EXPLOITATION DE COMPTE-RENDUS MÉDICAUX GRÂCE AUX WORD EMBEDDINGS

Mikaël Dusenne

2020-02-04

INTRODUCTION

80% des données cliniques pertinentes sont non structurées

Apprentissage automatique et langage naturel

Approches classiques : un mot / n-gram = une variable

Problèmes :

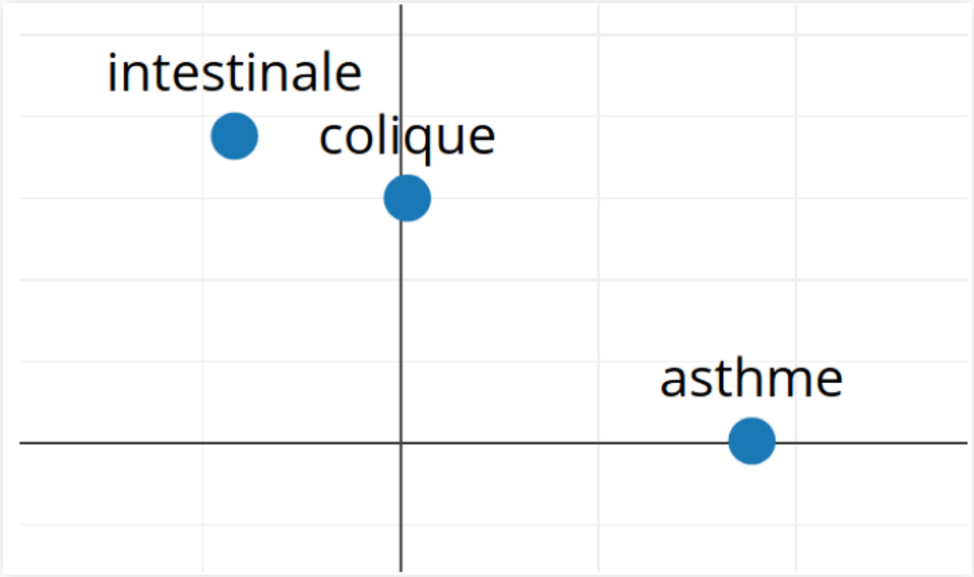
- Pas de notion de **distance sémantique**
- Très grand **nombre de variables**
- Données **éparses**

mot	hospitalisé	asthme	occlusion	...	colique	intestinale	aigüe
asthme	0	**1**	0	...	0	0	0
colique	0	0	0	...	**1**	0	0
intestinale	0	0	0	...	0	**1**	0

Word Embeddings

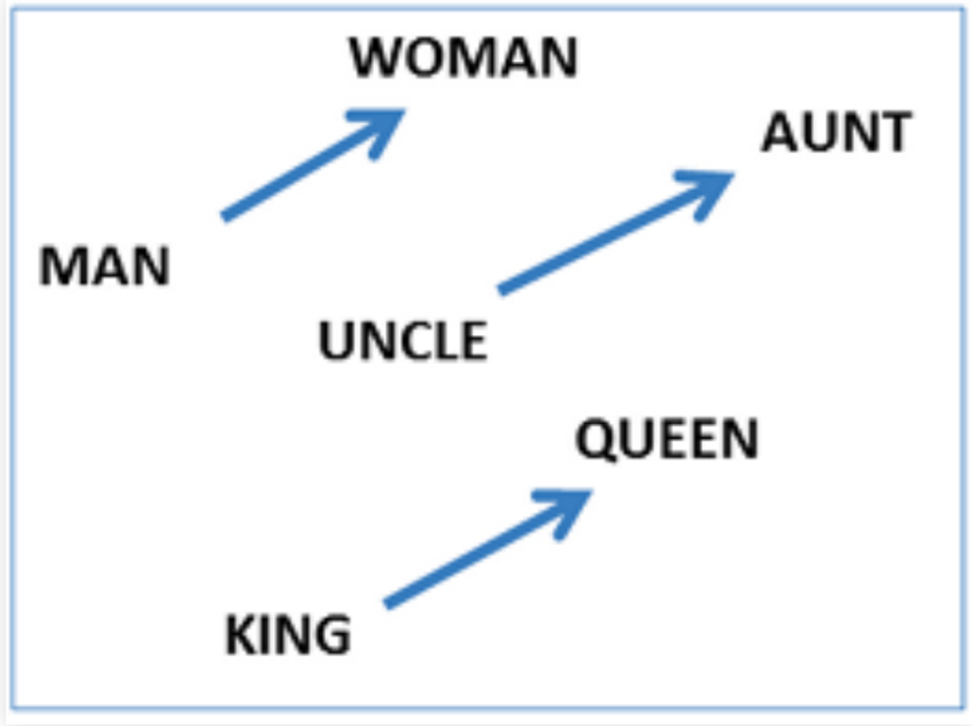
- Représentation **dense** des mots
- Vecteurs de **nombres réels**
- Dimension **indépendante de la taille du vocabulaire**
- Proximité dans l'espace vectoriel corrélée à la **similarité sémantique**

