EXPLOITATION DE COMPTE-RENDUS MÉDICAUX GRÂCE AUX WORD EMBEDDINGS

Mikaël Dusenne

2020-03-04

INTRODUCTION

80% des données cliniques pertinentes sont non structurées

Comment représenter efficacement les données textuelles pour l'apprentissage automatique?

Apprentissage automatique et langage naturel

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

Problèmes :

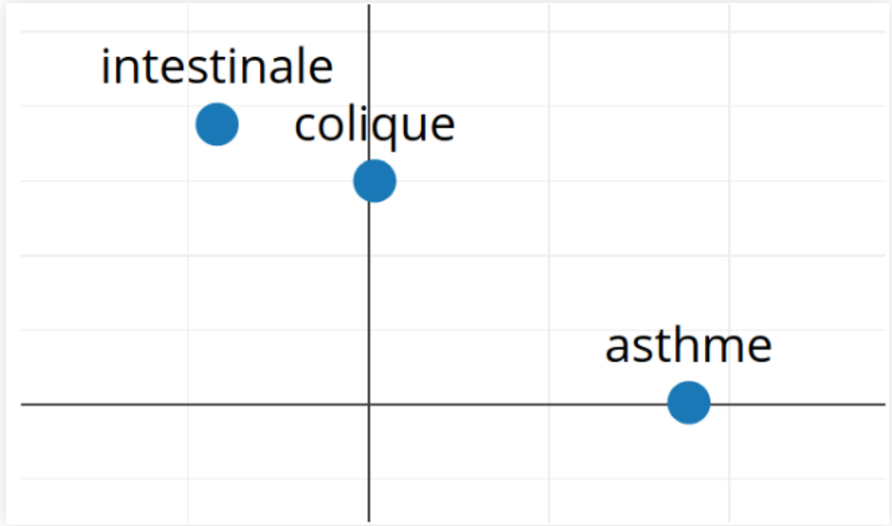
- Pas de notion de **distance sémantique**
- Très grand **nombre de variables**
- Données **éparses**

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

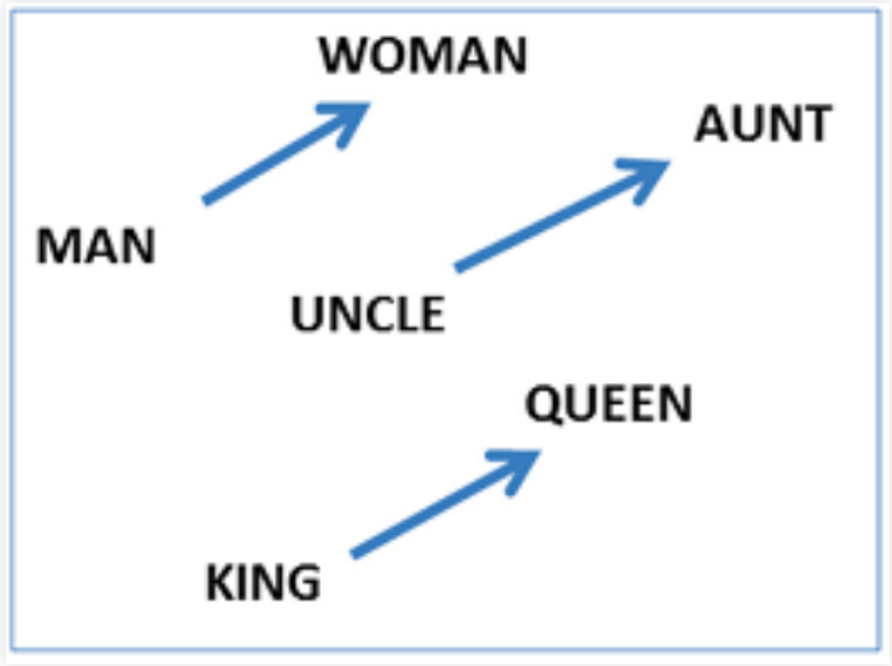
- Représentation **dense** des mots
- Vecteurs de **nombres réels**
- Dimension **indépendante de la taille du vocabulaire**
- Proximité dans l'espace vectoriel corrélée à la **similarité sémantique**

