : Similarité cosinus



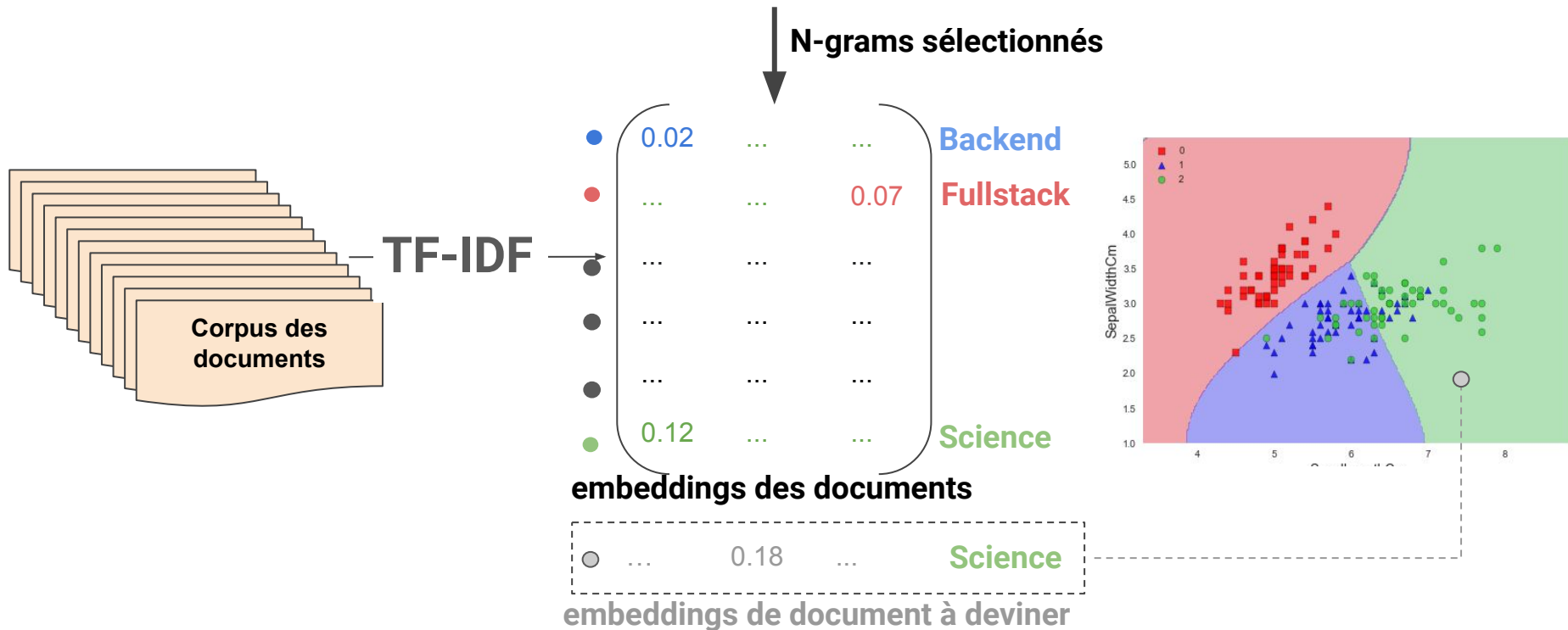
- Similarité cosinus est généralement utilisée comme métrique pour mesurer la distance entre documents
- Calculer la similarité cosinus entre vecteur document et chaque vecteur domaine
- Pondérer et combiner les matrices de similarité cosinus de chaque champs
- Attribuer le domaine dont la somme de similarité cosinus est la plus élevée

Classification : K-NN



- Attribuer le domaine le plus fréquent dans les K plus proches voisins (ou dans le rayon de R)

