Institut
québécois
d'intelligence
artificielle

Applications : Extraction de texte

Joseph Paul Cohen, PhD (joseph@josephpcohen.com)
Margaux Luck, PhD (margaux.luck@mila.quebec)

1er décembre 2018

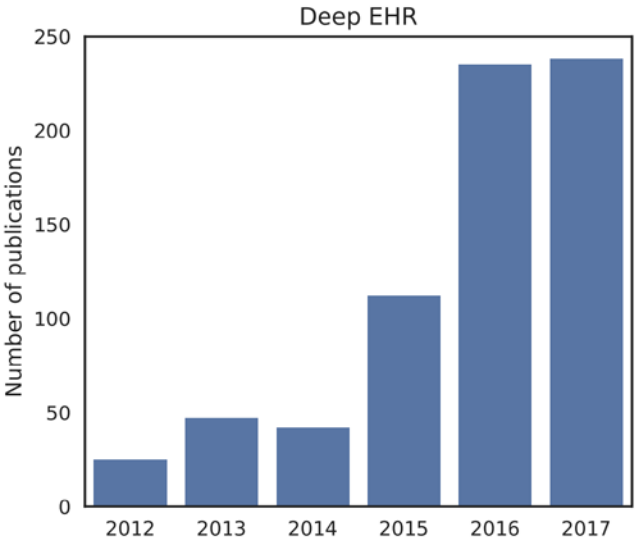
Extraction de texte

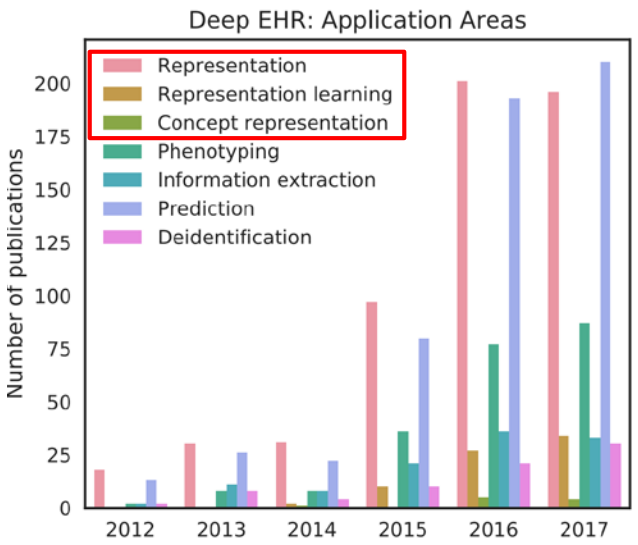
Revue par Joseph Paul Cohen, PhD
Mila - Institut Québécois d'Intelligence Artificielle

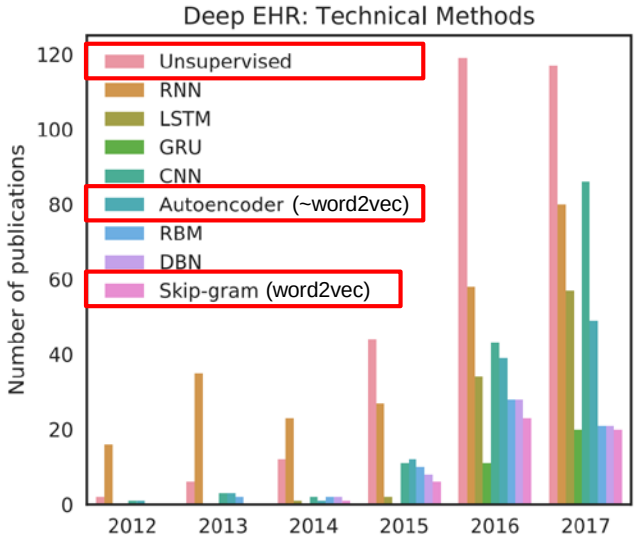
Sujets:

1. Représentation de concepts médicaux
2. Word2Vec

Où en est la recherche en apprentissage profond ?







Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis

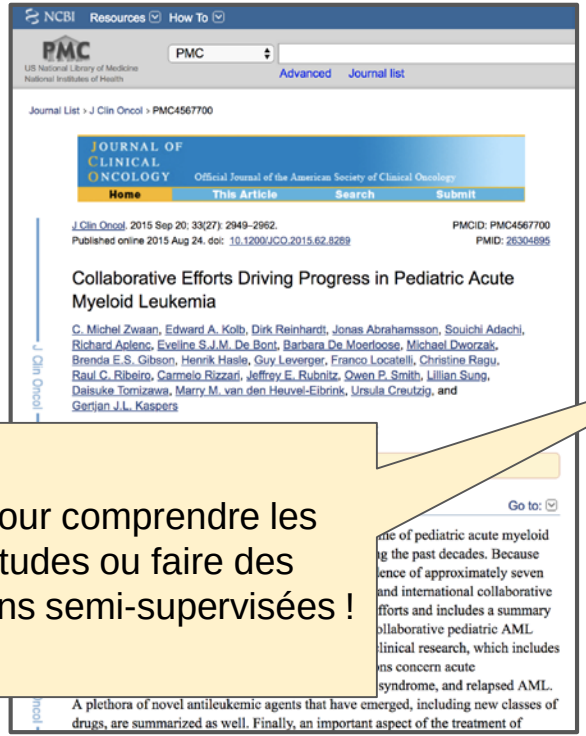
[Shickel, Deep EHR : A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record, 2018]

Représentation de concepts

Note clinique

Mr. Smith is a 63-year-old gentleman with coronary artery disease, hypertension, hypercholesterolemia, COPD and tobacco abuse. He reports doing well. He did have some more knee pain for a few weeks, but this has resolved. He is having more trouble with his sinuses. I had started him on Flonase back in December. He says this has not really helped. Over the past couple weeks he has had significant congestion and thick discharge. No fevers or headaches but does hav diffuse upper right-sided teeth **He denies any chest pains, palpitations, PND, orthopnea, ed syncope**. His breathing is doing No cough. He continues to smoke half-a-pack per day. He plans or **trying the patches again**.

Publication clinique



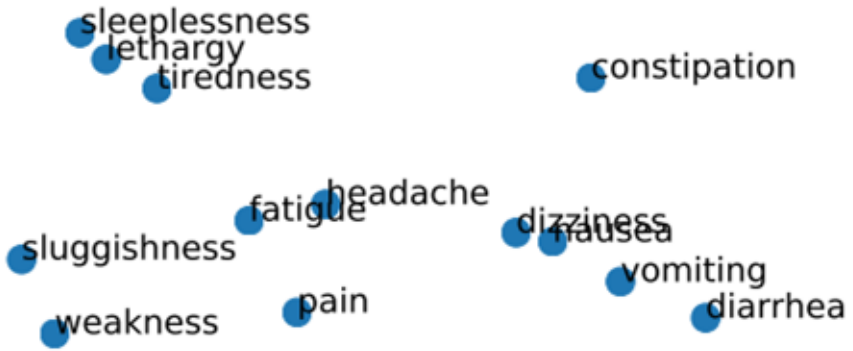
Utile pour comprendre les similitudes ou faire des prédictions semi-supervisées !

Représentations pour:

Les patients
Les médecins
Les visites
Les maladies
Les médicaments
Les symptômes

Encodage de mots dans le langage biomédical

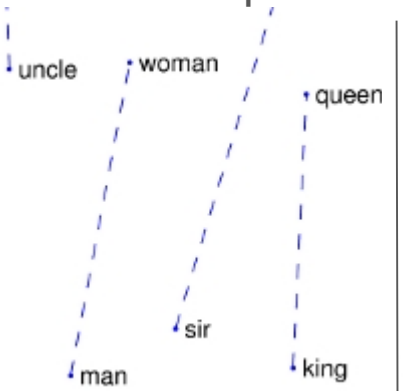
Extraire les relations entre les mots et produire une représentation latente



Que faire avec l’encodage de mots?

