Soutenance de Projet de Fin d'Études

Master Web Intelligence et Science des Données (WISD) en double diplomation avec l'Université Sorbonne Paris Nord

Normalisation automatique de variables issues de bases de données en agroécologie

Membres du jury :

- M. Ali YAHYAOUY
- M. Badraddine Aghoutane
- M. Abdelouahed Sabri
- M. Mohammed Adnane Mahrez

Réalisé par :

Oussama MECHHOUR

Sous la direction de :

- Mme. Sandrine AUZOUX
- M. Mathieu ROCHE
- M. Benjamin HEUCLIN
- M. Ismail EL BATTEOUI

Plan

I. Introduction générale
II. Problématique et objectifs
III. Méthodologies utilisées
IV. Comparaison des résultats
V. Interface web
VI. Conclusion et perspectives

I. Introduction général

cirad

- **1 700**
- **200**
- **220 M€**
- **900**

#DigitAg

- +570
- 16
- **30**
- **9,9 M€**

II. Problématique et objectifs

❑ Agroécologie :

Une approche holistique de l'agriculture qui intègre des principes écologiques, économiques et sociaux pour développer des systèmes agricoles durables

❑ Problématique

Comment pouvons-nous faire correspondre les variables sources et candidates ?

❑ Objectifs

1. Automatisation de la mise en correspondance des variables

2. Développement d'une interface web destinée aux chercheurs

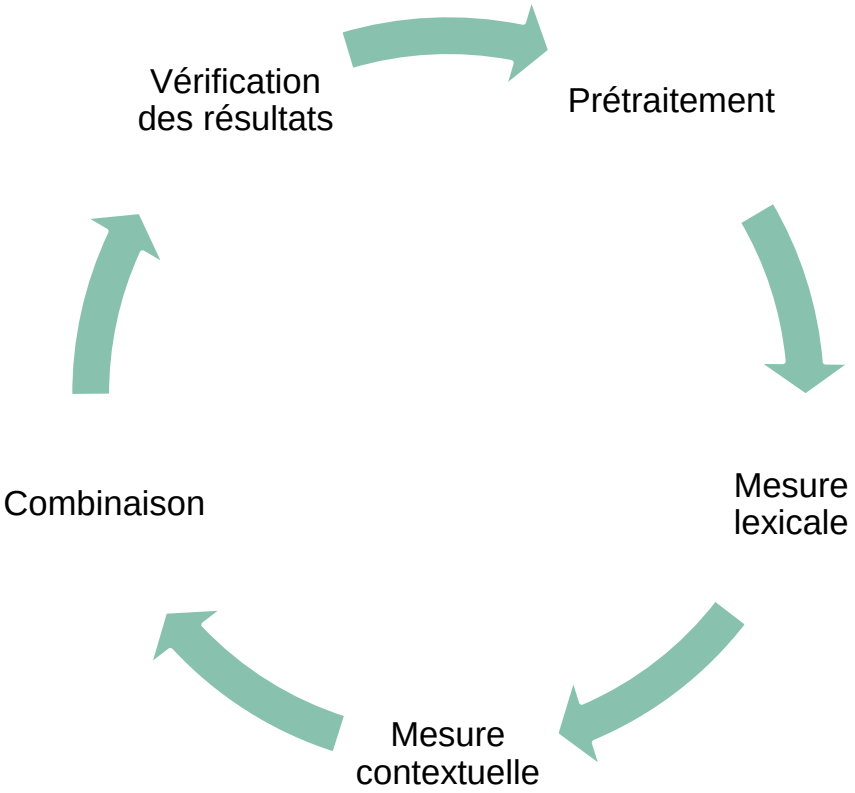
Exemples de données

Noms des variables sources	Descriptions des variables sources
Yield_CAS_t.ha-1	Cane yield (in fresh machinable stem)
Sugar_CAS_%	Sugar content of fresh stem mass
….	….

Noms des variables candidates	Descriptions des variables candidates
root_crop_yield_dm_t.ha-1	measurement of root dry biomass at plot level
root_crop_yield_fm_t.ha-1	measurement of root fresh biomass at plot level
….	….