mots	0	1
asthme	0.888	0.014
colique	0.017	1.500
intestinale	-0.420	1.880



Word Embeddings

Les Embeddings permettent d'utiliser le calcul vectoriel pour effectuer des transformations sémantiques



King + (Woman - Man) = Queen

Embeddings et TAL : implémentations

- **2013** : Word2Vec 2013 :
  - réseau de neurones pour créer les embeddings
- **2014** : GloVe
  - "global vectors", matrice de co-occurence utilisant le corpus entier
- **2014** : Doc2Vec
  - Vecteurs de Documents
- **2016** : FastText
  - Décomposition des mots en n-grams de caractères
- **2018** : ELMo
  - utilise l'ordre des mots (LSTM bi-directionnel)
- **2018** : BERT
  - utilise des "attention network" (Transformer)
  - gestion des homonymes
- **2018** : Flair
  - Zalando Research
  - PoS tagging, named entities recognition
- **2019** : ALBERT
  - Améliore BERT : moins de paramètres, entraînement plus rapide

Application des embeddings aux compte-rendus médicaux

- **Word Embeddings** *(non supervisé)* :
  Enrichissement de l'annotation sémantique

- **Document Embeddings** *(supervisé)* :
  Prédiction du type de document

- **"Séjour Embeddings"** *(supervisé)* :
  Aide au codage de l'activité hospitalière

- **"Patient Embedding"** *(non supervisé)* :
  Création de cohortes pour la recherche, aide au diagnostic

Thèse Emeric Dynomant

- Co encadrement Pr. Stéphane Canu & Stéfan Darmoni
- Bourse CIFRE : société OMICX
- Sujet : Bioinformatics articles structuring with an end-to-end processing pipeline
- Soutenance prévue : premier semestre 2020
- Machine Learning for NLP; word & document embeddings for text
- Word embeddings
- Comparaison de cinq algorithmes sur 11,8 M de documents de santé d’un EDS
- Document embeddings
- Doc2Vec2PubMed vs. algorithme actuel Related Articles

Medical word embeddings querying page

12M | endocardite | Search

GloVe	infectieuse, myocardite, eto, native, streptocoque, bovis, bactériémie, faecalis, _endocardite
FastText (CBOW)	proprio_septive, myopericardite, septo_optique, endo_aortique, endocartite, endoculaire, acrodermite, rhino_septale, salmonellose, épidermolyse
Word2Vec (Skip-Gram)	bovis, sanguinis, eto, gordonii, gallolyticus, aorto_mitrale, mutans, infectieuse, streptoccoque, salivarius
FastText (Skip-Gram)	endocartite, endocardique, proctologique, extancilline, septo_basale, prolongements, recanalisée, précentrale, dantrolene, podoscopique
Word2Vec (CBOW)	endocartite, _endocardite, native, bovis, médiastinite, myocardite, mutans, gallolyticus, myopéricardite, tamponnade

Home

Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, Canu S, Darmoni SJ. Word Embedding for the French Natural Language in Health Care: Comparative Study. JMIR Med Inform. 2019 Jul 29;7(3):e12310.

Word embeddings dans deux contextes différents

QUERY : "facebook"

LiSSa corpus (300k)	internet, twitter, web, blog, e_learning, blogs, internautes, tic, game, …
RUH documents (12M)	reproches, injures, messages, insultes, rumeurs, ex_conjointe, menaces, insultant, …

Espace vectoriel disponible pour la communauté scientifique

Valorisation ???

