Modélisation de recommandation de code SH par IA

Dr Seon Yeong Han

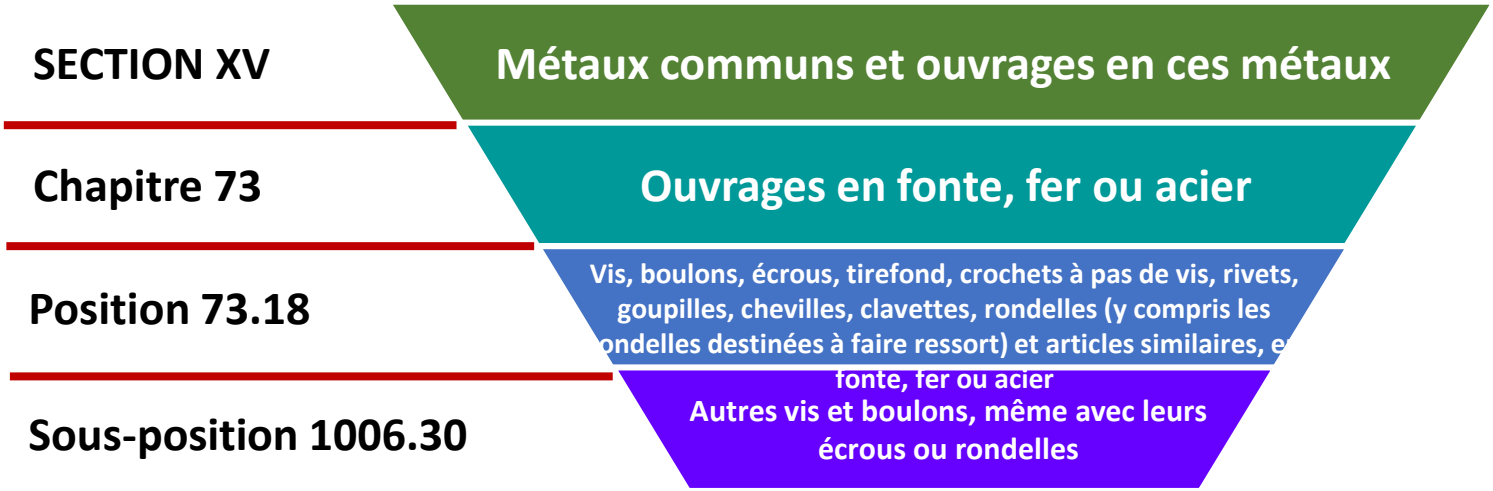
I . Vue d'ensemble

❍ **Code SH (Code du Système harmonisé ; Système harmonisé de désignation et de codification des marchandises)**

– Écarts entre les descriptions commerciales des marchandises et les descriptions figurant dans la Nomenclature

