Séminaire HiPhiS

Univ. de Montpellier, Univ. Paul Valéry, IRES, CNRS

Montpellier, France

Épistémologie de l'apprentissage machine

Pour un formalisme critique

Juan Luis Gastaldi

www.jlgastaldi.com

ETH zürich

15 Avril, 2025

Plan

Introduction

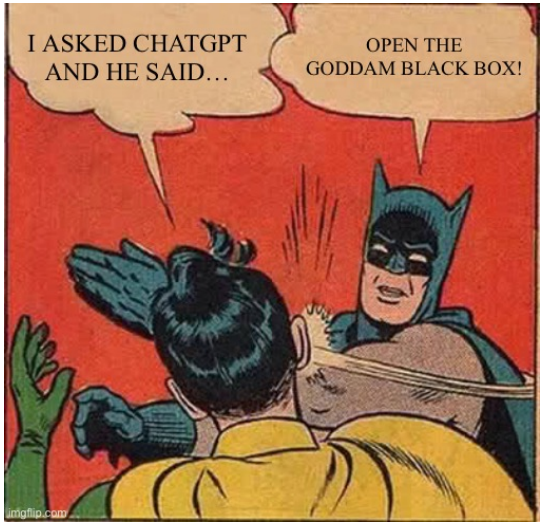
LLMs comme des objets formels

La structure des ‘embeddings’
- L’algèbre derrière les embeddings
- La structure derrière l’algèbre
- Les categories derrière la structure

Conclusion

Plan

L’épistémologie de l’IA



Vous avez dit “critique”?

Héritage kantien:

- Portant sur des phénomènes (non-dogmatique)
- Conscient des conditions de possibilité (fondant une legitimité)
- Concerné par d’autres possibilités (identifiant des limites)

Vous avez dit “critique”?

Vous avez dit “critique”?

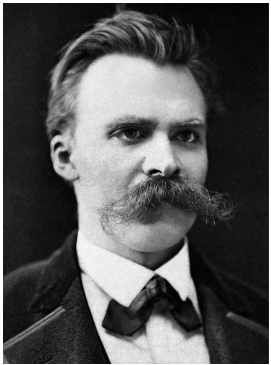
- Bonne critique “externaliste”
  - (cf. Ali et al., 2023)

- Critique “internaliste” médiocre
  - La référence “critique” principale reste celle des “Stochastic Parrots” de (Bender & Koller, 2020; Bender et al., 2021)

  - Kirschenbaum (2023):
    Le papier de Bender et al. (2021) “offers a disarmingly linear account of how language, communication, intention, and meaning work, one that would seem to sidestep decades of scholarship around these same issues in literary theory [...] the passage would be red meat for a graduate critical-theory seminar.”

  - Underwood (2023):
    “The beautiful irony of this situation [...] is that a generation of humanists trained on Foucault have now rallied around “On the Dangers of Stochastic Parrots” to oppose a theory of language that their own disciplines invented, just at the moment when computer scientists are reluctantly beginning to accept it.”

La naissance de la critique contemporaine



“Dans quelque coin reculé de l’univers ruisselant du scintillement d’innombrables systèmes solaires, il y eut un jour un astre sur lequel des animaux intelligents inventèrent le connaître. Ce fut la minute la plus orgueilleuse et la plus menteuse de l’« histoire universelle »…”

De la vérité et du mensonge au sens extra-moral
(Nietzsche, 1873)

La matrice argumentative de la critique

La connaissance dépend du langage

↓

La relation entre la langue et le monde est essentiellement arbitraire

↓

Toute régularité dans le language/la connaissance
n'est pas naturelle mais culturelle/sociale/politique

↓

Nous devons résister aux régularités existantes et en créer de nouvelles

Critique et Formalisme

- La matrice argumentative articulée par Nietzsche…
  - …représenta un renouvellement radical du projet critique du 19ème siècle dans la pensée occidentale

Critique et Formalisme

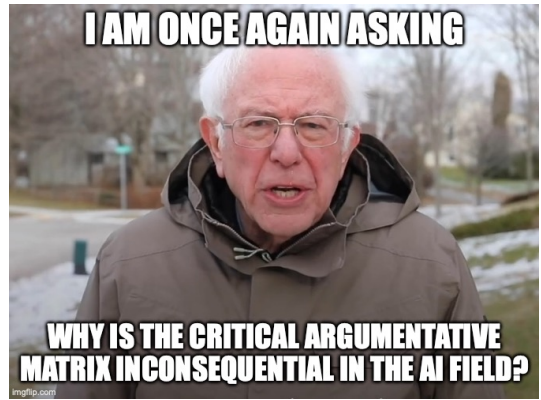
- La matrice argumentative articulée par Nietzsche…
  - …représenta un renouvellement radical du projet critique du 19ème siècle dans la pensée occidentale
  - …est devenue le modèle de la plupart des efforts critiques au 20ème siècle

Critique et Formalisme

- La matrice argumentative articulée par Nietzsche…
  - …représenta un renouvellement radical du projet critique du 19ème siècle dans la pensée occidentale
  - …est devenue le modèle de la plupart des efforts critiques au 20ème siècle
  - …semble contenir tout ce dont nous avons besoin pour élaborer une approche critique de l'IA et les LLMs au 21ème siècle

Critique et Formalisme

- La matrice argumentative articulée par Nietzsche…
  - …représenta un renouvellement radical du projet critique du 19ème siècle dans la pensée occidentale
  - …est devenue le modèle de la plupart des efforts critiques au 20ème siècle
  - …semble contenir tout ce dont nous avons besoin pour élaborer une approche critique de l'IA et les LLMs au 21ème siècle



La matrice argumentative de la critique

La connaissance dépend du langage

↓

La relation entre la langue et le monde est essentiellement arbitraire

↓

Toute régularité dans le language/la connaissance
n'est pas naturelle mais culturelle/sociale/politique

↓

Nous devons résister aux régularités existantes et en créer de nouvelles

La matrice argumentative de la critique

La connaissance dépend du langage
(Épistémologie)

↓

La relation entre la langue et le monde est essentiellement arbitraire

↓

Toute régularité dans le language/la connaissance
n'est pas naturelle mais culturelle/sociale/politique
(Politique)

↓

Nous devons résister aux régularités existantes et en créer de nouvelles
(Esthétique)

La matrice argumentative de la critique

La connaissance dépend du langage
(Épistémologie)

[La relation entre la langue et le monde est essentiellement arbitraire?]

Toute régularité dans le language/la connaissance
n'est pas naturelle mais culturelle/sociale/politique
(Politique)

↓

Nous devons résister aux régularités existantes et en créer de nouvelles
(Esthétique)

Critique et Formalisme

- A l'origine de cette situation se trouve le nouveau rôle fondationnel joué par les sciences formelles au 20ème siècle
  - Pour une théorie du langage: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...

Critique et Formalisme

- A l'origine de cette situation se trouve le nouveau rôle fondationnel joué par les sciences formelles au 20ème siècle
  - Pour une théorie du langage: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...
- La tradition critique s'est soit retirée des domaines conquis par les approches formelles, soit a fait des approches formelles la cible de la critique.