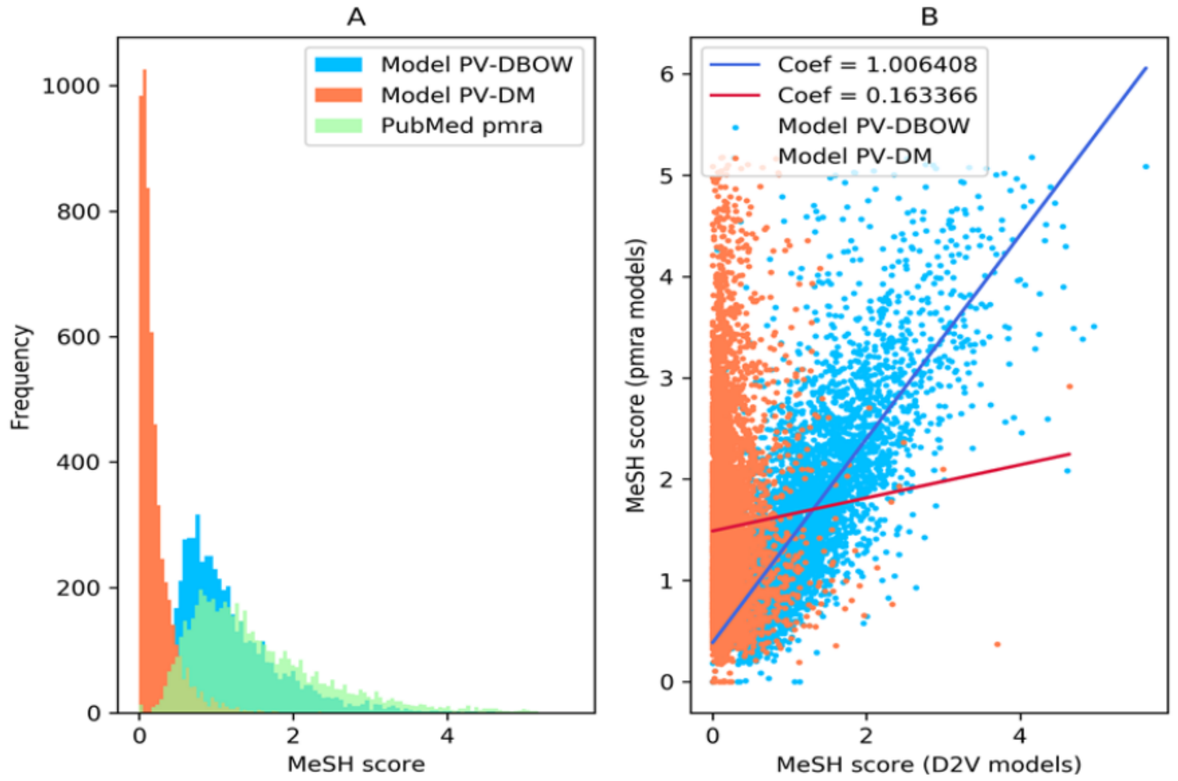
Doc2VecPubMed

Emeric Dynomant et coll.

Doc2Vec on the PubMed corpus: study of a new approach to generate related articles,

HAL

https://arxiv.org/abs/1911.11698



Thèse de science

Co-directeurs :

- Professeur S.J. Darmoni
- Professeur S. Canu

Co-Encadrant :

- Docteur J. Grosjean

Objectifs :

Exploration des applications des Embeddings aux documents médicaux :

évaluation, application en vie réelle sur les documents de l'entrepôt du CHU de Rouen