III. Méthodologies utilisées

Récapitulatif des tâches réalisées



Prétraitement

clean_text()

- measurement of plant (or species) recovery by ceptometry
- measurement of plant recovery by ceptometry

remove_stopwords()

- measurement of root dry biomass at plot level
- measurement root dry biomass plot level

lemmatize()

- total number leaf blades tracking labeled stems
- total number leaf blade tracking labeled stem

remove_punctuation()

- Acid Detergent Fiber : lignin, cellulose

replace_synonyms()

- level -> degree, yield -> output, stem -> root

Méthodes

1

Mesure lexicale

2

Mesure contextuelle

3

Combinaison

Méthodes

1

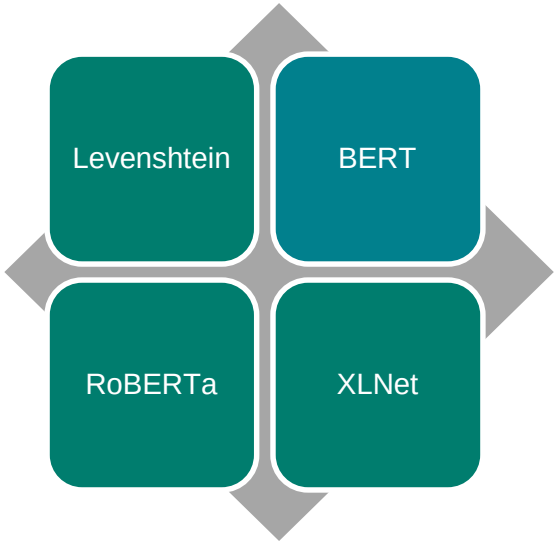
Mesure lexicale

2

Mesure contextuelle

3

Combinaison

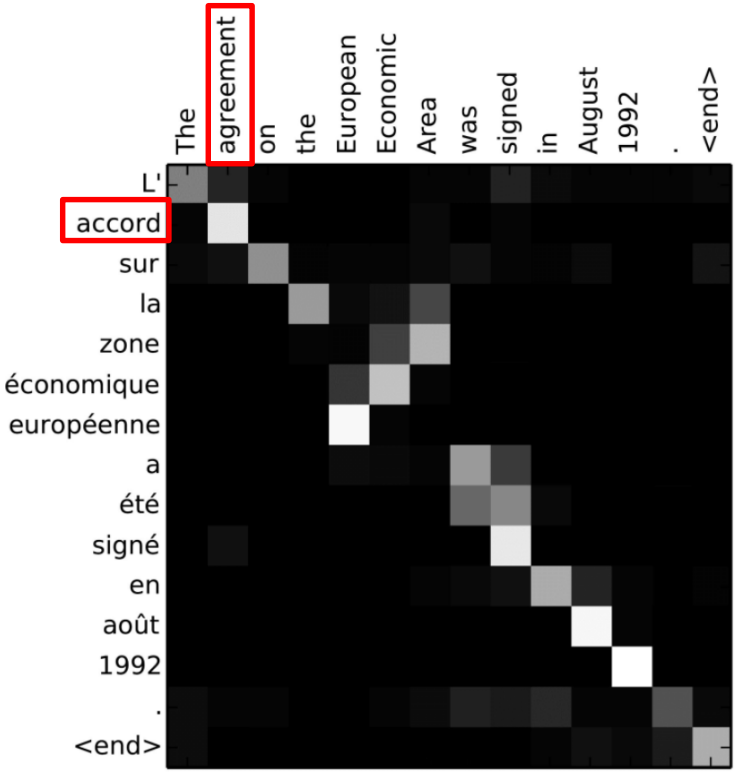


Variable source	Variable candidate
rendement des cultures	production agricole par hectare