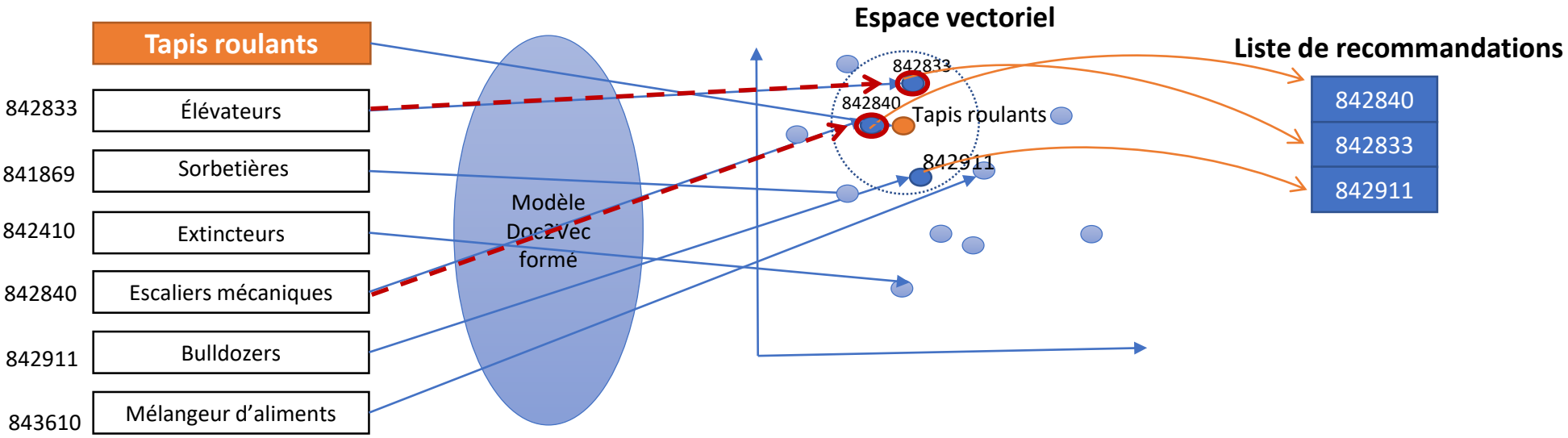
VIS – CTNS PLTS

?



I . Vue d'ensemble

❍ Diagramme de conception de modèle par IA



✓ **Intégration** – Convertir le langage humain en valeurs vectorielles

✓ **Classement** – Classer le nouveau vecteur de description dans les vecteurs de code SH existants

Ⅱ. Prétraitement des données

❍ **Données**

✓ **Importations américaines 2020 (Envois dans le système automatisé de manifeste – AMS) (de janvier à septembre 2020)**

CODE SH	Description
2102201000	PRODUITS NATURELS – RACINES DE GINGEMBRE MOULU NON SULFITÉES (45 3 g)
6 chiffres	Chaîne de caractères : lettres, chiffres, caractères spéciaux (long. moy. = 100)

✓ **Prétraitement**

– **Suppression des doubles espaces**

– **Suppression des descriptions constituées de chiffres uniquement**

– **Suppression des descriptions comportant deux caractères ou moins**

– **Mise en majuscules**

– **Suppression des données redondantes**

Ⅱ. Prétraitement des données

❍ **Problèmes détériorant la performance du modèle**

1) Données mal étiquetées

- **Descriptions associées à des codes SH erronés (cas de fraude au classement)**

2) Hors vocabulaire

- **Descriptions contenant des mots qui ne figurent pas dans le vocabulaire du modèle**
- **Erreurs de frappe, mots mal orthographiés**

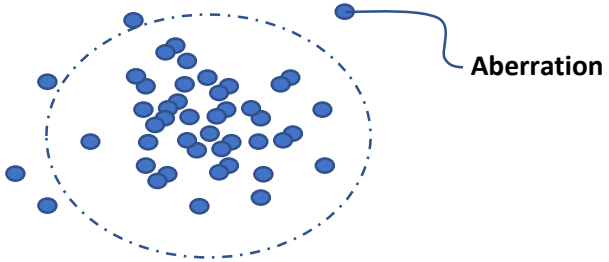
3) Données déséquilibrées

- **Chaque code du SH associé à un volume de données déséquilibrées**

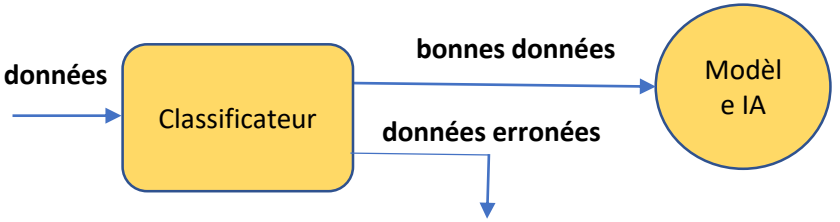
Ⅱ. Prétraitement des données

❍ Techniques de prétraitement – Données mal étiquetées

- ✓ **Détection d'aberration basée sur la distance**
  - **– Détecte les données qui sont éloignées du même groupe de codes du SH**



- ✓ **Détection d'aberration basée sur le classement**
  - **– Utilise un classificateur distinct pour supprimer l'aberration**



- ✓ **Ignorer le problème**
  - **– Utilise un modèle très peu sensible aux aberrations mineures**

Ⅱ. Prétraitement des données

❍ Techniques de prétraitement

✓ **Caractère « n-gram » – Hors vocabulaire**

Ex. : 2-gram

téléphone cellulaire intelligent (cellular smartphone) → {ce, el, ll, lu, ul, la, ar, sm, ma, ar, rt, tp, ph, ho, on, ne}: 16 mots secondaires

téléphones cellulaires intelligents (cellular smartphones) → {ce, el, lu, ul, la, ar, sm, ma, ar, rt, tp, ph, ho, on, ne, es}: 16 mots secondaires