Date de début : Octobre 2019

Thèse de science

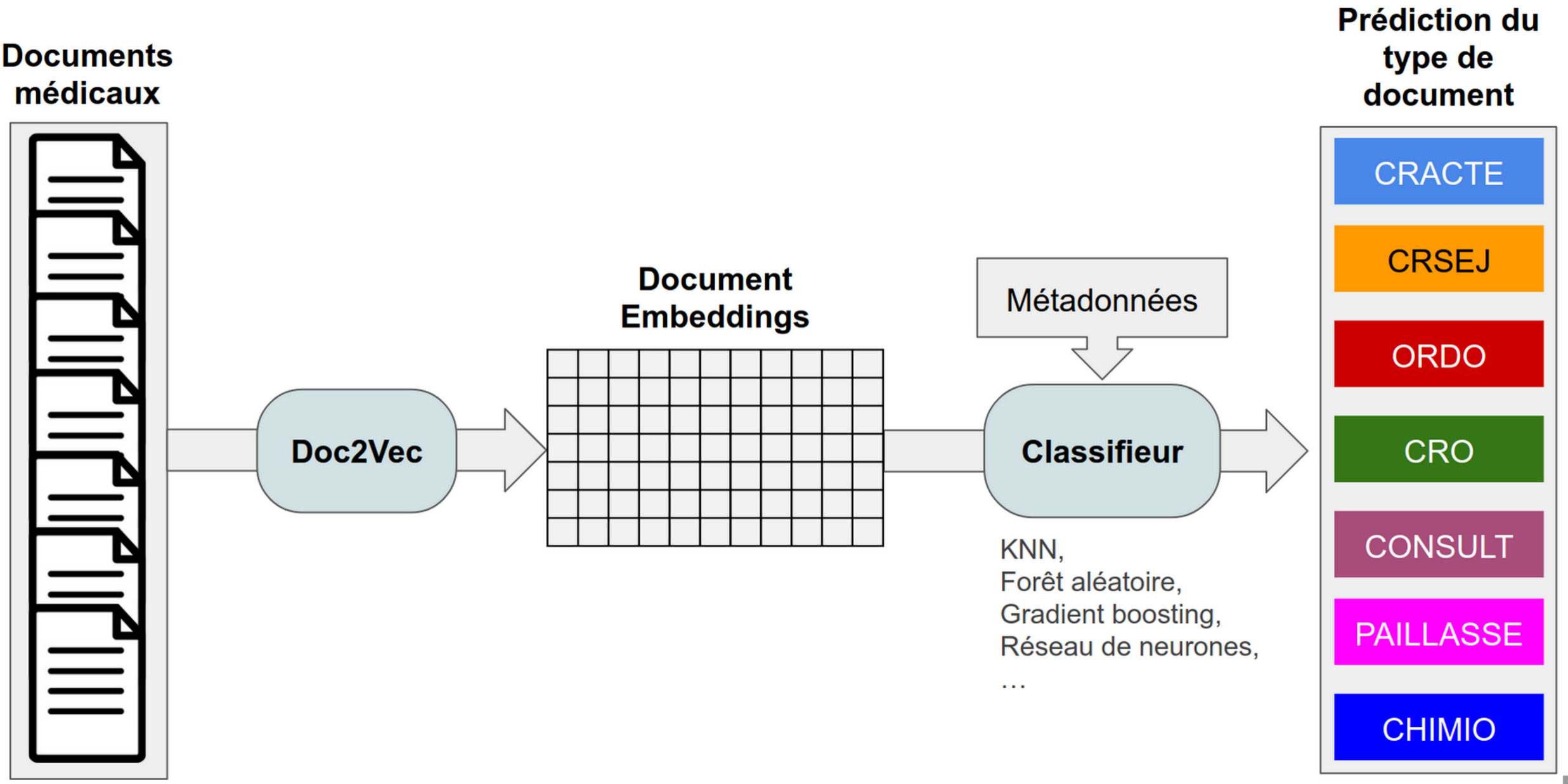
- Documents médicaux au CHU de Rouen :
  - *"Big Data"* : ≈ **17 millions de documents**

Problématique 1 :

Types de documents :

- Compte rendu de séjour / d'acte / opératoire, ordonnance, consultation, …
- Métadonnée existante dans le système d'information hospitalier
- Incomplète : ≈ **10% non typés**