Classification : Linear SVC



Evaluation

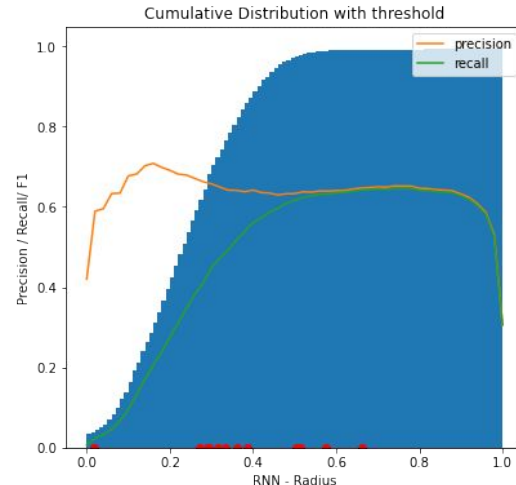
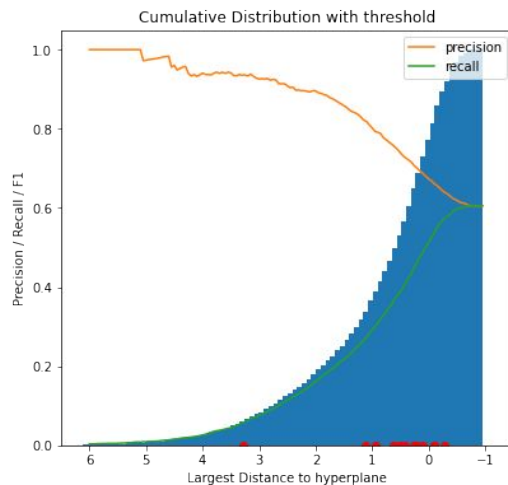
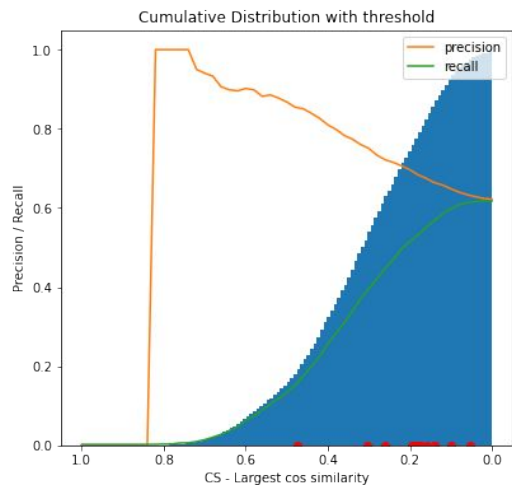
Tradeoff : Precision, Recall, Seuil de confiance

Matrice d'évaluation	Prediction	Similarité Cosinus	Rayon de KNN	Distance à l'hyperplan de SVC	Groud Truth
<div>0.02</div> <div>... ... 0.07</div> <div>...</div> <div>...</div> <div>0.12</div>	Backend	0.8	0.1	5	Backend
	Fullstack	0.5	0.5	3	Fullstack
	Science	0.4	0.3	1	Science

	QA	0.7	0.6	4	QA
		Similarité Cosinus	Rayon de KNN	Distance à l'hyperplan de SVC	
	Gamme	[0,1]	[0,1]	$[-\infty +\infty]$	
	Confiance	<div>→+ </div>	<div>+ ← </div>	<div>→+ </div>	

Evaluation

Tradeoff : Precision, Recall, Seuil de confiance



On baisse (augmente pour RNN) le seuil afin de devenir sûr plus de profils, en même temps, on risque d'endommager la précision, car on a moins en moins de confiance pour les profils suivants.

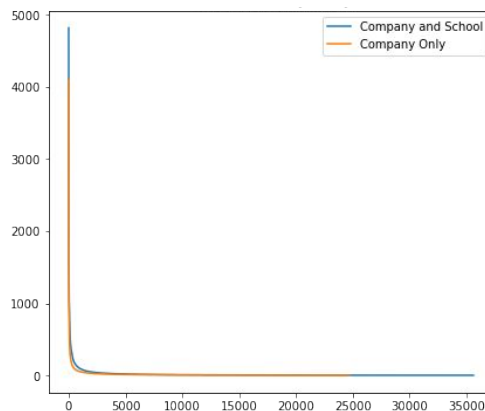
Temps (s)	BoW	Similarité Cosinus	K-NN	Linear SVC
Entraînement	/	/	0.02	0.5
Prédiction	1.95	0.02	1.19	0.02

M2. Recommandation d'entreprise pour les coders

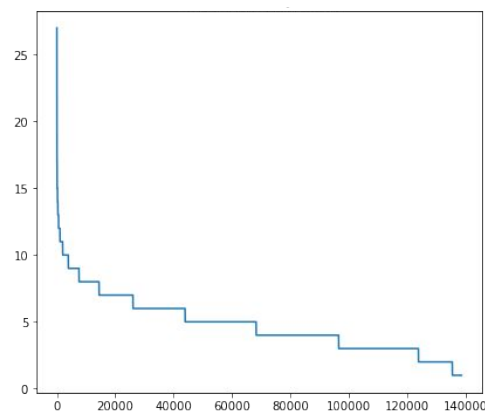


Coder_id	Company_id
C1	E10
C1	E55
C2	E8
C3	E100
...	...