Critique et Formalisme

- A l'origine de cette situation se trouve le nouveau rôle fondationnel joué par les sciences formelles au 20ème siècle
  - Pour une théorie du langage: Carnap, Gödel, Turing, Shannon, Harris, Chomsky...
- La tradition critique s'est soit retirée des domaines conquis par les approches formelles, soit a fait des approches formelles la cible de la critique.
- Nous avons besoin d'une nouvelle stratégie: élaborer un formalisme critique

Pour un formalisme critique

- Dans le cas de l'IA, un formalisme critique peut fournir:
  - De nouveaux outils épistémologiques pour contrer les perspectives dogmatiques provenant de l'intérieur du domaine
  - De nouveaux outils théoriques contribuant à la production non dogmatique de connaissances positives

Plan

Introduction

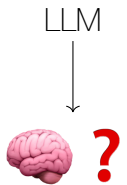
LLMs comme des objets formels

La structure des ‘embeddings’
- L’algèbre derrière les embeddings
- La structure derrière l’algèbre
- Les categories derrière la structure

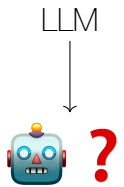
Conclusion

LLMs comme des fonctions calculables

LLMs comme des fonctions calculables



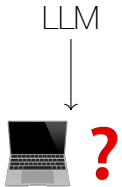
LLMs comme des fonctions calculables



LLMs comme des fonctions calculables

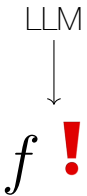
LLM

LLMs comme des fonctions calculables



LLMs comme des fonctions calculables

LLMs comme des fonctions calculables



LLMs comme des fonctions calculables

LLM

↓

Fonction

$$f\colon \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

LLMs comme des fonctions calculables

LLM

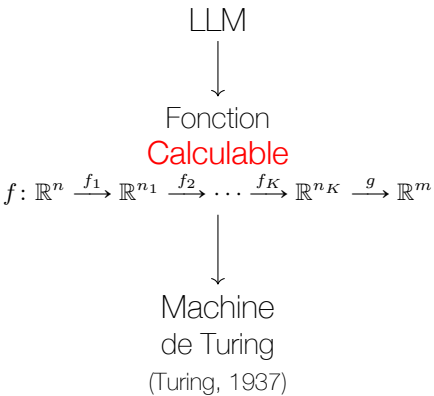
↓

Fonction

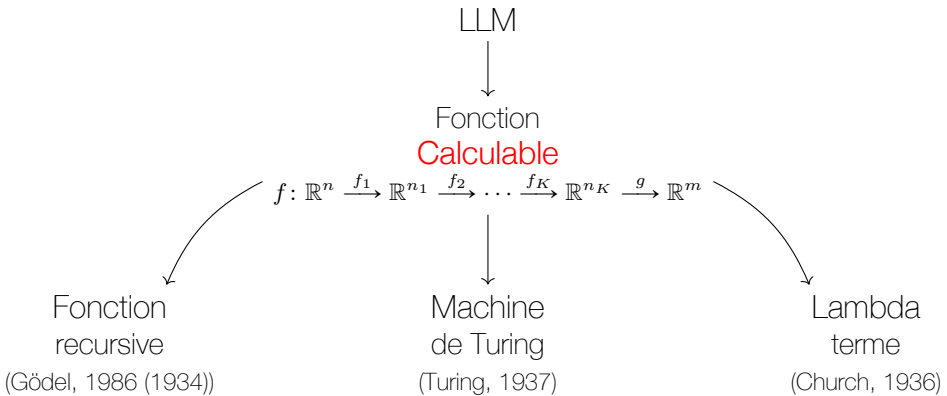
Calculable

$$f\colon \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \cdots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$$

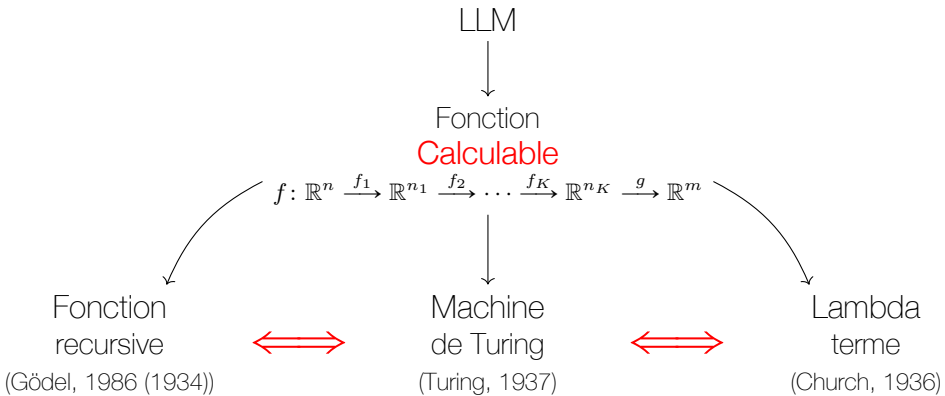
LLMs comme des fonctions calculables



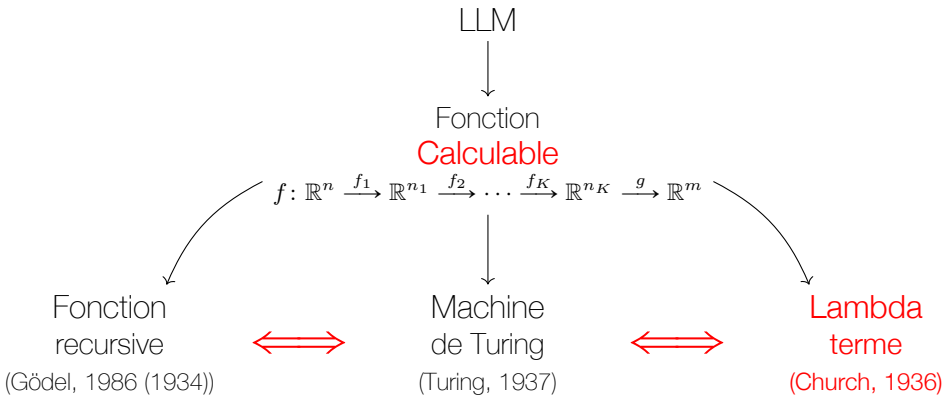
LLMs comme des fonctions calculables



LLMs comme des fonctions calculables



LLMs comme des fonctions calculables



LLMs comme des fonctions calculables

LLM

Fonction
Calculable

$f : \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^{n_1} \xrightarrow{f_2} \dots \xrightarrow{f_K} \mathbb{R}^{n_K} \xrightarrow{g} \mathbb{R}^m$

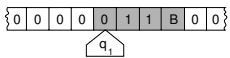
Fonction
recursive
(Gödel, 1986 (1934))

$\Longleftrightarrow$

Machine
de Turing
(Turing, 1937)

$\Longleftrightarrow$

Lambda
terme
(Church, 1936)



$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

credit: Nynexman4464

β-réduction dans le λ-calcul

$$yxz$$

β-réduction dans le λ-calcul

$$\lambda x.yxz$$

β-réduction dans le λ-calcul

$$(\lambda x.yxz)t$$

β-réduction dans le λ-calcul

$$(\lambda x.yxz)t$$

$$\downarrow$$

$$ytz$$

Évaluation empirique

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

Évaluation empirique

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

Évaluation empirique

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$

Évaluation empirique

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$

$\lambda f.\lambda x.f(f(f(f(fx))))$

Évaluation empirique

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$
$P' := \lambda r.\lambda s.\lambda f.\lambda x.f(f(f(f(fx))))$

0: $\lambda f.\lambda x.x$
1: $\lambda f.\lambda x.fx$
2: $\lambda f.\lambda x.f(fx)$
3: $\lambda f.\lambda x.f(f(fx))$
4: $\lambda f.\lambda x.f(f(f(fx)))$
5: $\lambda f.\lambda x.f(f(f(f(fx))))$
...
n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

$\lambda r.\lambda s.\lambda f.\lambda x.f(f(f(f(fx))))(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$