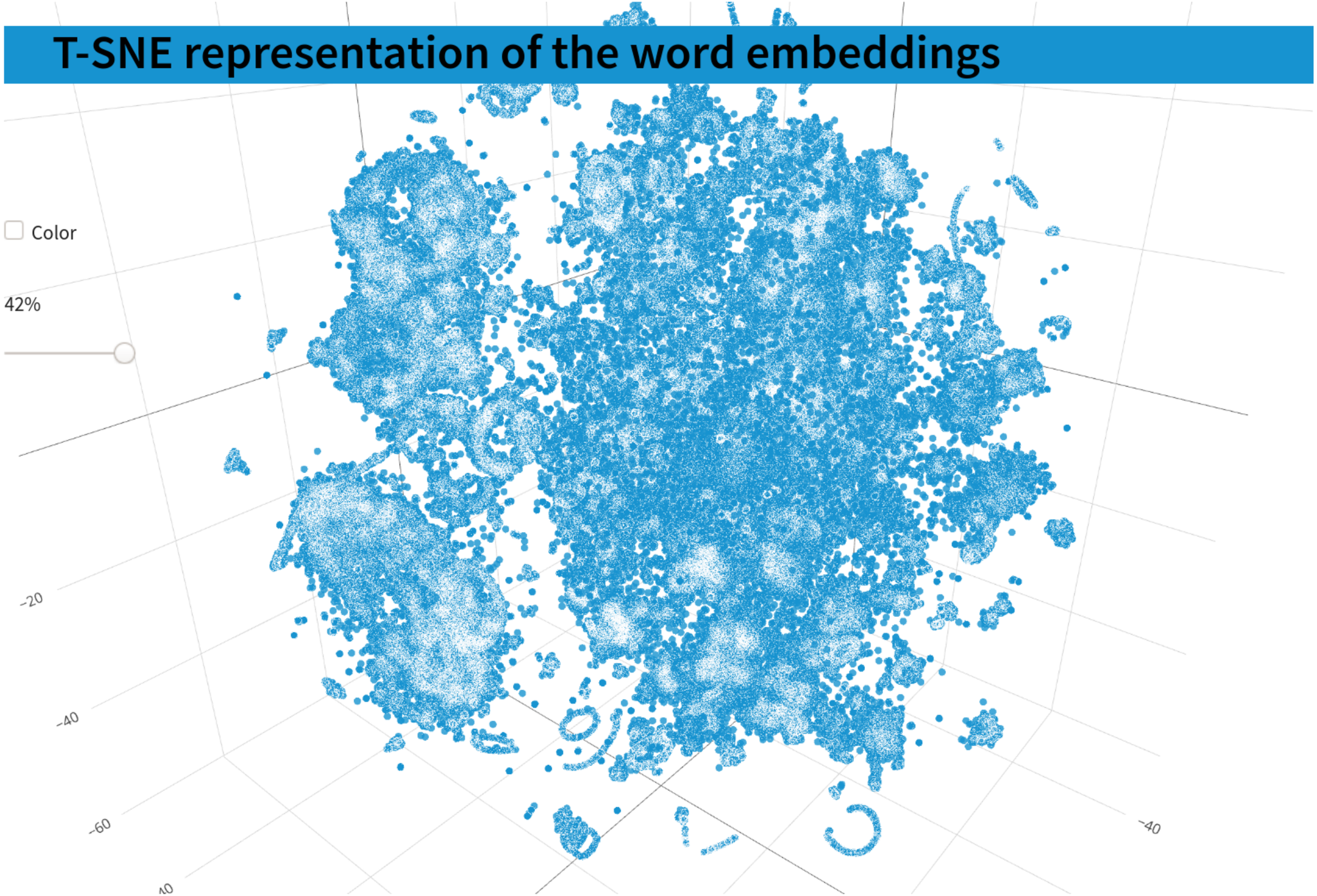
Classification des documents médicaux



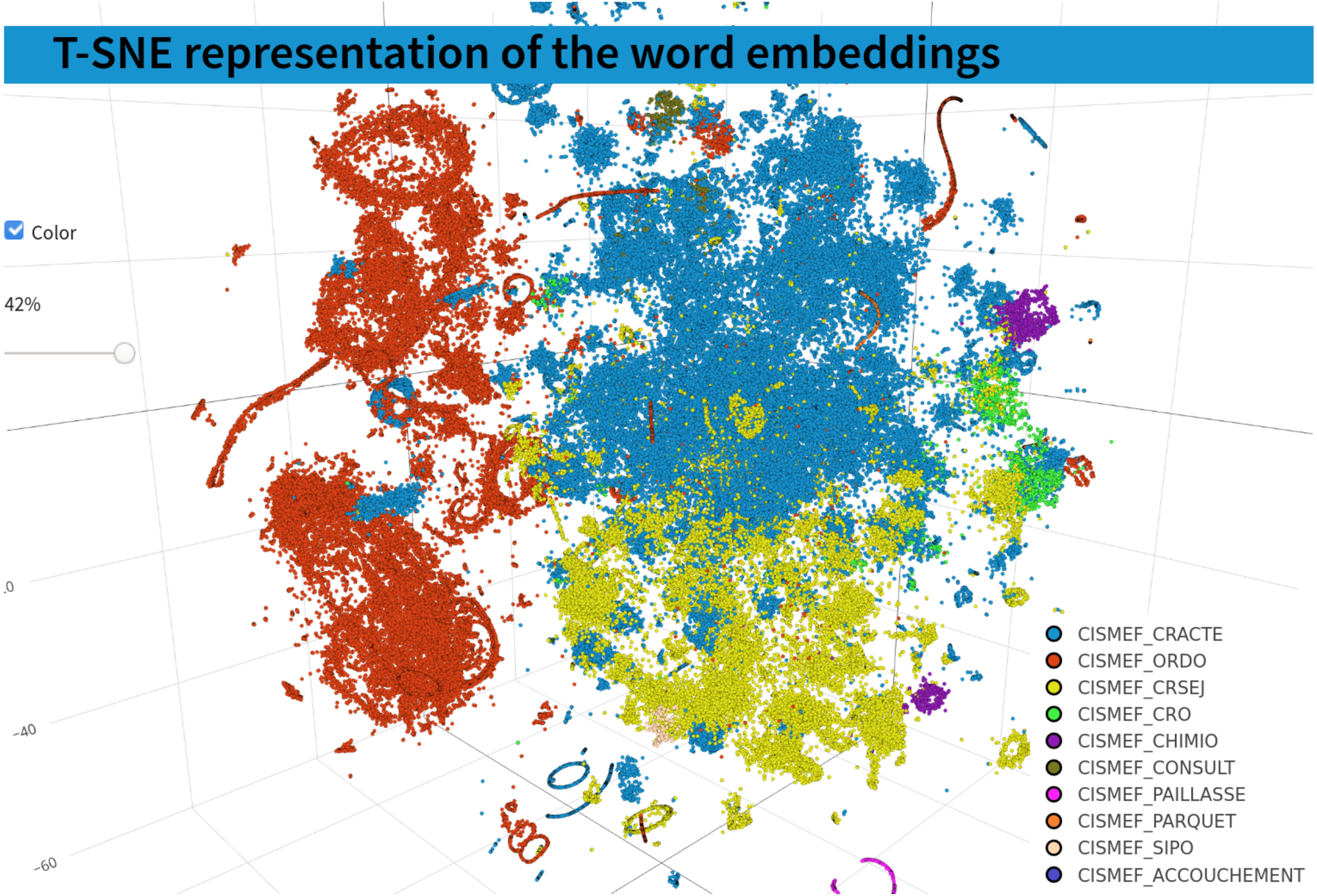
Accuracy: 98.57 %

pred \ actual	ACCOUCHEMENT	CHIMIO	CONSULT	CRACTE	CRO	CRSEJ	ORDO	PAILLASSE	PARQUET	SIPO
ACCOUCHEMENT	7	0	0	0	0	0	0	0	0	0
CHIMIO	0	60	0	0	0	0	0	0	0	0
CONSULT	0	0	30	6	0	0	1	0	0	0
CRACTE	0	0	6	4989	23	22	3	0	0	0
CRO	0	0	0	2	173	3	0	0	0	0
CRSEJ	0	2	1	54	7	1778	1	0	0	0
ORDO	0	0	0	11	0	0	2783	0	0	0
PAILLASSE	0	0	0	0	0	0	0	17	0	0
PARQUET	0	0	0	0	0	0	0	0	16	0
SIPO	0	0	0	0	0	1	0	0	0	4
Class Accuracy	**100**	**96.77**	**81.08**	**98.56**	**85.22**	**98.56**	**99.82**	**100**	**100**	**100**

T-SNE representation of the word embeddings



T-SNE representation of the word embeddings



Conclusion

- Exploration de l'exploitation des documents médicaux par les techniques d'embeddings
- Premiers résultats satisfaisants pour la classification des documents
  - bonnes performances
  - permettant la complétion des données manquantes + correction potentielle de données erronées (en cours d'évaluation)
  - possibilité d'utiliser d'autres métadonnées pour la classification
- Suite:
  - évaluation manuelle de la classification des documents
  - mise en oeuvre des autres applications des Embeddings

MERCI

mikaeldusenne@gmail.com