Coder_id	School_id
C1	S6
C2	S99
C2	S12
C3	S4
...	...



Nombre d'coder par entité

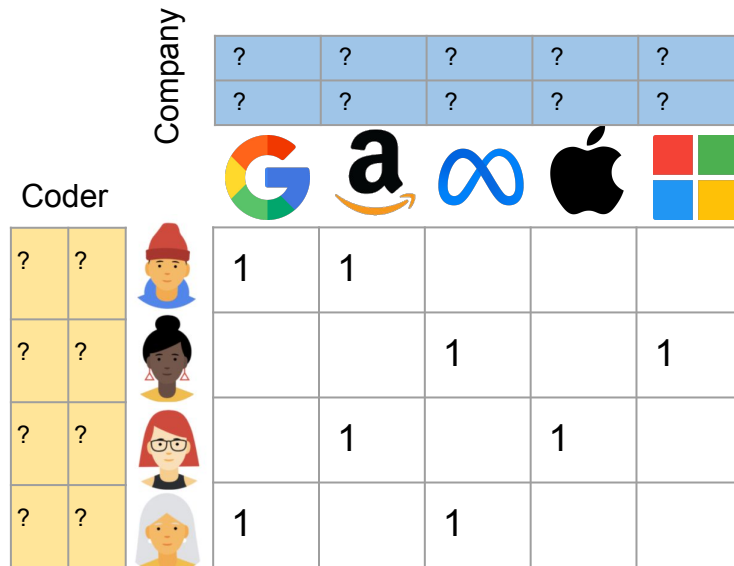
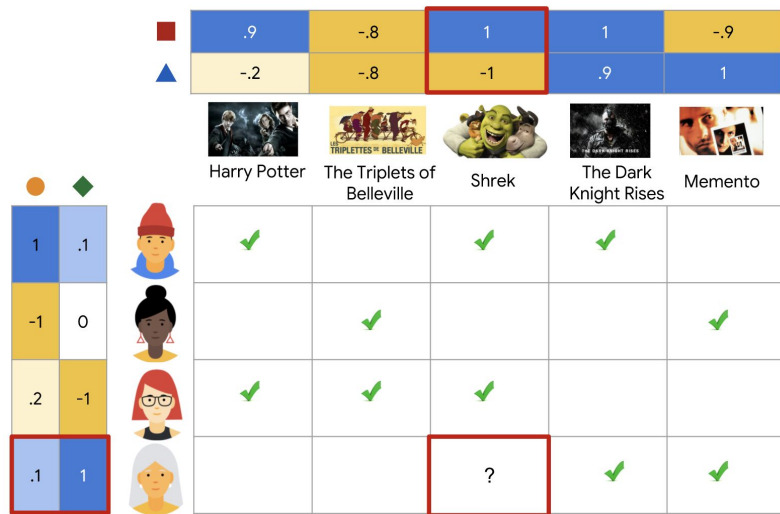


Nombre d'entité par coder

- Le minimum d'occurrence de coder : 3
- Le minimum d'occurrence de company: 3

Collaborative Filtering

Matrice Feedback



1 signifie le coder avait travaillé dans cette entreprise; 0 le contraire

Objectif: représenter la matrice par le produit de 2 matrices de faible dimension (celle de coder et de company) de sorte que chaque vecteur de faible dimension représente un coder ou une entreprise

Collaborative Filtering

Alternating Least Squares

$$\min_{x_{\star}, y_{\star}} \sum_{u,i} c_{ui} (p_{ui} - x_u^T y_i)^2 + \lambda \left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2 \right)$$