→ Correspondance 15/16 → valeurs de vecteur similaires

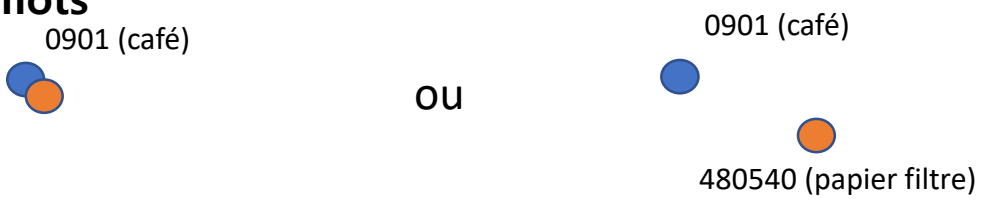
✓ **Mot « n-gram »**

– ***N* mots consécutifs -> une unité d'information**

– **Tient compte de l'ordre des mots**

Ex. : Filtre à café

café filtre



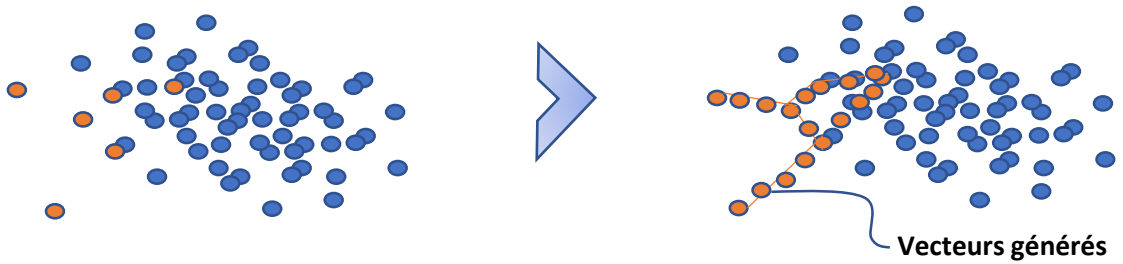
Ⅱ. Prétraitement des données

❍ Techniques de prétraitement – Données déséquilibrées

✓ **Échantillonnage équilibré**

– Technique dite « Synthetic Minority Oversampling Technique » (SMOTE)

: Génère des données pour la classe minoritaire



Ⅲ. Modèle IA

❍ Modèle IA – Scénario

Word2Vec →

Doc2Vec →

FastText →

Scénario	Intégration	Classement	Prétraitement
1	CBOW	SVM	
2	Skipgram	SVM	
3	PV-DBOW	ms	
4	PV-DM	ms	
5	Doc2Vec	SVM	
6	FastText	SVM	
7	FastText-cl		
8	FastText-cl		d-outlier
9	FastText-cl		cl-outlier
10	FastText-cl		bigram
11	FastText	SVM	b-sampling

Ⅲ. Modèle IA

❍ Modèle IA – Intégration

- ✓ **Word2Vec**
  - **– Convertit un mot en vecteur en maintenant la similarité sémantique**
  - **– Les mots utilisés dans les mêmes contextes ont tendance à avoir une signification similaire**

- ✓ **Doc2Vec**
  - **– Word2Vec avec ID du document**
  - **– Reflète la relation entre les mots et le document**

- ✓ **FastText**
  - **– Word2Vec + Subword**
  - **– Classement FastText : met en œuvre l'intégration et le classement en même temps**

Ⅲ. Modèle IA

❍ Modèle IA – Classement, prétraitement

Scénario	Intégration	Classement	Prétraitement
1	CBOW	SVM	
2	Skipgram	SVM	
3	PV-DBOW	ms	
4	PV-DM	ms	
5	Doc2Vec	SVM	
6	FastText	SVM	
7	FastText-cl		
8	FastText-cl		d-outlier
9	FastText-cl		cl-outlier
10	FastText-cl		bigram
11	FastText	SVM	b-sampling

Support Vector Machine

most_similar()

Ⅳ. Évaluation

❍ Évaluation des modèles IA

- ✓ **Crée 10 modèles pour chaque chapitre de la Nomenclature**
- ✓ **Évalue la performance pour 5 chapitres sélectionnés de manière aléatoire**