$\lambda f.\lambda x.f(f(f(f(fx))))$

Interprétabilité

$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}\, x)\ldots)$

$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$

$\vdots$

$\lambda f.\lambda x.f(f(f(f(fx))))$

$$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$$

0: $\lambda f.\lambda x.x$

1: $\lambda f.\lambda x.fx$

2: $\lambda f.\lambda x.f(fx)$

3: $\lambda f.\lambda x.f(f(fx))$

4: $\lambda f.\lambda x.f(f(f(fx)))$

5: $\lambda f.\lambda x.f(f(f(f(fx))))$

$\ldots$

n: $\lambda f.\lambda x.\underbrace{f(\ldots(f}_{n \text{ times}}x)\ldots)$

$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda f.\lambda x.f(fx))(\lambda f.\lambda x.f(f(fx)))$$
$$\lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)(\lambda g.\lambda y.g(gy))(\lambda h.\lambda z.h(h(hz)))$$
$$\lambda n.\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f(nfx)(\lambda h.\lambda z.h(h(hz)))$$
$$\lambda n.\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f(nfx)(\lambda h.\lambda z.h(h(hz)))$$
$$\lambda f.\lambda x.(\lambda g.\lambda y.g(gy))f((\lambda h.\lambda z.h(h(hz)))fx)$$
$$\lambda f.\lambda x.(\lambda y.f(fy))((\lambda h.\lambda z.h(h(hz)))fx)$$
$$\lambda f.\lambda x.(\lambda y.f(fy))((\lambda z.f(f(fz)))x)$$
$$\lambda f.\lambda x.(\lambda y.f(fy))(f(f(fx)))$$
$$\lambda f.\lambda x.f(f(f(f(fx))))$$

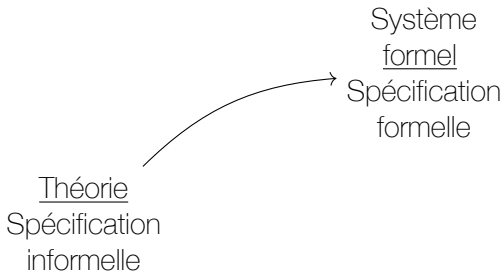
$P := \lambda m.\lambda n.\lambda f.\lambda x.mf(nfx)$

$P'' :=$ λRófÄÒêÑ5È|Àxñ=∞ù ÿmWf286ëy'SÒú>v&ìÂ ¬2 óÉ7öç∞{ã>2flB°µG#À9ÇU ∞btYBô¥ Ù ë‰3;5 å[l–èuô¨Ü ˚ 7–Ù.λ:ˆ4mÓØ¥´è—+ÌsÖ,$+gï„B™÷o–#ïŸê Ûv –gÓÿ/ëiijO‡Œfi • J1«€ø¸Ï hãêt‡æY$ˆ6 FïW»RÙKg¢¨.λ‡d¯… D2÷¸ò ˚ xêÈy. Ó”¢b Bé£NÈ1Ê‡/Û9Ñµ–/JYÇõË9ÿÀÈ.λÁÍ À^öÇ,»fq∞±îˇB5Ì>O˜g™“6Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨.λÆà€fUòfEÙ·Í m#„¸4\r√¬÷Îpò»y*vtÄJÃF1ûÁóz«ñM”DjŒ BËèÍT _ Êa‰ÁÇΩ @\Ø^˜]Î›h‡.:ˆ4mÓØ¥´è—+ÌsÖ,$+gï„B™÷o–#ïŸê Ûv–gÓÿ /ëiijO‡Œfi • J1«€ø¸Ï hãêt‡æY$ˆ6 FïW»RÙKg¢¨ÁÍ À^öÇ,»fq∞±îˇB5Ì>O˜g™“6 Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨(‡d¯… D2÷¸ò ˚ xêÈy. Ó”¢bBé£NÈ1Ê‡/Û9Ñµ–/ JYÇõË9ÿÀÈÁÍ À^öÇ,»fq∞±îˇB5Ì>O˜g™“6Ωe“æëC/ã… Ö · fÓ Å]ÑåyÊN°Ê ›¨Æ à€fUòfEÙ·Í m#„¸4\r√¬÷Îpò»y*vtÄJÃF1ûÁóz«ñM”DjŒBËèÍT _ Êa‰ÁÇΩ @\ Ø^˜]Î›h‡)(ËÎÛeí4WµÍ }w„$Ω“K5èÃ¶%3[m ’˙BÃfifÒ; óJçCËîõŸØcB,ñ$Ãå}ÓÅØ3¡ ˇ?ö¯oŒ@ fl8ˇR CÆoˇ* &<Ÿ ¬o12À‰ àÓÜ#ï ″,ú¨ «ô¸„∞Îâä“øÃd|ˆÑ’Ê yØ;ˆW ›″wò[}\»ÖÈúw’6<ùˇ=ãÕ ¯Ì‘Dz?2±|é‘ 3Â/rxµ∞µ$ÀëÂ*lf˜îû’+ÍVìy^{a}Gæßägô/,uÑ)