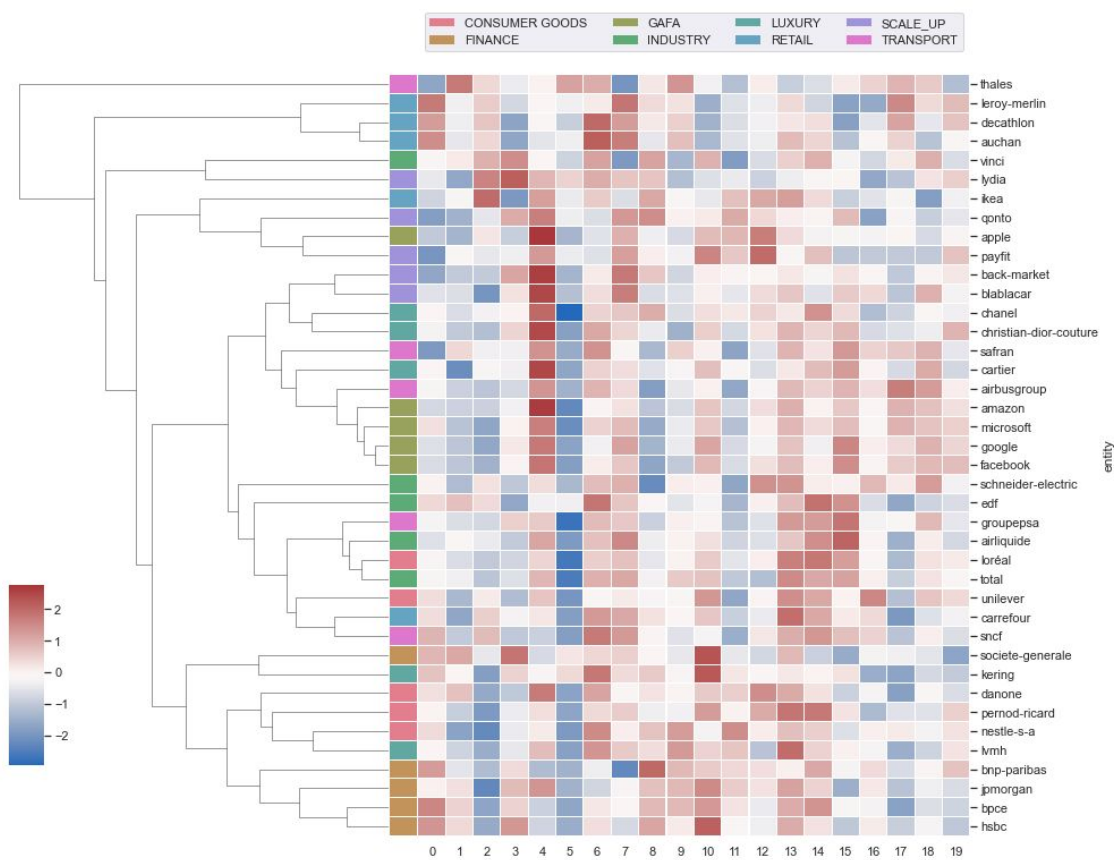
- Sparsity of matrix : 0.00013
- Alternier entre le calcul des User embeddings et Company embeddings, et chaque étape est garantie pour réduire la valeur de cost function.
- Dimension de company/coder vecteur: 20 ; Regularization: 0.2