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

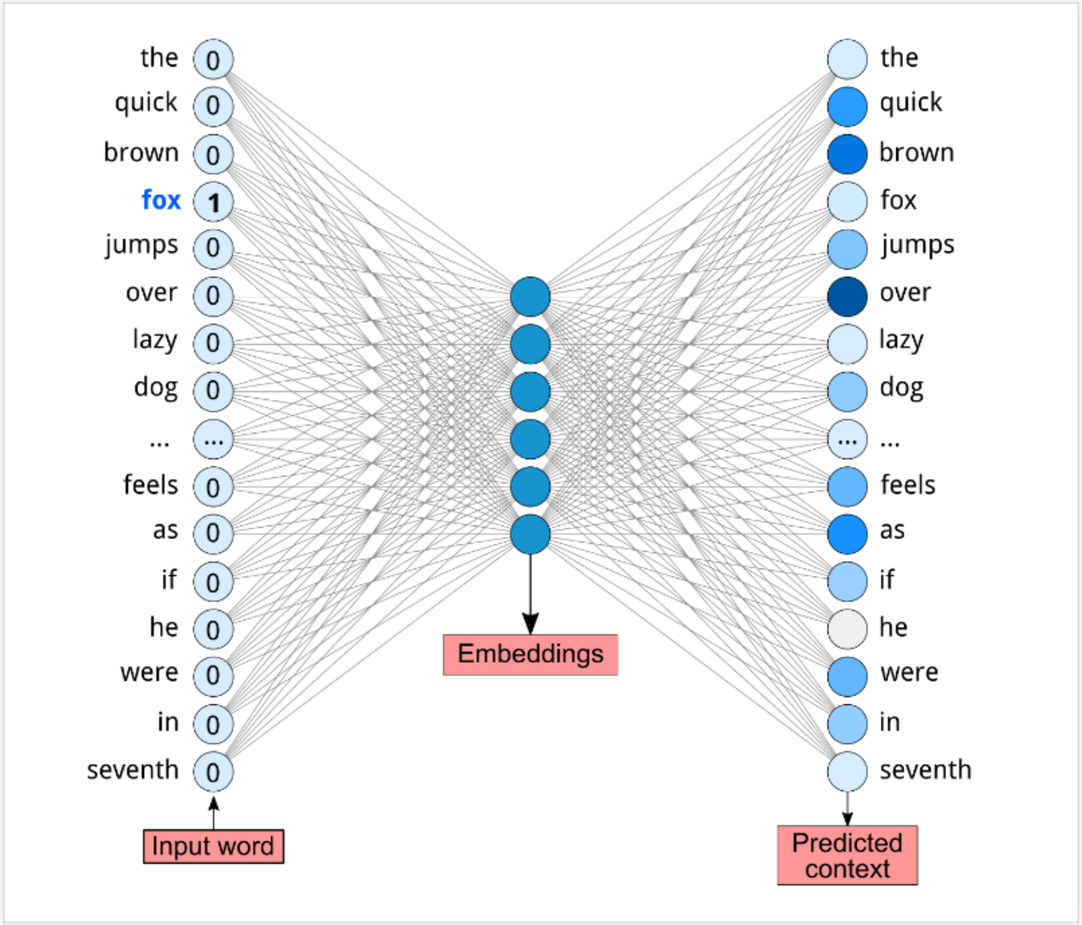


King + (Woman - Man) = Queen

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

Word Embeddings : word2vec



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 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