Levenshtein(var_s, var_c) = 16 ⟺ Les 2 variables ne sont pas similaires

Oussama MECHHOUR

Méthodes / BERT



Mécanisme d'Attention

Méthodes

1

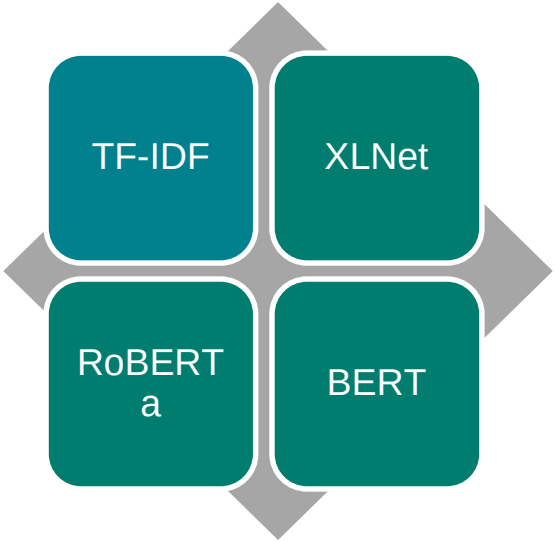
Mesure lexicale

2

Mesure contextuelle

3

Combinaison



Méthodes

1

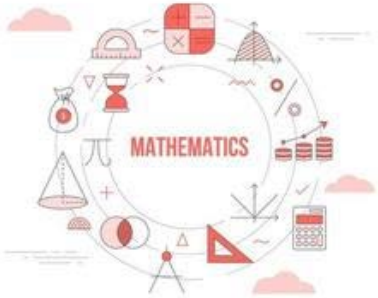
Mesure lexicale

3

Combinaison

2

Mesure contextuelle



$$\text{combinaison} = \alpha \cdot X + (1 - \alpha) \cdot Y$$

Oussama MECHHOUR

Sans contexte

IV. Comparaison des résultats

Sans contexte

❖ Résultats précédents (TF-IDF et Levenshtein) :

	Nombre de bons résultats après lemmatisation (sur les 84 variables sources)											
	Levenshtein				TF-IDF + Cosinus				Combinaison			
	précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10
0	15.48%	19.05%	23.81%	42.86%	33.33%	42.86%	51.19%	60.71%	41.67%	51.19%	64.29%	73.81%

❖ Résultats actuels (BERT-base) :

	Nombre de bons résultats après lemmatisation (sur les 84 variables sources)											
	BERT-base + cosinus(noms des variables)				BERT-base + cosinus(descriptions des variables)				Combinaison			
	précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10
0	11.90%	28.57%	36.90%	53.57%	33.33%	44.05%	52.38%	57.14%	41.67%	60.71%	69.05%	79.76%

Discussion

Pour la méthode combinaison :

1. La précision à la position 3 a été augmentée de 9,52%
2. La précision à la position 5 a été augmentée de 4,76%
3. La précision à la position 10 a été augmentée de 5,95%

La similarité entre les noms des variables :

1. La précision à la position 3 a été augmentée de 9,52%
2. La précision à la position 5 a été augmentée de 1,09%
3. La précision à la position 10 a été augmentée de 10,71%

La similarité entre les descriptions des variables :

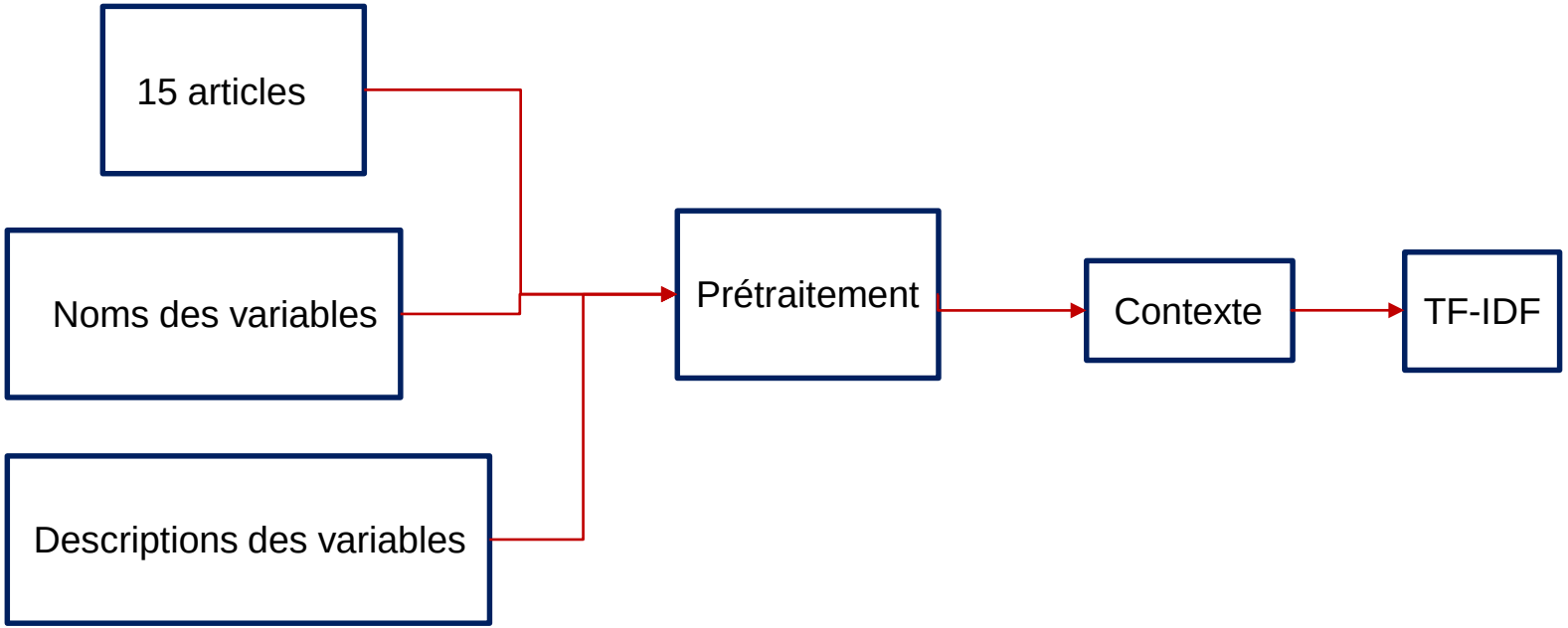
1. La précision à la position 3 et 5 a été augmentée de 1,19%