Visualization d'embeddings

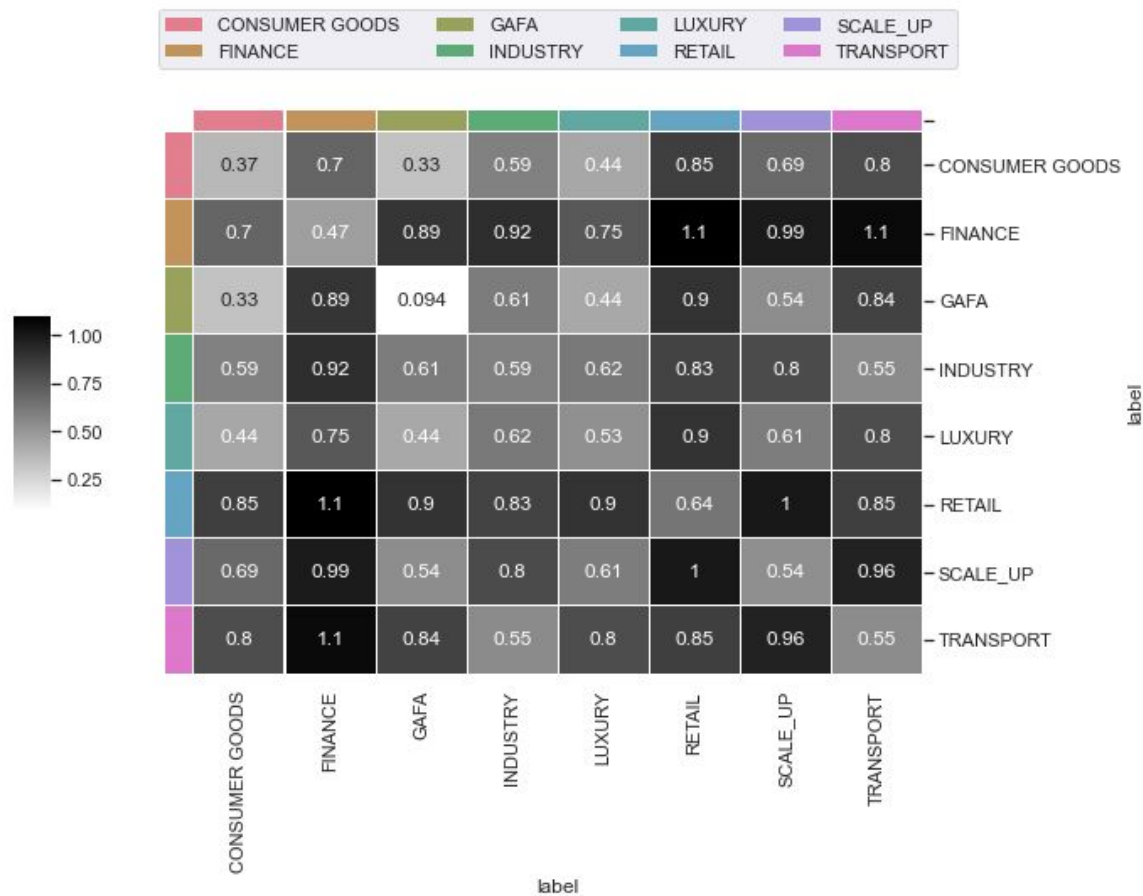
Consumer Goods	Finance	GAFA	Industry	Luxury	Retail	Scale Up	Transport
L'oréal	HSBC	Google	Total	Kering	IKEA	Payfit	Airbus
Danone	BPCE	Amazon	EDF	Dior	Carrefour	Lydia	Safran
Nestlé	BNP Parisbas	Facebook	Air Liquide	Channel	Auchan	Back Market	Groupe PSA
Unilever	JP Morgan	Apple	Vinci	LVMH	Decathlon	Blablacar	Thales
Pernod Ricard	Société Générale	Microsoft	Schneider Electric	Cartier	Leroy Merlin	Qonto	SNCF

On a regroupé des entreprises dans 8 catégories, chacune contenant 5 entreprises, soit 40 au total.