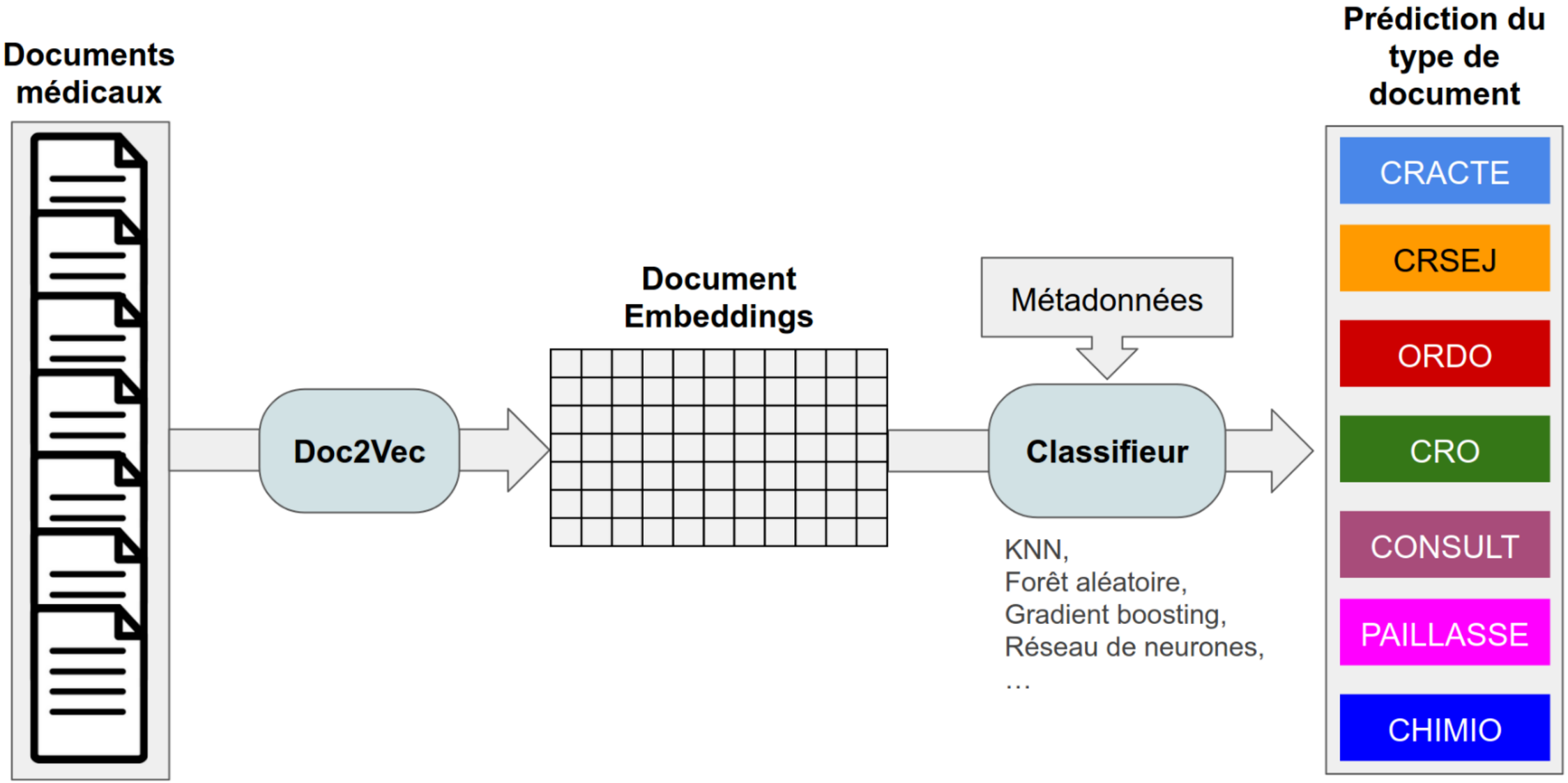
- Documents médicaux au CHU de Rouen :
  - *"Big Data"* : ≈ **17 millions de documents**