Avec contexte

IV. Comparaison des résultats

Processus

- Travail actuel :



Avec contexte

❖ Résultats précédents (TF-IDF et Levenshtein) :

	Nombre de bons résultats après lemmatisation (sur les 84 variables sources)											
	Levenshtein				Cosinus				Combinaison			
	précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10
0	15.48%	19.05%	23.81%	42.86%	33.33%	42.86%	51.19%	60.71%	41.67%	51.19%	64.29%	73.81%

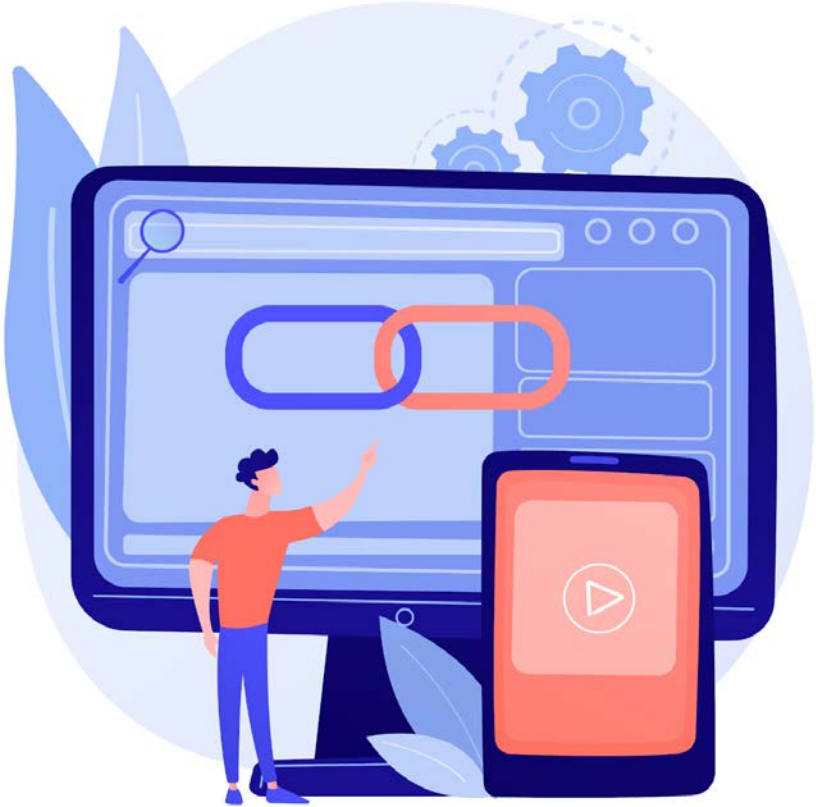
❖ Résultats actuels (BERT-base et TF-IDF) :

	Nombre de bons résultats après lemmatisation (sur les 84 variables sources)											
	BERT-base + cosinus(nom de variables)				TF-IDF + cosinus(description de variables)				Combinaison			
	précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10	précision au rang 1	Rang 3	Rang 5	Rang 10
0	11.90%	28.57%	36.90%	53.57%	29.76%	42.86%	51.19%	64.29%	52.38%	66.67%	71.43%	80.95%