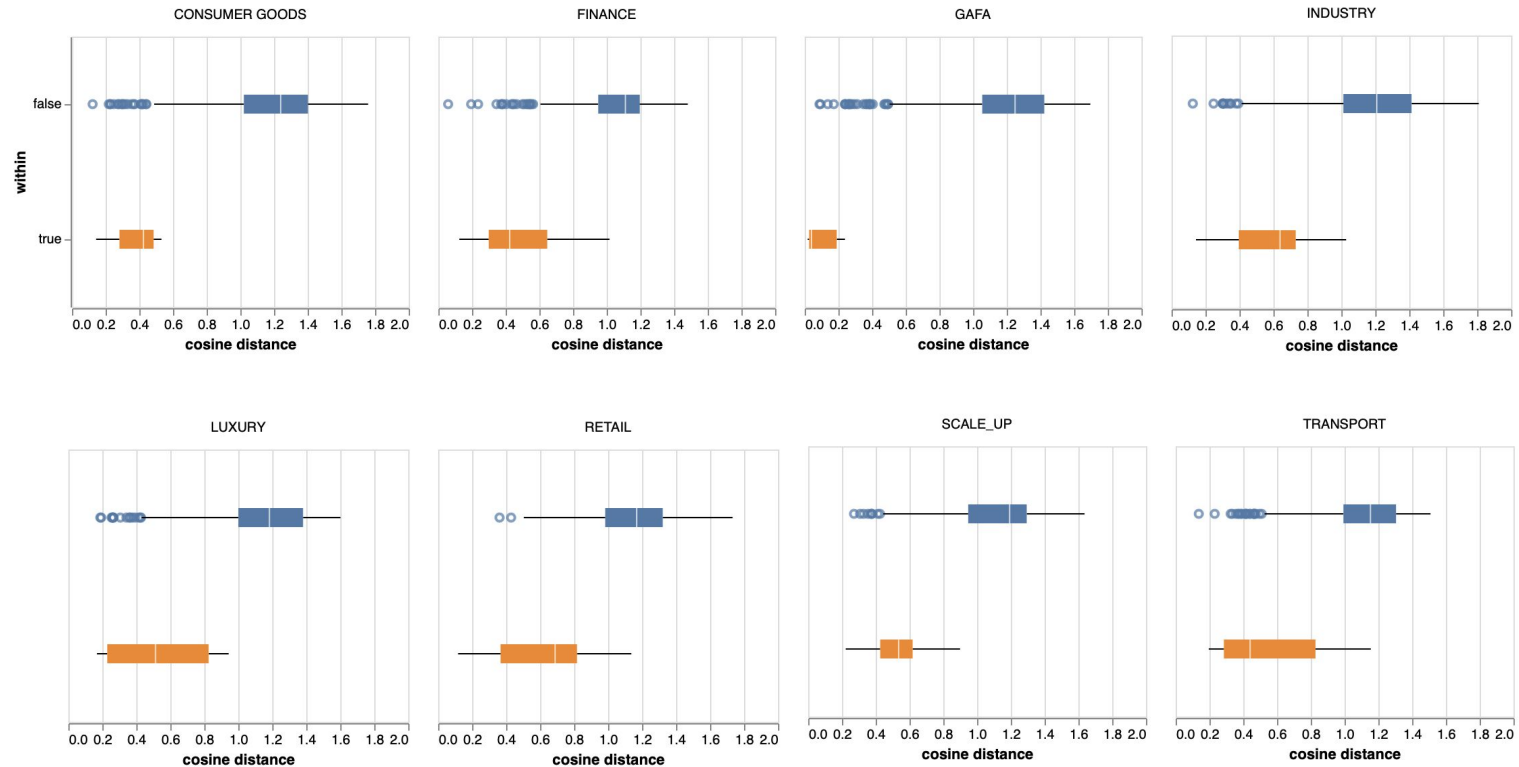
Graph Dendrogramme de clustering



Graph Heatmap - Distance Cosinus: Couples de catégories

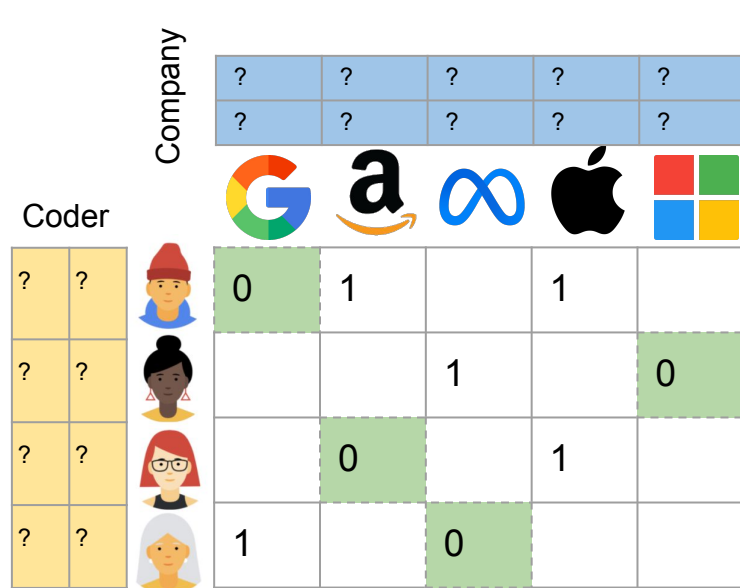


Graph Boxplot - Distance Cosinus à l'intérieur et à l'extérieur

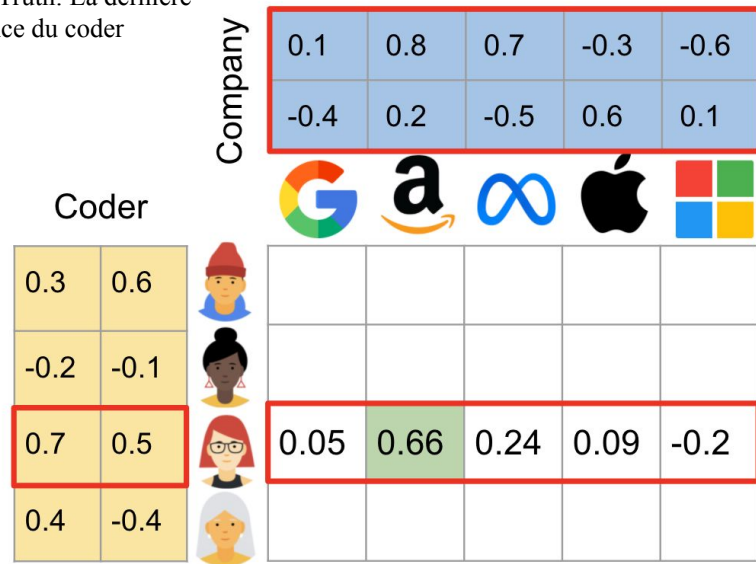


Evaluation

Sur les dernières expériences des coders



Ground Truth: La dernière expérience du coder