- Nous pouvons les composer pour créer des encodages de paragraphe.
- Utilisation à la place des mots pour un RNNs
- Augmenter les représentations apprises sur de petits ensembles de données

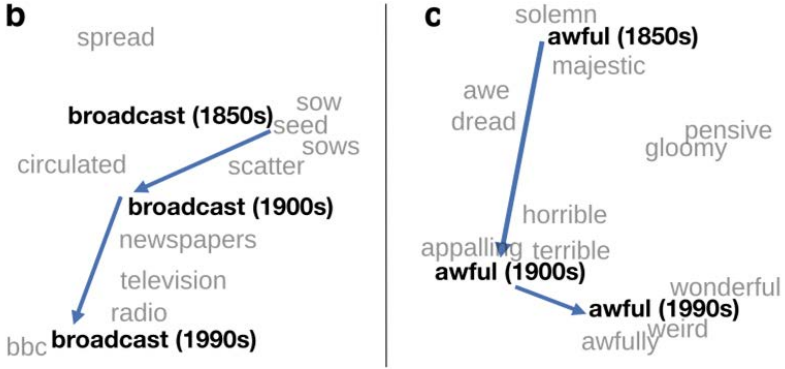
Étudier la compositionnalité de l’espace latent appris



[Mikolov, 2013]

[Pennington, 2014]

Étudiez la variation de sens entre deux textes (ou hôpitaux, ou médecins) ?



[Cultural Shift or Linguistic Drift, Hamilton, 2016]

Représentation des mots

Encodage One-hot : vecteur binaire par mots

Exemple :

chat = `[0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 … 0]`

chien = `[0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 … 0]`

maison = `[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 … 0]`

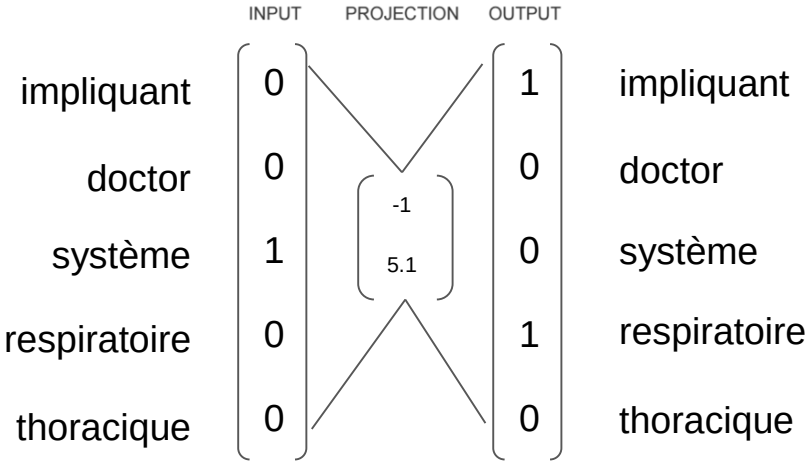
word2vec

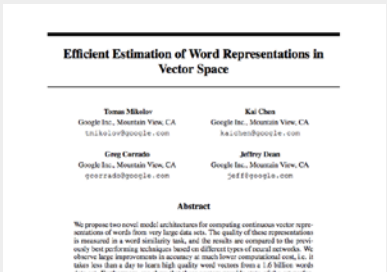
contexte fenêtre

Mot contextuel | Mot cible | Mot contextuel

impliquant le système respiratoire et d'autres symptômes thoraciques

1. Chaque mot est un exemple d'entraînement
2. Chaque mot est utilisé dans de nombreux contextes
3. Le contexte définie chaque mots