❍ Précision : la plus élevée

SH	Scénario	précision	recall	F1-score	acc@3top	acc@5top
Chapitre 40	**FastText-cl**	0,89	0,89	0,88	0,96	0,97
Chapitre 62	**FastText+SVM**	0,71	0,7	0,69	0,91	0,96
Chapitre 73	**FastText-cl**	0,75	0,74	0,74	0,87	0,91
Chapitre 61	**FastText-cl+bigram**	0,76	0,76	0,76	0,89	0,93
Chapitre 33	**FastText-cl+bigram**	0,8	0,8	0,79	0,93	0,96

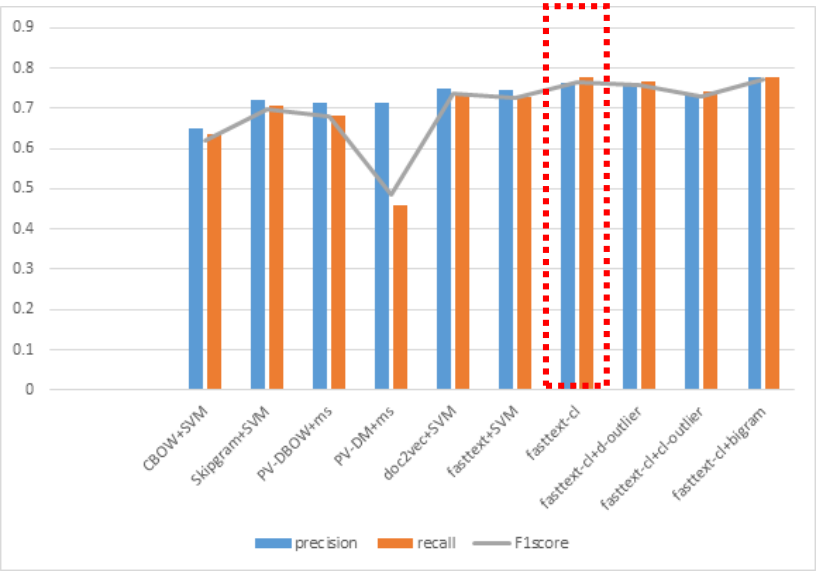
❍ Précision : la plus faible

SH	Scénario	précision	recall	F1-score	acc@3top	acc@5top
Chapitre 40	**PV-DM+ms**	0,79	0,49	0,52	0,68	0,76
Chapitre 62	**CBOW+SVM**	0,56	0,57	0,54	0,76	0,83
Chapitre 73	**CBOW+SVM**	0,66	0,6	0,59	0,8	0,86
Chapitre 61	**CBOW+SVM**	0,63	0,61	0,6	0,81	0,87
Chapitre 33	**CBOW+SVM**	0,58	0,58	0,55	0,8	0,88

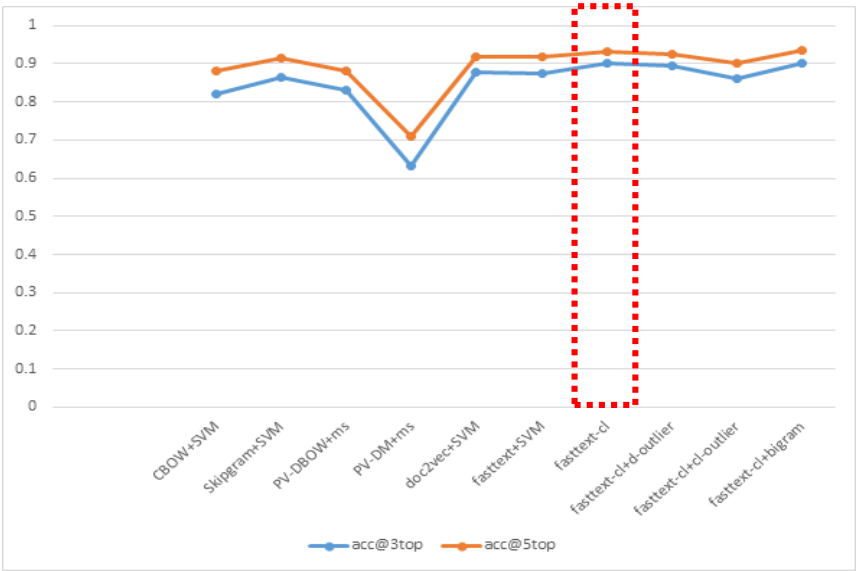
Ⅳ. Évaluation

❍ Résultats de l'évaluation

Précision moyenne, recall, F1-score



Moyenne acc@3top, acc@5top



V. Conclusion

❍ Classement dans le SH en utilisant l'IA