- Supprimer les dernières expériences de chaque coder dans la matrice d'expérience et recalculer les embeddings

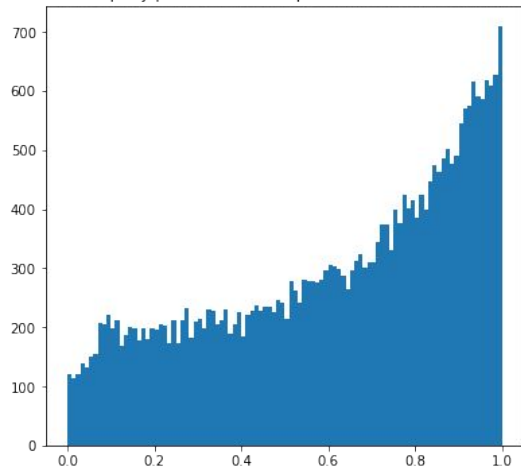
- Reproduire les similarité cosinus des deux embeddings et récupérer les éléments sur ces ground-truths

Evaluation

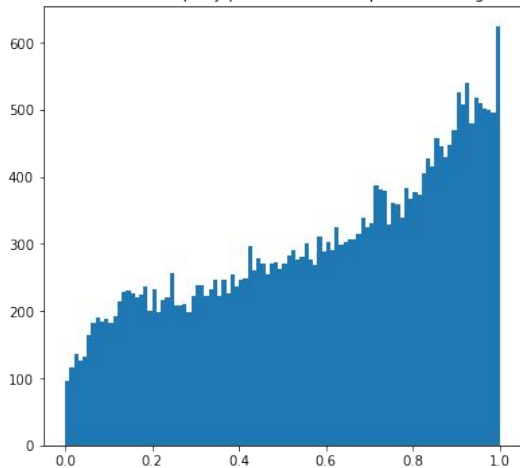
	Avec écoles	Sans écoles
Quantile Moyen	0.6203	0.5934
Similarité Moyen	0.2789	0.2188