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**

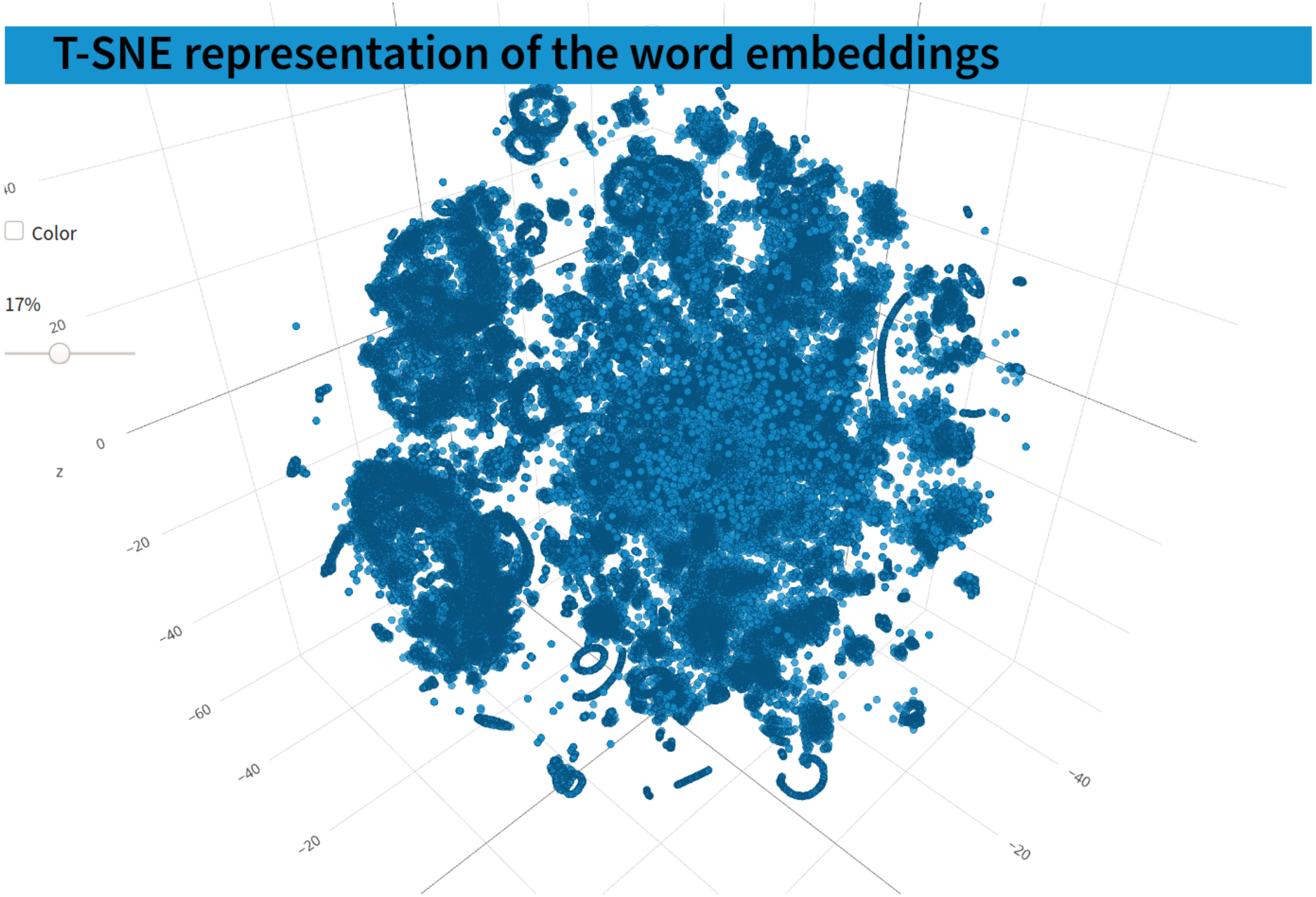
Classification des documents médicaux



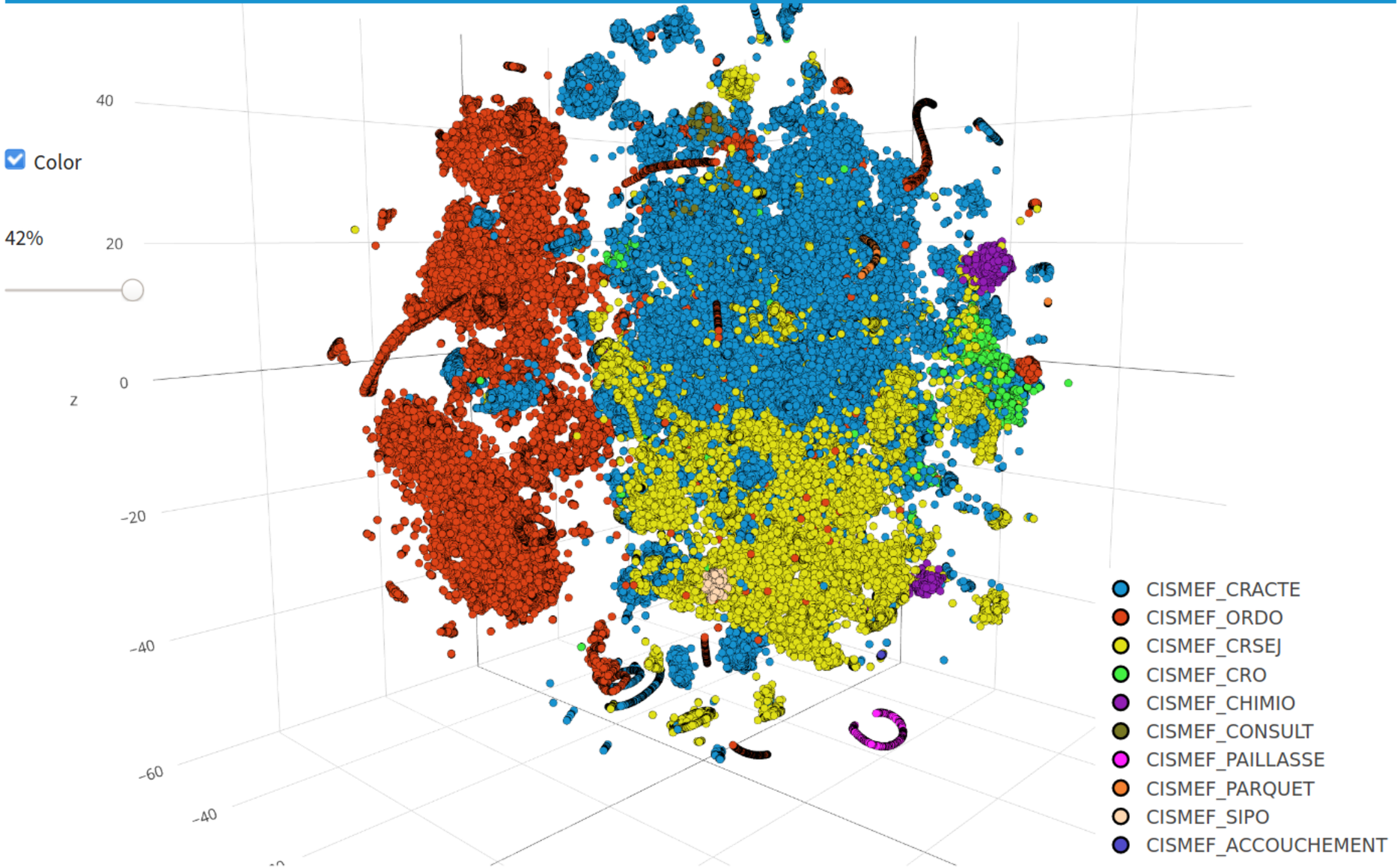
Accuracy: 98.37 %

pred \ actual	ACCOUCHEMENT	CHIMIO	CONSULT	CRACTE	CRO	CRSEJ	ORDO	PAILLASSE	PARQUET	SIPO
ACCOUCHEMENT	82	0	0	0	0	1	0	0	0	0
CHIMIO	0	1892	0	0	0	8	0	0	0	0
CONSULT	0	0	948	147	0	0	7	0	0	0
CRACTE	1	2	231	141236	530	721	98	2	16	0
CRO	0	0	1	178	4723	32	0	0	0	0
CRSEJ	5	81	13	1951	163	49221	21	0	0	6
ORDO	0	0	0	338	0	17	77872	0	0	0
PAILLASSE	0	0	0	0	0	0	0	636	0	0
PARQUET	0	0	0	5	0	0	0	0	472	0
SIPO	0	0	0	0	0	20	0	0	0	205
Class Accuracy	**93.18**	**95.8**	**79.46**	**98.18**	**87.2**	**98.4**	**99.84**	**99.69**	**96.72**	**97.16**

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

1. Mikolov, Tomas; et al. (2013). Efficient Estimation of Word Representations in Vector Space
2. Pennington, Jeffrey, et al. “Glove: Global Vectors for Word Representation.”
3. Le, Q. V. & Mikolov, T. "Distributed Representations of Sentences and Documents"
4. Bojanowski, Piotr, et al. "Enriching Word Vectors with Subword Information."
5. Peters, Matthew E., et al. "Deep Contextualized Word Representations."
6. Devlin, Jacob, et al. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.”
7. Akbik, A.; Blythe, D. & Vollgraf, R. Contextual string embeddings for sequence labeling
8. Lan, Zhenzhong, et al. “ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations.”

MERCI

mikaeldusenne@gmail.com