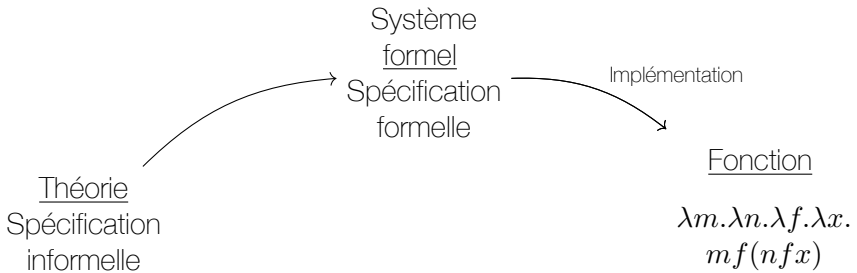
La structure implicite des données

Théorie
Spécification
informelle

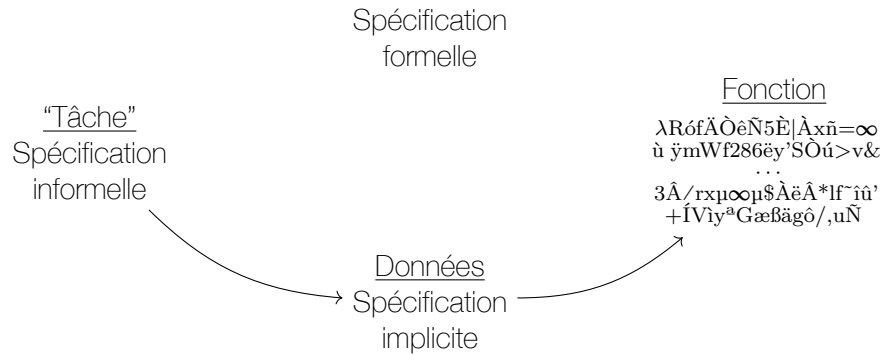
La structure implicite des données



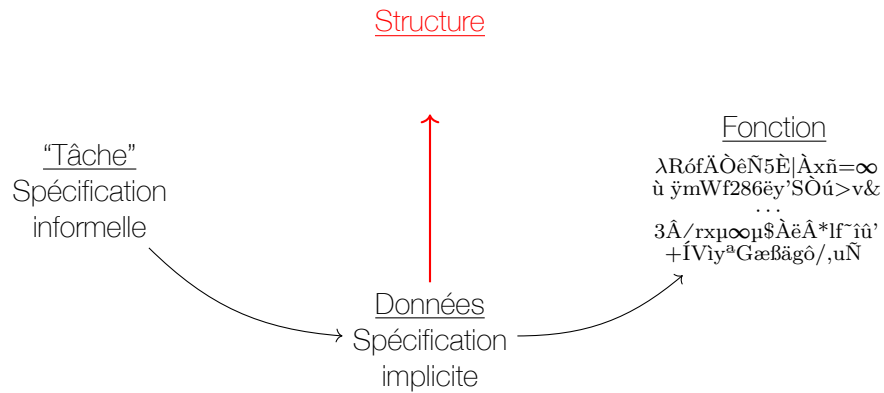
La structure implicite des données



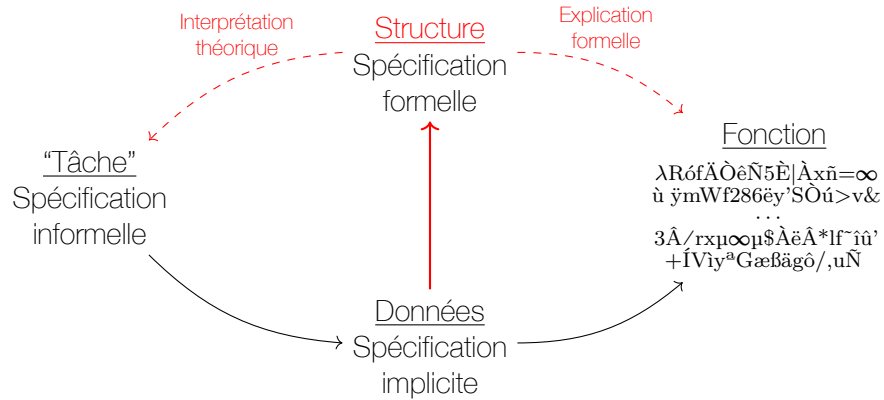
La structure implicite des données



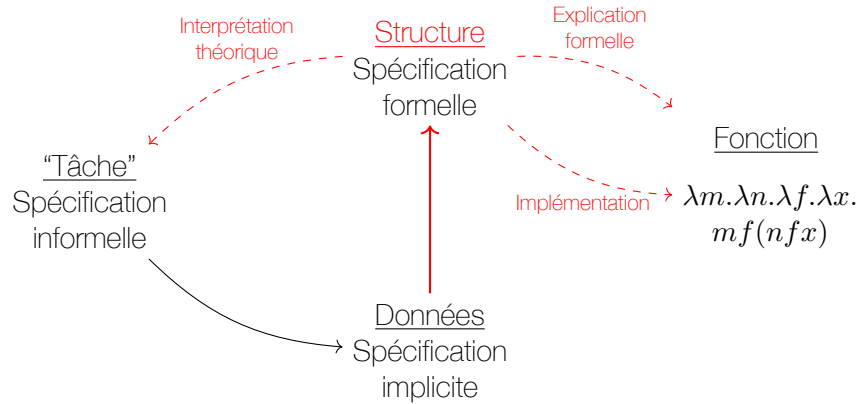
La structure implicite des données



La structure implicite des données



La structure implicite des données



Plan

Plan

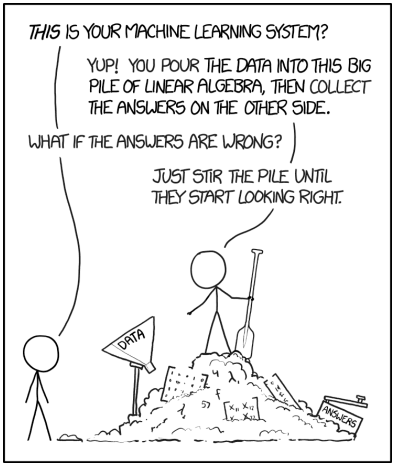
Introduction

LLMs comme des objets formels

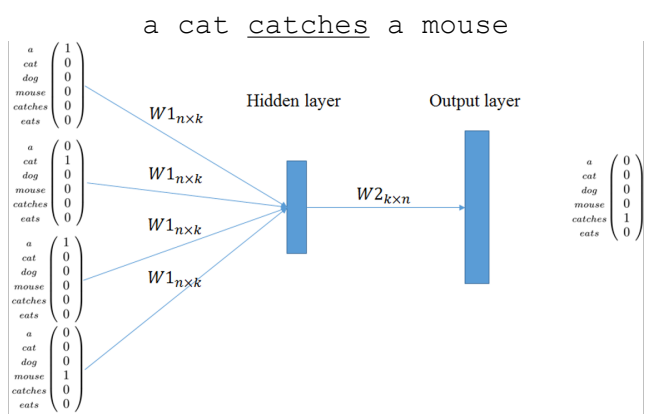
La structure des ‘embeddings’
- L’algèbre derrière les embeddings
- La structure derrière l’algèbre
- Les categories derrière la structure

Conclusion

Apprentissage machine dans l'espace de plongement

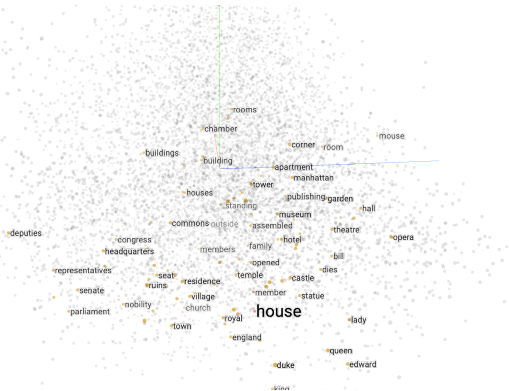


Credit: xkcd.com

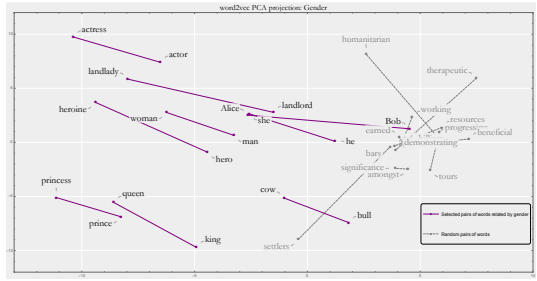


Credit: Ferrone et al., 2017

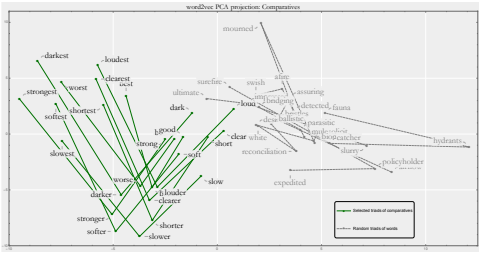
Espace des vecteurs de mot: similarité et analogie

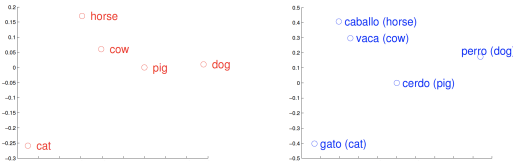


(https://projector.tensorflow.org)

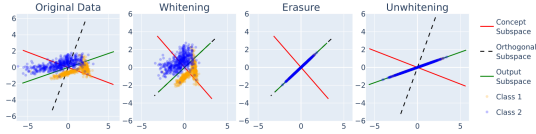


Espace des vecteurs de mot: autres applications

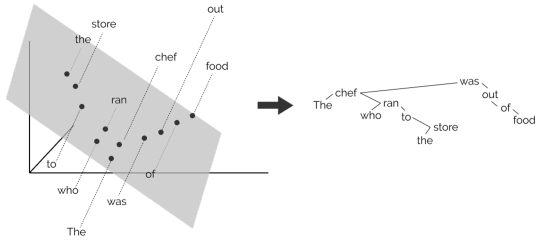




(Mikolov et al., 2013)



(Belrose et al., 2024)



(https://nlp.stanford.edu/~johnhew/structural-probe.html)

word2vec expliqué

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une factorisation implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

word2vec expliqué

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une factorisation implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

word2vec expliqué

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une factorisation implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

word2vec expliqué

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une factorisation implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

word2vec expliqué

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une factorisation implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

word2vec expliqué

(Levy and Goldberg, 2014)

$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une factorisation implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).

word2vec expliqué

(Levy and Goldberg, 2014)

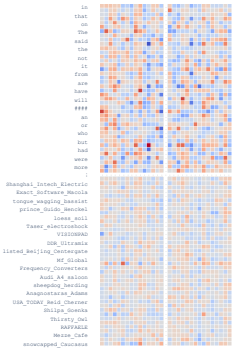
$$\ell = \sum_{w \in V_w} \sum_{c \in V_c} \#(w,c)(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D}[\log \sigma(-\vec{w} \cdot \vec{c}_N)])$$

$$\frac{\partial \ell}{\partial(\vec{w} \cdot \vec{c})} = 0 \quad \text{when} \quad \vec{w} \cdot \vec{c} \quad = \log \left(\frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)} \right) - \log k$$

- Word2vec effectue une factorisation implicite, de basse dimension, d'une matrice mot-contexte à information mutuelle ponctuelle (pmi).
- La Décomposition en Valeurs Singulières fournit une solution exacte à ce problème.