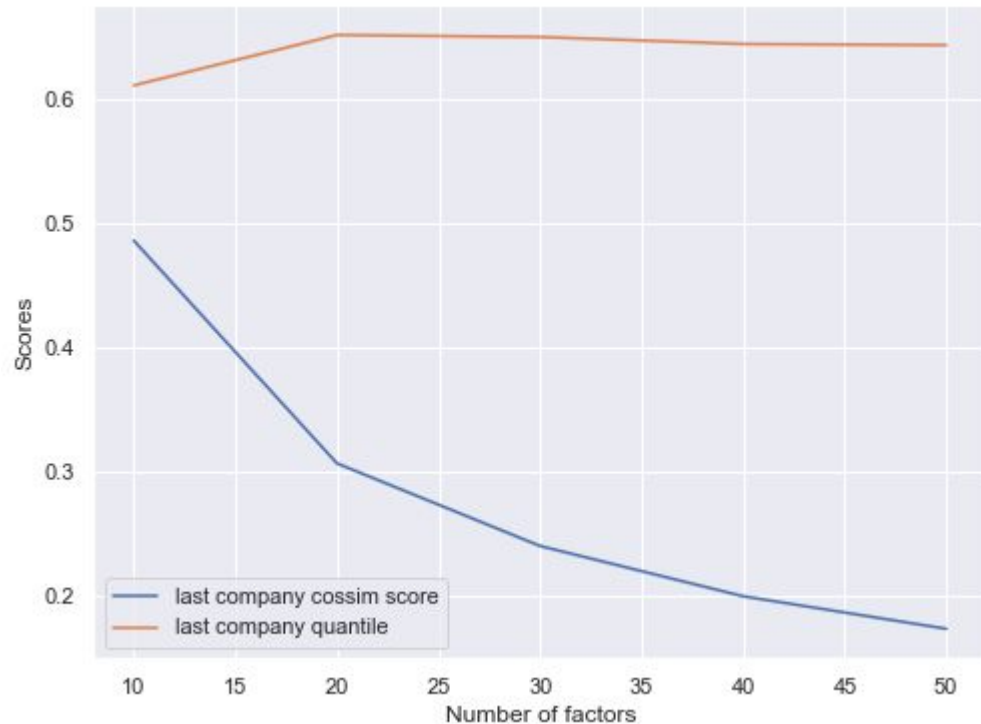
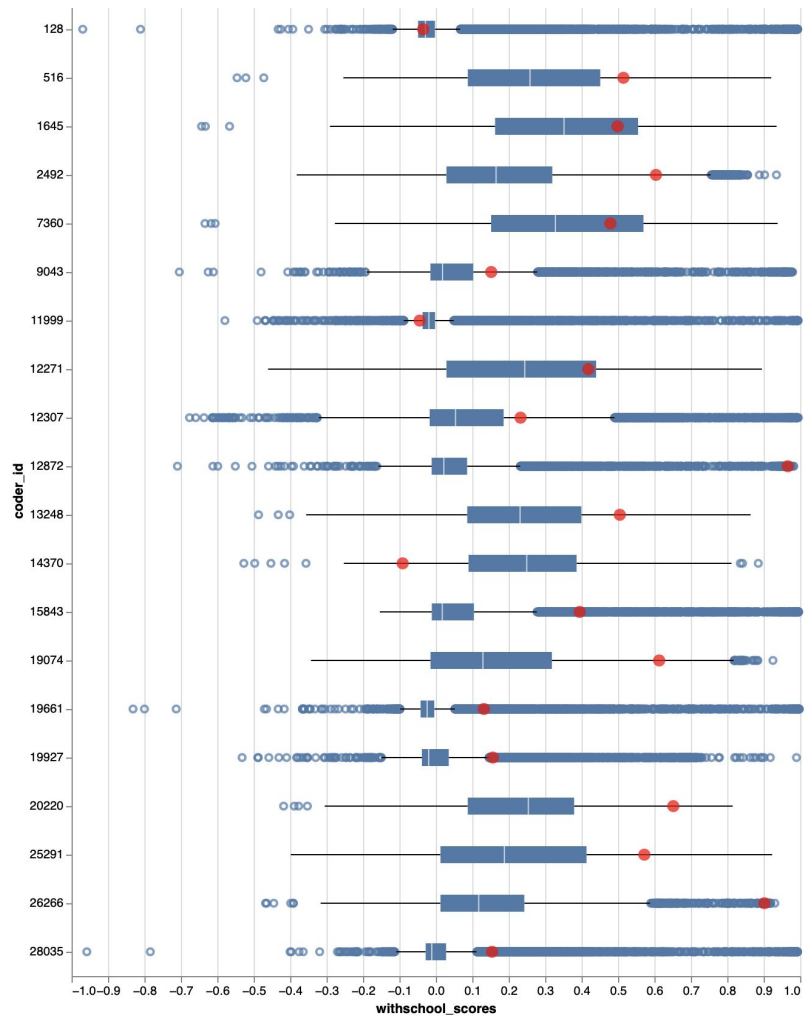
- Les écoles nous aident à augmenter le quantile et le score considérablement
- Le percentile moyen à 65 est bien supérieur que la base de référence 50

Coder's last company prediction score quantile (WithSchool Vecs) histogram



Coder's last company prediction score quantile histogram







Merci