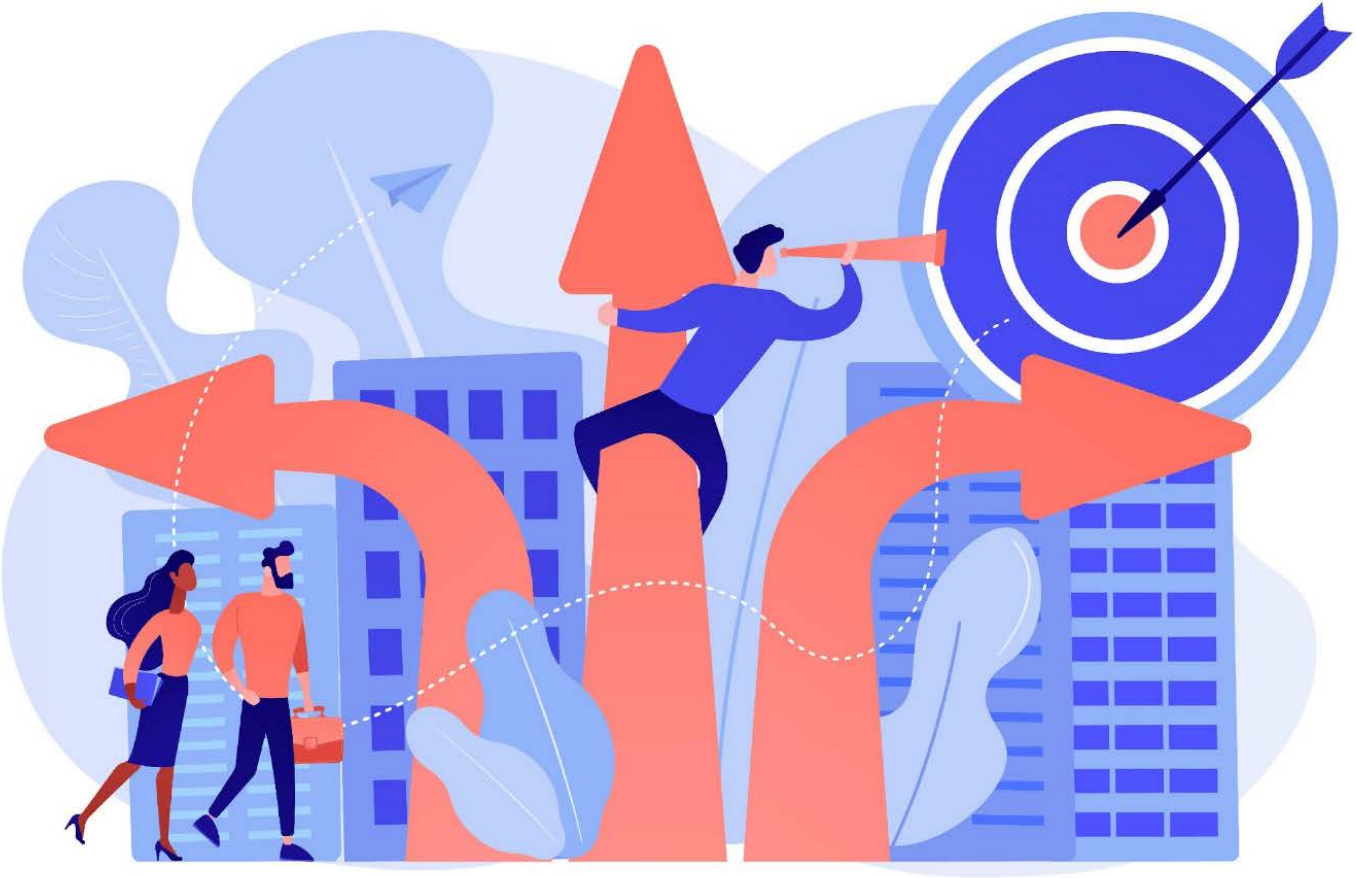
Pour la méthode combinaison :

1. La précision à la position 1 a été augmentée de 8,33%
2. La précision à la position 3 a été augmentée de 10,72%
3. La précision à la position 5 et 10 a été augmentée de 7,14%

V. Interface web/Démonstration



VI. Conclusion et perspectives



Limites et défis

01 Un nombre limité de variables en anglais (84 variables)

02 Les variables sont souvent formulées de manière non canonique

03 Certaines variables sont en français (non utilisées dans ce stage)

04 Les descriptions des variables sont très courtes

05 Les ontologies associées aux variables n'ont pas été abordées dans le cadre du stage

MERCI DE
VOTRE
ATTENTION