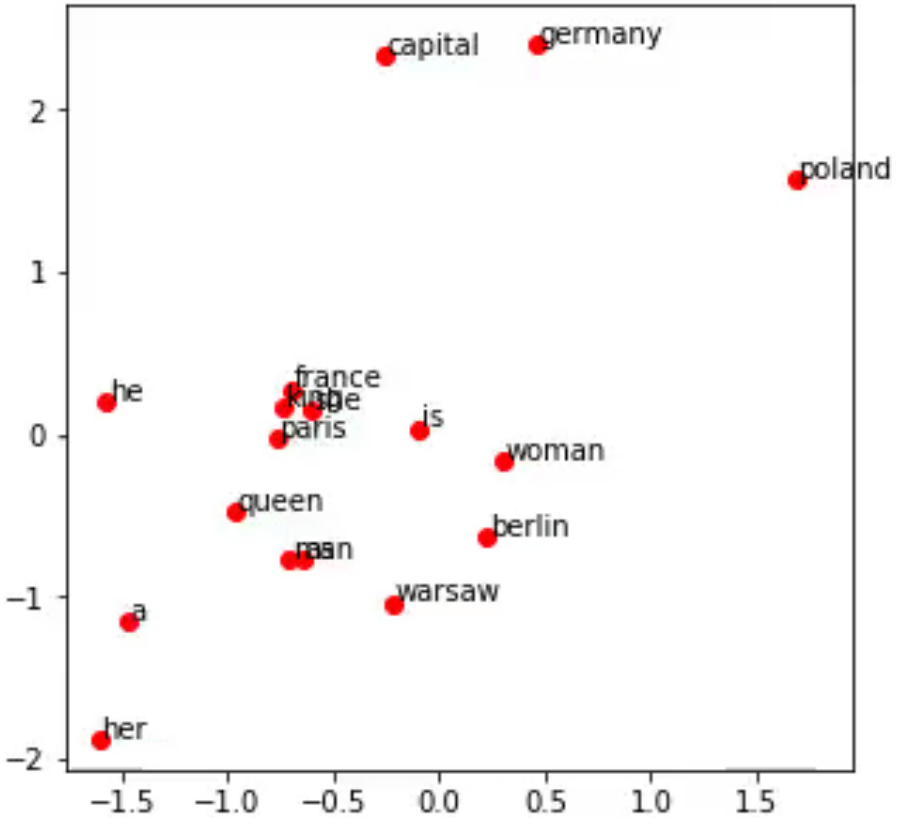
Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov, Google Inc., Mountain View, CA, tmikolov@google.com
Kai Chen, Google Inc., Mountain View, CA, kaichen@google.com
Greg Corrado, Google Inc., Mountain View, CA, gcorrado@google.com
Jeffrey Dean, Google Inc., Mountain View, CA, jeff@google.com

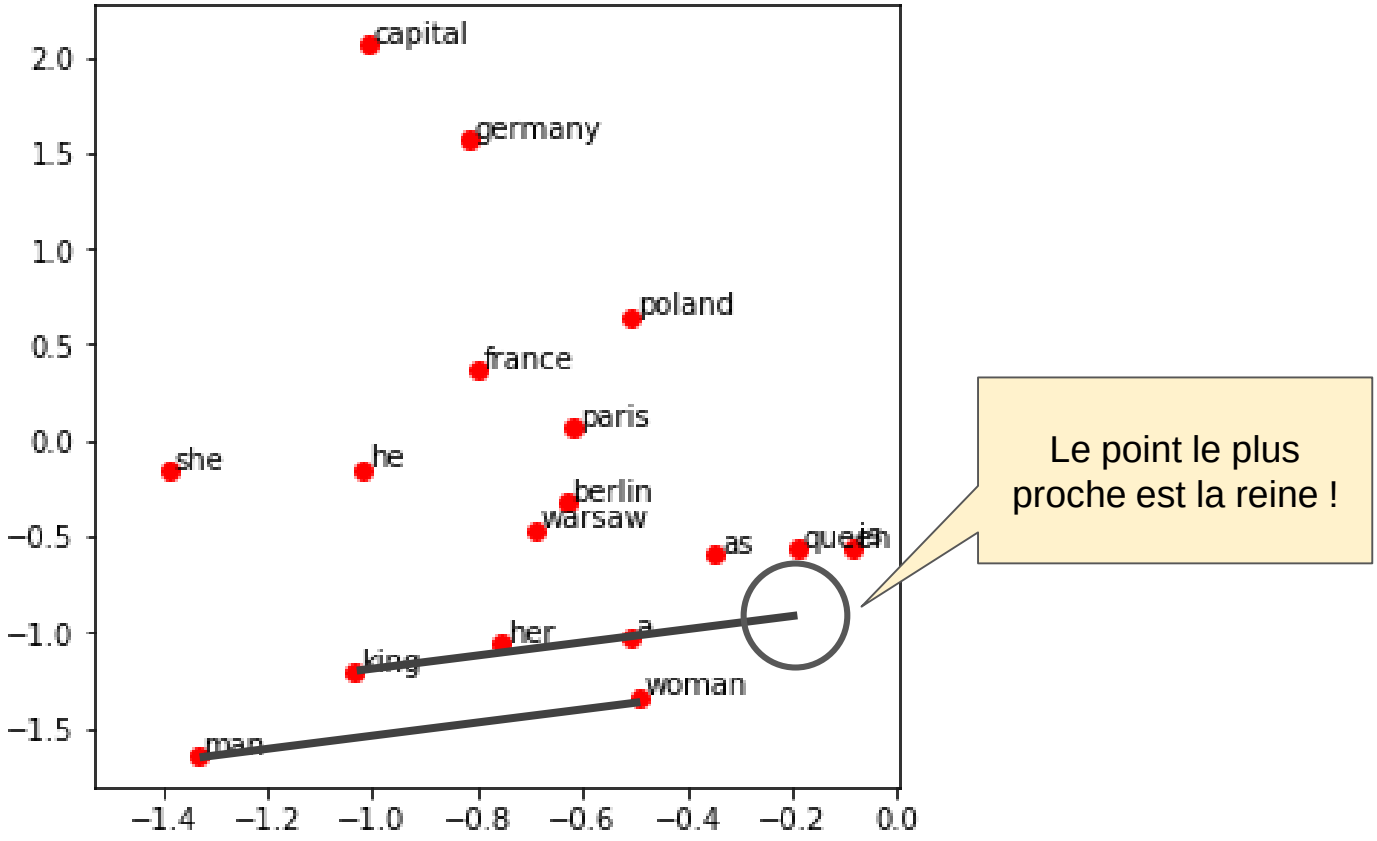
Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words