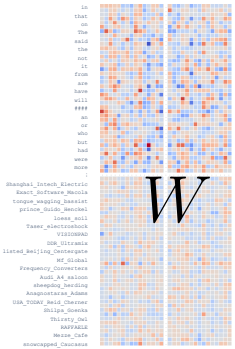
word2vec comme factorization implicite de matrice

(Levy and Goldberg, 2014)



word2vec comme factorization implicite de matrice

(Levy and Goldberg, 2014)



W

$C?$

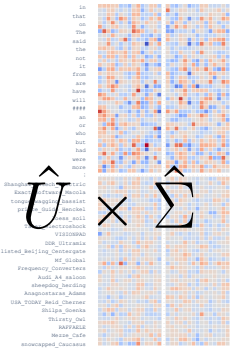
$M =$

$\text{pmi}(w, c) - \log k$

$W \times C \approx M$

Vecteurs de mot comme SVD

(Levy and Goldberg, 2014)



$$\hat{U} \times \hat{\Sigma}$$

$$\hat{V}^{\mathrm{T}}$$

$$M = \mathrm{pmi}(w, c) - \log k$$

$$M \approx \hat{U} \times \hat{\Sigma} \times \hat{V}^{\mathrm{T}}$$

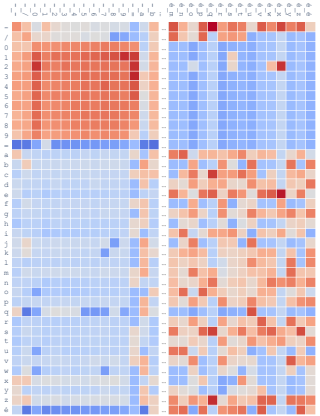
Exemple: Caractères dans Wikipédia

$$W = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$C = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

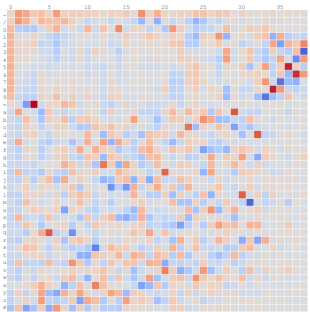
$$M_{wc} = \text{pmi}(w, c)$$

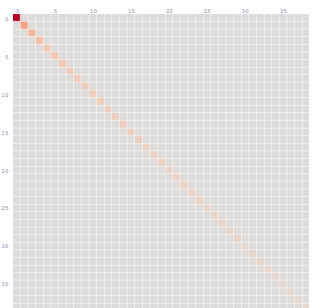
$$= \log \frac{p(w, c)}{p(w)p(c)}$$

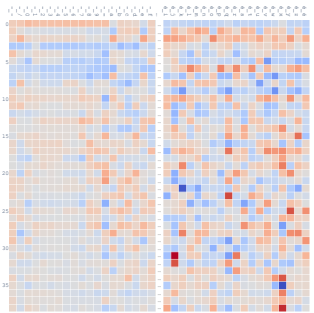


SVD d'une matrice pmi des caractères dans Wikipédia

$$U \qquad \Sigma \qquad V^T$$

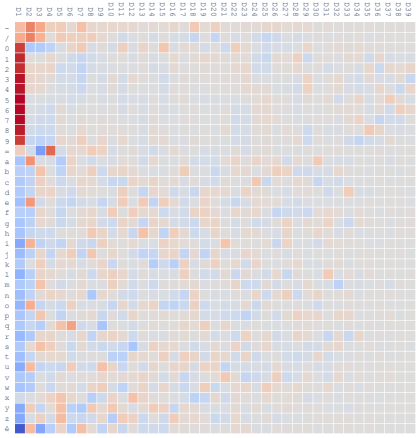






Tronquer

$$U \times \Sigma$$

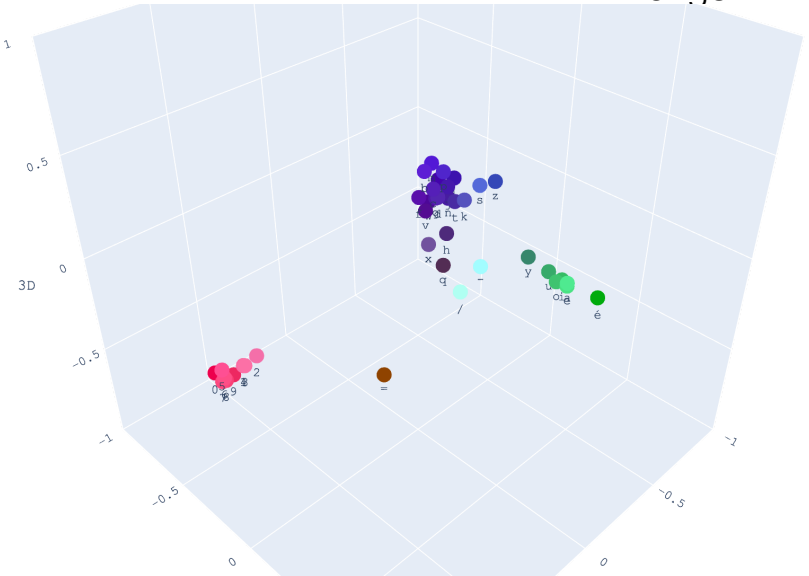


Tronquer

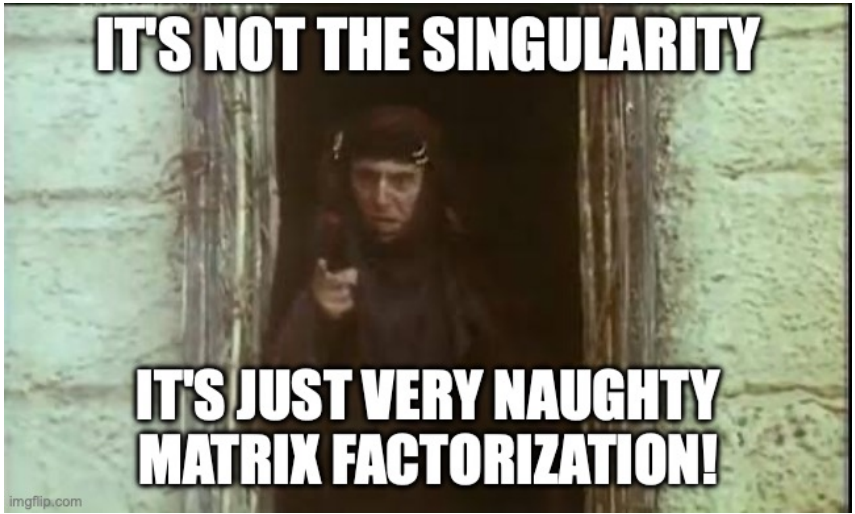
$$\hat{U} \times \hat{\Sigma}$$

Plonger

$$\hat{U} \times \hat{\Sigma}$$



Que conclure?



4 Why does this produce good word representations?

Good question. We don't really know.