Mikolov, Efficient Estimation of Word Representations in Vector Space, 2013



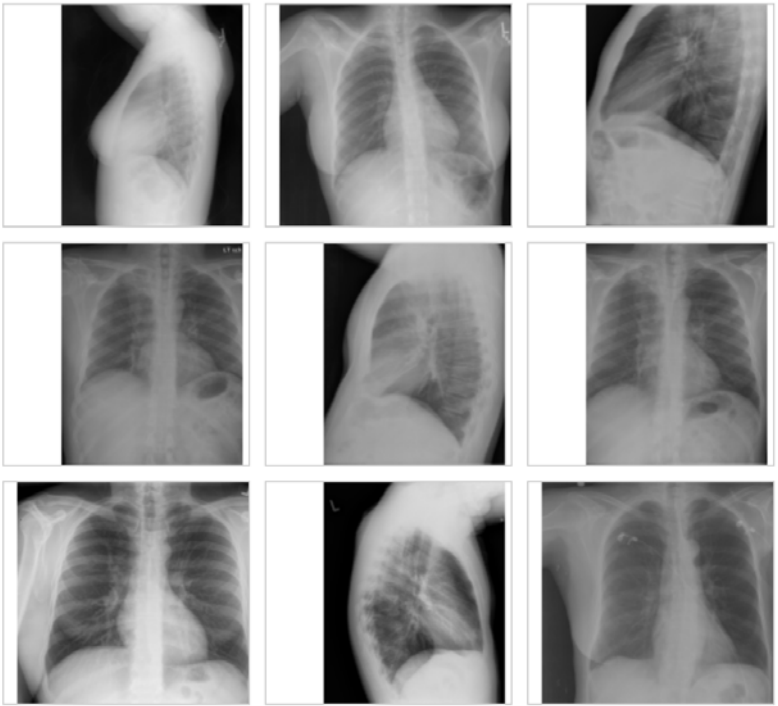
Apprentissage en cours



roi + (femme - homme) = ?

Rapports de l’hôpital universitaire de l'Indiana

Images de radiographie pulmonaire du réseau hospitalier de l'Université de l'Indiana



Indiana University Chest X-ray Collection

Kohli MD, Rosenman M - (2013)

Affiliation: Indiana University

ABSTRACT

Comparison: None.

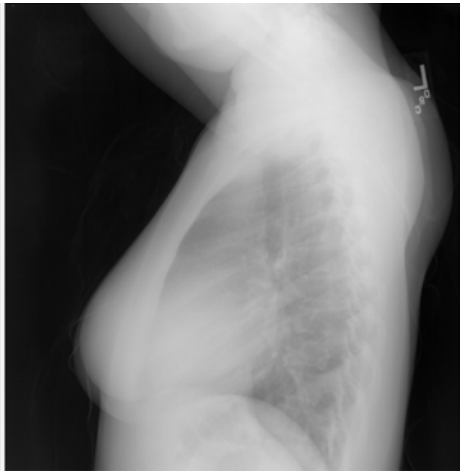
Indication: Positive TB test

Findings: The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.

Impression: Normal chest x-XXXX.

NOTE: The data are drawn from multiple hospital systems.

Show MeSH
Related in: MedlinePlus Request Collection



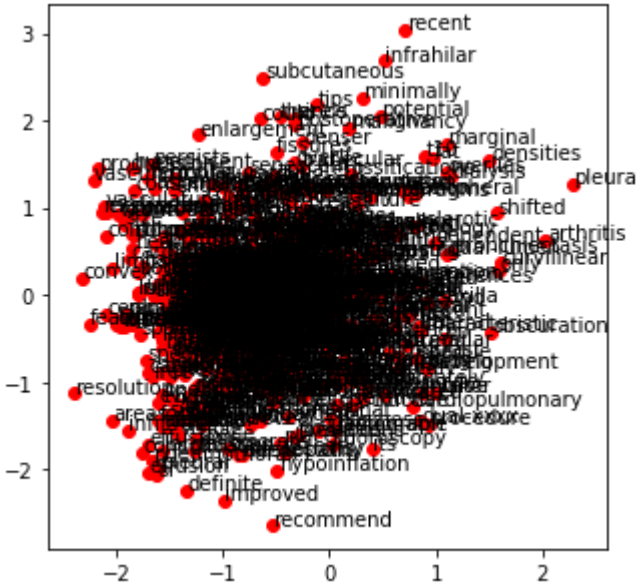
4000 rapports disponibles en format XML !

OPEN i

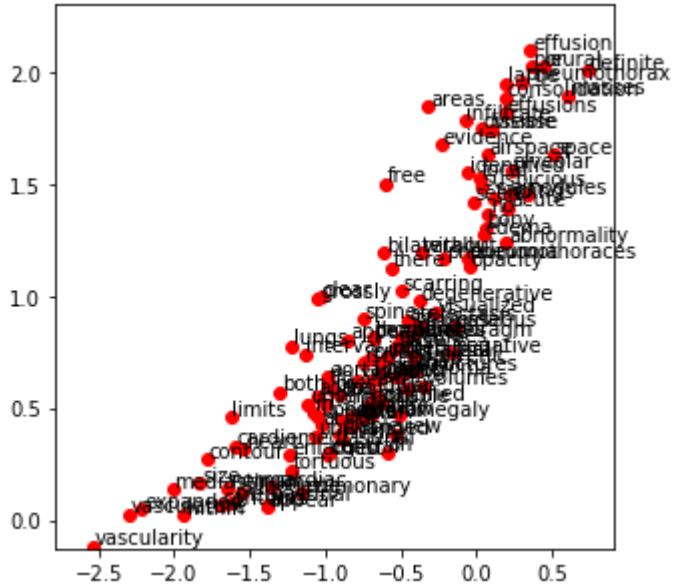
Hyperparamètres !

En fonction de la configuration du modèle, les encodages peuvent varier :

Aucune information



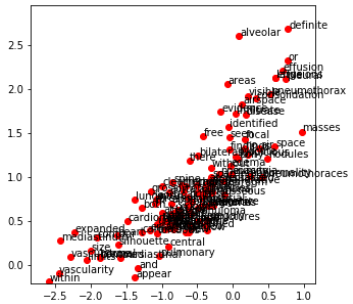
Compression



Nous pouvons varier :

la dimension de l'encodage, le taux d'apprentissage, les mots-clés, la taille de la fenêtre, etc.

La taille de l’encodage à de l’importance

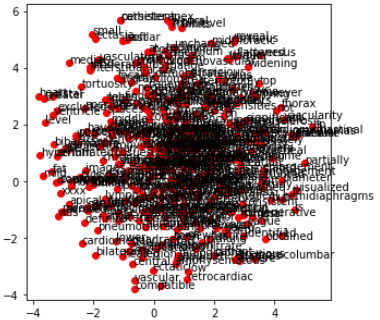


0	spine	0.000000
1	seen	6.632637
2	evidence	8.230038
3	infiltrate	13.233962
4	findings	17.762381
5	disease	19.574038
6	focal	22.267534
7	aortic	23.346777
8	suspicious	25.558247
9	acute	29.176380
10	within	29.798399

Dim = 2

0	spine	0.000000
1	severe	3.304547
2	changes	3.469765
3	anterior	3.474130
4	mild	3.731760
5	also	3.828588
6	thoracic	3.838329
7	noted	3.871284
8	calcifications	3.934170
9	vertebral	4.004414
10	stable	4.022851