The distributional hypothesis states that words in similar contexts have similar meanings. The objective above clearly tries to increase the quantity $v_w \cdot v_c$ for good word-context pairs, and decrease it for bad ones. Intuitively, this means that words that share many contexts will be similar to each other (note also that contexts sharing many words will also be similar to each other). This is, however, very hand-wavy.

Can we make this intuition more precise? We'd really like to see something more formal.

(Goldberg and Levy, 2014)

Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’
- L’algèbre derrière les embeddings
- La structure derrière l’algèbre
- Les categories derrière la structure

Conclusion

Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

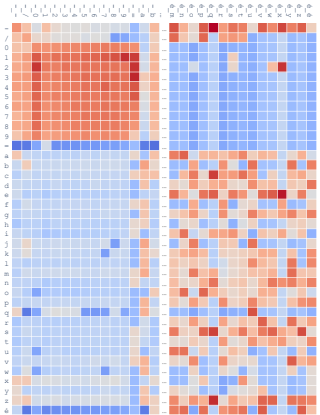
Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$

$$(x, y) \mapsto \mathrm{pmi}(x, y)$$



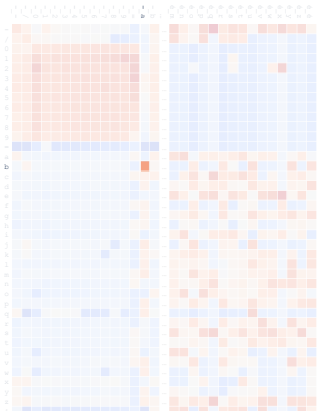
Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$

$$(x, y) \mapsto \mathrm{pmi}(x, y)$$



Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

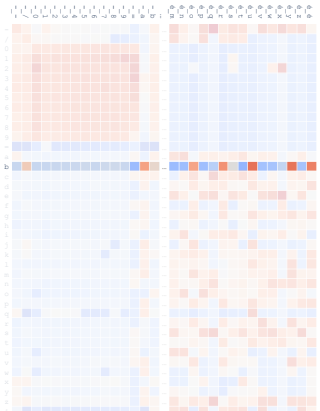
$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$

$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$

$$x \mapsto M(x, -)$$



Vecteurs de mots comme fonctions sur des ensembles

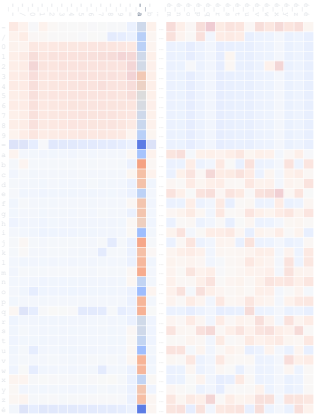
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



Vecteurs de mots comme fonctions sur des ensembles

$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$

$$X \xrightarrow{M_x} \mathbb{R}^Y$$

$$\mathbb{R}^X \xleftarrow[M_y]{} Y$$

Vecteurs de mots comme fonctions sur des ensembles

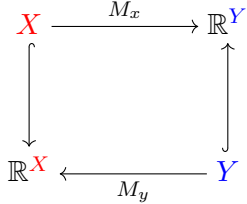
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



Vecteurs de mots comme fonctions sur des ensembles

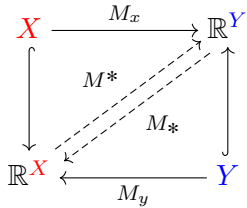
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\,\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\,\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

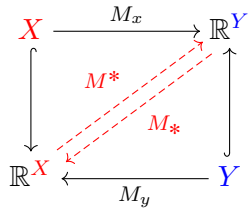
$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



$$M^* \colon \mathbb{R}^X \to \mathbb{R}^Y$$
$$M_* \colon \mathbb{R}^Y \to \mathbb{R}^X$$

Vecteurs de mots comme fonctions sur des ensembles

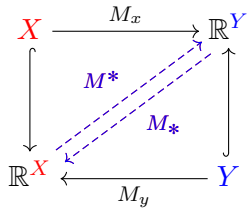
$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$



Vecteurs de mots comme fonctions sur des ensembles

$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$

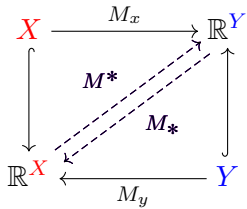
$M^* M_* : \mathbb{R}^Y \to \mathbb{R}^Y$



Vecteurs de mots comme fonctions sur des ensembles

$$M_* M^* \colon \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* \colon \mathbb{R}^Y \to \mathbb{R}^Y$$

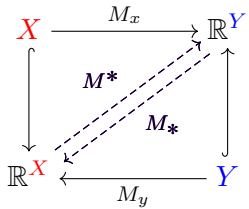
$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$



Vecteurs de mots comme fonctions sur des ensembles

$$M_* M^* \colon \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* \colon \mathbb{R}^Y \to \mathbb{R}^Y$$

$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$



$$M = U \Sigma V^T \qquad U := [u_1, \ldots, u_m] \qquad V := [v_1, \ldots, v_n]$$
$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

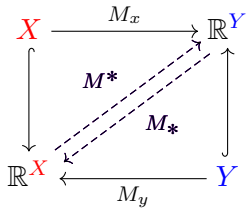
Vecteurs de mots comme fonctions sur des ensembles

$$M_* M^* : \mathbb{R}^X \to \mathbb{R}^X$$
$$M^* M_* : \mathbb{R}^Y \to \mathbb{R}^Y$$

$$\{u_1, \ldots, u_m\} \subset \mathbb{R}^X$$
$$\{v_1, \ldots, v_n\} \subset \mathbb{R}^Y$$
$$\{\lambda_1, \ldots, \lambda_{\min(m,n)}, 0, \ldots, 0\}$$

$$M_* M^* u_i = \lambda_i u_i$$
$$M^* M_* v_i = \lambda_i v_i$$

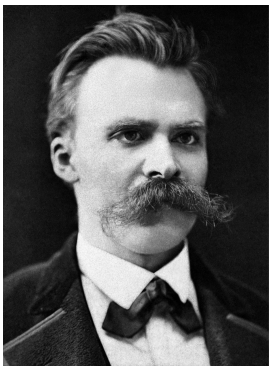
Les u_i and v_i sont des points fixes (linéaires)!



$$M = U \Sigma V^T$$

$$U := [u_1, \ldots, u_m]$$
$$V := [v_1, \ldots, v_n]$$
$$\Sigma := \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_r} \end{bmatrix}$$

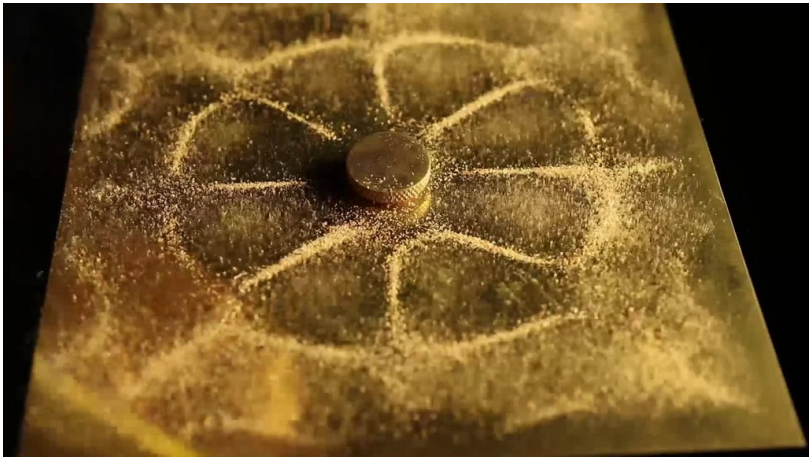
Figures de Chladni



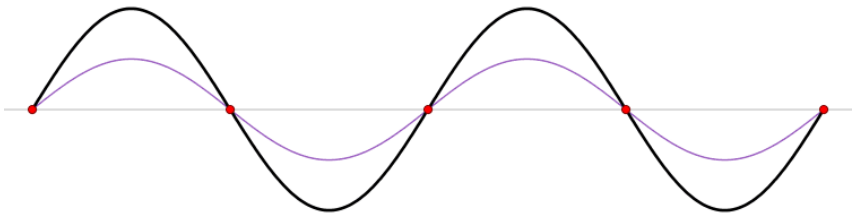
“On peut imaginer un homme qui soit totalement sourd, et n’ait jamais ressenti une sensation sonore et musicale: tout comme cet homme considérera avec stupéfaction les figures acoustiques de Chladni dans le sable, trouvera leur cause dans la vibration de la corde et jurera ensuite qu’il sait nécessairement à présent ce que les hommes appellent le son, c’est ainsi qu’il en va pour nous tous avec le langage.”