Dim = 100

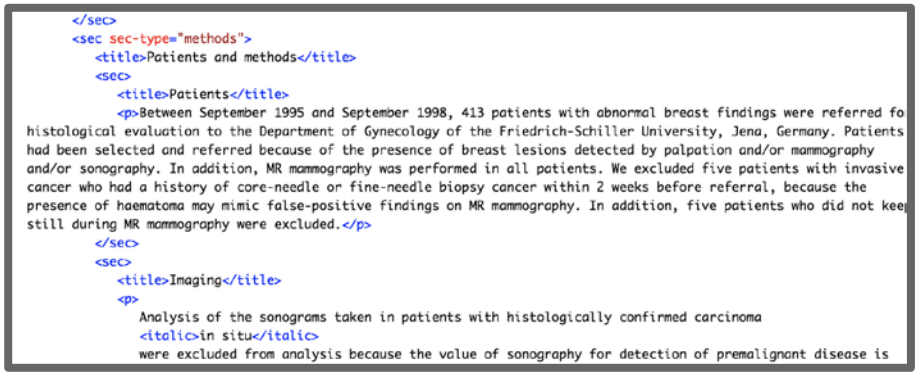


*via t-sne

Sous-ensemble à accès libre de PubMedCentral

Sous-ensemble de fichiers disponibles en libre accès

- Documents disponibles en format PDF ou XML
- 1,25 million d'articles biomédicaux et 2 millions de mots distincts
- Disponible par FTP pour téléchargement en masse
- Les métadonnées comprennent le nom de la revue et l'année.



```
      </sec>
      <sec sec-type="methods">
         <title>Patients and methods</title>
         <sec>
            <title>Patients</title>
            <p>Between September 1995 and September 1998, 413 patients with abnormal breast findings were referred fo
histological evaluation to the Department of Gynecology of the Friedrich-Schiller University, Jena, Germany. Patients
had been selected and referred because of the presence of breast lesions detected by palpation and/or mammography
and/or sonography. In addition, MR mammography was performed in all patients. We excluded five patients with invasive
cancer who had a history of core-needle or fine-needle biopsy cancer within 2 weeks before referral, because the
presence of haematoma may mimic false-positive findings on MR mammography. In addition, five patients who did not kee
still during MR mammography were excluded.</p>
         </sec>
         <sec>
            <title>Imaging</title>
            <p>
               Analysis of the sonograms taken in patients with histologically confirmed carcinoma
               <italic>in situ</italic>
               were excluded from analysis because the value of sonography for detection of premalignant disease is
```

Exemple de données XML

```
Breast Cancer Res.
Genome Biol.
Arthritis Res.
BMC Cell Biol.
…
```

Noms des journaux

https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Encodage de mots dans le langage biomédical

Les représentations sont biaisées par les données.

Nous pouvons l'utiliser à notre avantage pour contrôler le domaine.

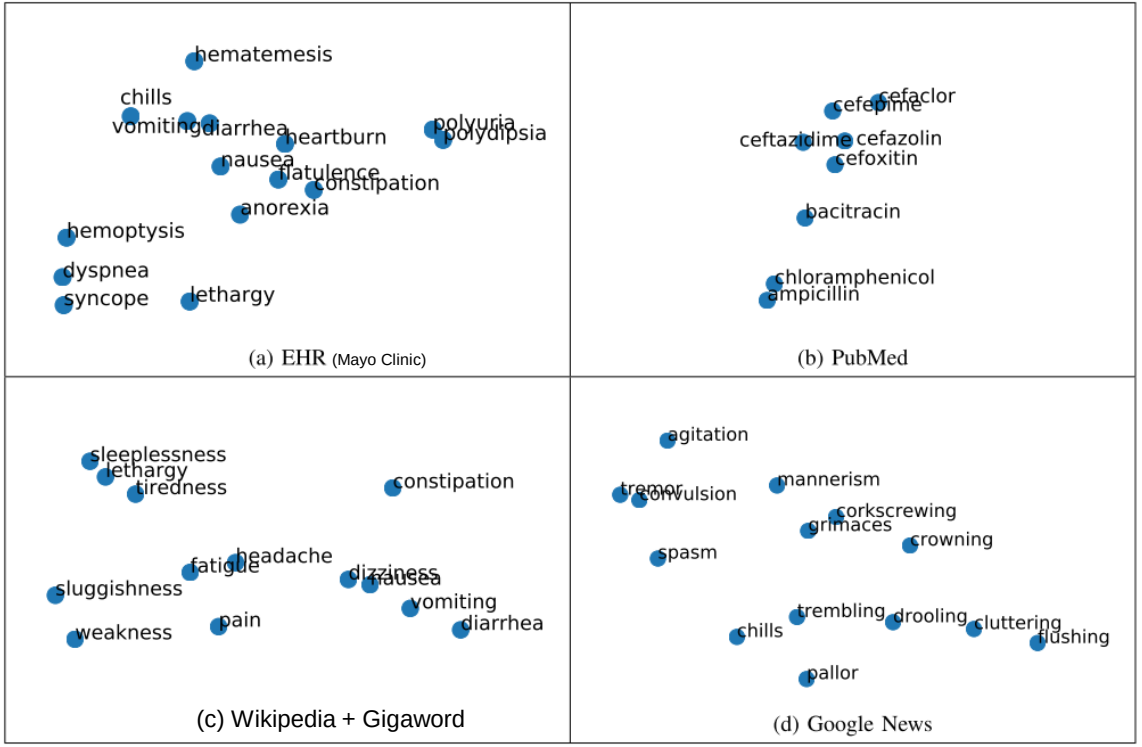
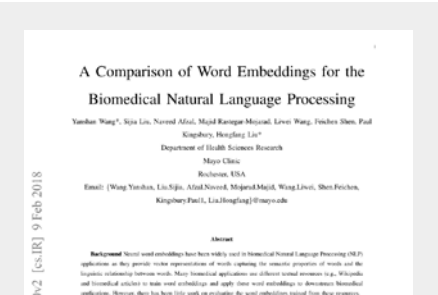


Fig. 1: Example clusters in different word embeddings.


A Comparison of Word Embeddings for the Biomedical Natural Language Processing

Wang, A Comparison of Word Embeddings for the Biomedical Natural Language Processing, 2018

Exemples de mots médicaux et leurs cinq mots similaires basés sur chaque corpus de texte de formation

Le nom complet du diabète est "diabète mellitus"

Selon certains articles, les diabétiques courent un risque accru d'hypertension (hypertension artérielle)

Un type d'ulcère gastro-duodénal

Un des symptômes est un ulcère

Le cancer du côlon est associé au cancer du sein !

	EHR (Mayo Clinic)	PubMed	Wikipedia + Gigaword	Google N
diabetes	mellitus,	cardiovascular,	hypertension,	diabetics,
	ntrolled,	nonalcoholic,	obesity,	hypertension,
	esterolemia,	obesity,	arthritis,	diabetic,
	pidemia,	mellitus,	cancer,	diabetes_mellitus,
	mellitis	polycystic	alzheimer	heart_disease
peptic ulcer disease	scleroderma,	gastritis,	ulcers,	ichen_planus,
	duodenal,	alcoholism,	arthritis,	Candida_infection,
	crohn,	rheumatic,	diseases,	vaginal_yeast_infections,
	gastroduodenal,	ischaemic,	diabetes,	oral_thrush,
	diverticular	nephropathy	stomach	dermopathy
colon cancer	breast,	breast,	breast,	breast,
	ovarian,	mcf,	prostate,	prostate,
	prostate,	cancers,	cancers,	tumor,
	postmenopausally,	tumor_suppressing,	tumor,	pre_cancerous_lesion,
	caner	downregulation	liver	cancerous_polyp

Spécifique ⟷ Général

Joseph Paul Cohen, PhD

Mila Medical

Myriam Côté, PhD

Pr. Yoshua Bengio, PhD

Mandana Samiei

Francis Dutil

Martin Weiss

Shawn Tan

Geneviève Boucher

Georgy Derevyanko, PhD

Tristan Sylvain

Margaux Luck, PhD

Sina Honari

Assya Trofimov

Vincent Frappier, PhD

Merci

Références

Ching, T., et al. **Opportunities And Obstacles For Deep Learning In Biology And Medicine**. Journal of The Royal Society Interface. 2018

Shickel, B. et al. **Deep EHR : A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record**. IEEE Journal of Biomedical and Health Informatics, 2018