(Nietzsche, 1873)

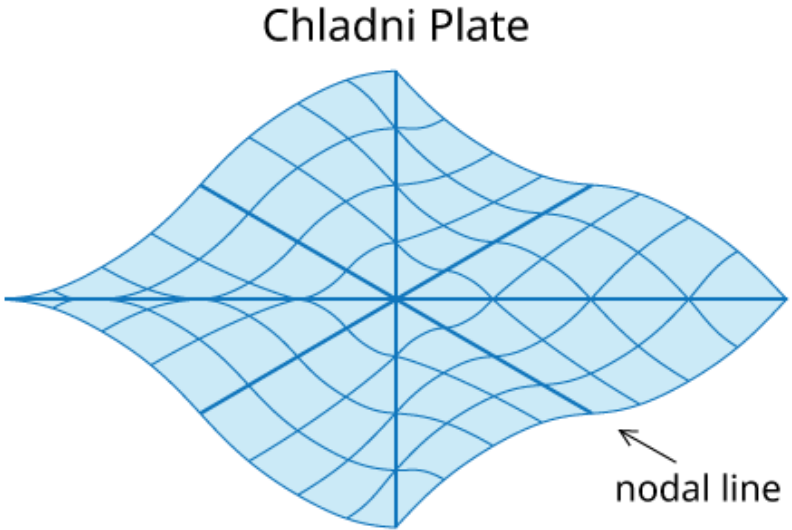
Figures de Chladni



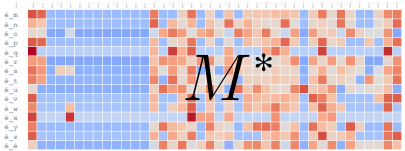
Figures de Chladni

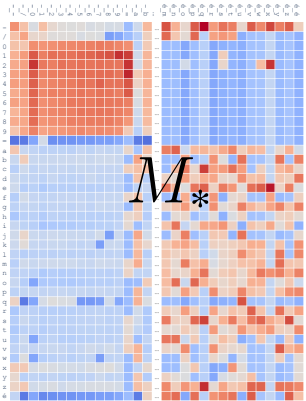


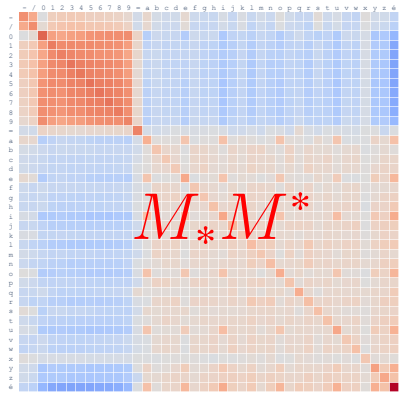
Figures de Chladni



M_*M^* comme matrice de covariance

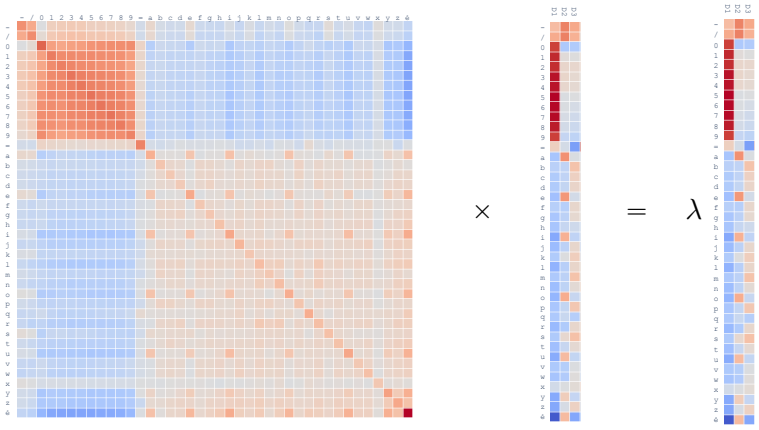






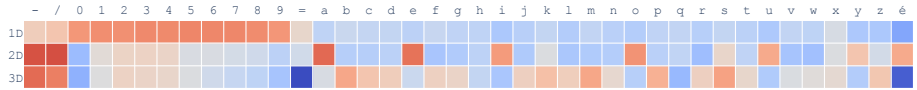
Vecteurs propres comme points fixes

$$M_* M^* u = \lambda u$$

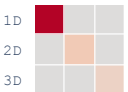


Traits structuraux

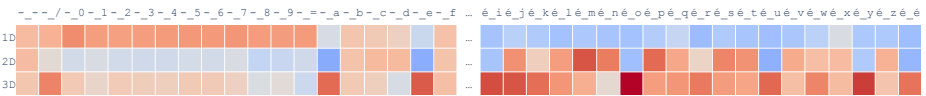
Eigenvectors of $M_* M^*$:



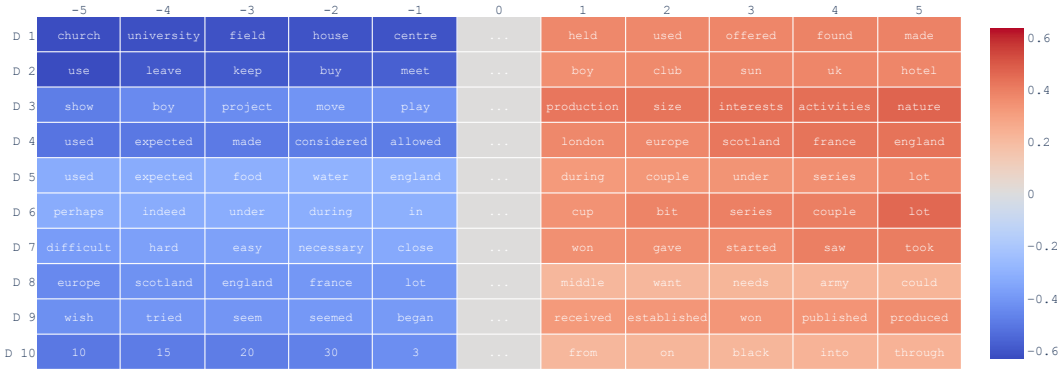
Eigenvalues of $M_* M^*$ and $M^* M_*$:



Eigenvectors of $M^* M_*$:



Mots


−5 −4 −3 −2 −1 0 1 2 3 4 5
D 1 church university field house centre ... held used offered found made
D 2 use leave keep buy meet ... boy club sun uk hotel
D 3 show boy project move play ... production size interests activities nature
D 4 used expected made considered allowed ... london europe scotland france england
D 5 used expected food water england ... during couple under series lot
D 6 perhaps indeed under during in ... cup bit series couple lot
D 7 difficult hard easy necessary close ... won gave started saw took
D 8 europe scotland england france lot ... middle want needs army could
D 9 wish tried seem seemed began ... received established won published produced
D 10 10 15 20 30 3 ... from on black into through
0.6
0.4
0.2
0
−0.2
−0.4
−0.6

Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’
- L’algèbre derrière les embeddings
- La structure derrière l’algèbre
- Les categories derrière la structure

Conclusion

Vecteurs de mots comme foncteurs sur des catégories

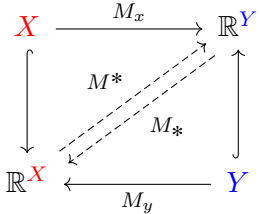
$$X = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$$

$$Y = X \times X = \{\texttt{(-,-)}, \texttt{(-,/)}, \texttt{(-,0)}, \ldots, \texttt{(é,z)}, \texttt{(é,é)}\}$$

$$M \colon X \times Y \to \mathbb{R}$$
$$(x, y) \mapsto \mathrm{pmi}(x, y)$$

$$M_x \colon X \to \mathbb{R}^Y$$
$$x \mapsto M(x, -)$$

$$M_y \colon Y \to \mathbb{R}^X$$
$$y \mapsto M(-, y)$$



$$M^* \colon \mathbb{R}^X \to \mathbb{R}^Y$$
$$M_* \colon \mathbb{R}^Y \to \mathbb{R}^X$$

Vecteurs de mots comme foncteurs sur des catégories

$\mathsf{C} = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$

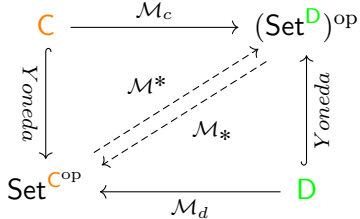
$\mathsf{D} = \mathsf{C} = \{\texttt{-}, \texttt{/}, \texttt{0}, \texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}, \texttt{6}, \texttt{7}, \texttt{8}, \texttt{9}, \texttt{=}, \texttt{a}, \texttt{b}, \texttt{c}, \ldots, \texttt{w}, \texttt{x}, \texttt{y}, \texttt{z}, \texttt{é}\}$

Profunctor

$$\mathcal{M} \colon \mathsf{C}^{\mathrm{op}} \times \mathsf{D} \to \mathsf{Set}$$
$$(c, d) \mapsto \mathcal{M}(c, d)$$

$$\mathcal{M}_c \colon \mathsf{C} \to \left(\mathsf{Set}^{\mathsf{D}}\right)^{\mathrm{op}}$$
$$c \mapsto \mathcal{M}(c, -)$$

$$\mathcal{M}_d \colon \mathsf{D} \to \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}}$$
$$d \mapsto \mathcal{M}(-, d)$$



$$\mathcal{M}^* \colon \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \to (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}}$$
$$\mathcal{M}_* \colon (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \to \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}}$$

Vecteurs de mots comme foncteurs sur des catégories

Adjonction d'Isbell

$$\mathcal{M}^* \colon \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \colon \mathcal{M}_*$$

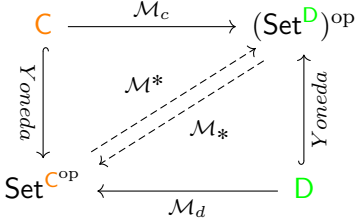
$$\mathcal{M}_*\mathcal{M}^* \colon \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} \to \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}}$$
$$\mathcal{M}^*\mathcal{M}_* \colon (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} \to (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}}$$

$$\mathrm{Fix}(\mathcal{M}_*\mathcal{M}^*) := \{f \in \mathsf{Set}^{\mathsf{C}^{\mathrm{op}}} | \mathcal{M}_*\mathcal{M}^*(f) \cong f\}$$
$$\mathrm{Fix}(\mathcal{M}^*\mathcal{M}_*) := \{g \in (\mathsf{Set}^{\mathsf{D}})^{\mathrm{op}} | \mathcal{M}^*\mathcal{M}_*(g) \cong g\}$$

Nucleus of $\mathcal{M} = \{(f_i, g_i)\}$, such that:

$\mathcal{M}^* f_i \cong g_i$ and $\mathcal{M}_* g_i \cong f_i$

Le noyeau (nucleus) est une catégorie complète et cocomplète



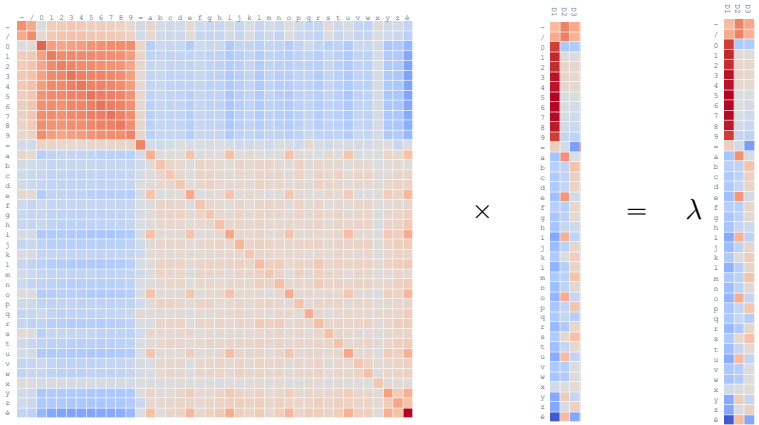
Les catégories C et D peuvent être enrichies!

E.g.:

$$\mathcal{M}^* \colon 2^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (2^{\mathsf{D}})^{\mathrm{op}} \colon \mathcal{M}_*$$
$$\mathcal{M}^* \colon \bar{\mathbb{R}}^{\mathsf{C}^{\mathrm{op}}} \leftrightarrows (\bar{\mathbb{R}}^{\mathsf{D}})^{\mathrm{op}} \colon \mathcal{M}_*$$

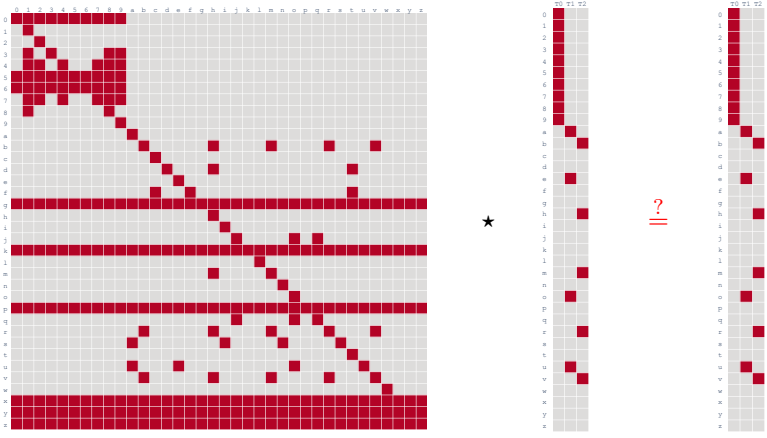
Points fixes booléens

$$M_* M^* u = \lambda u$$



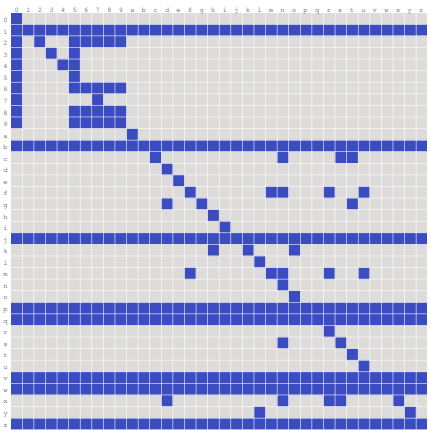
Points fixes booléens

$$\mathcal{M}_* \mathcal{M}^* f = f$$



Points fixes booléens

$$M_i^* M_*^i d = d$$

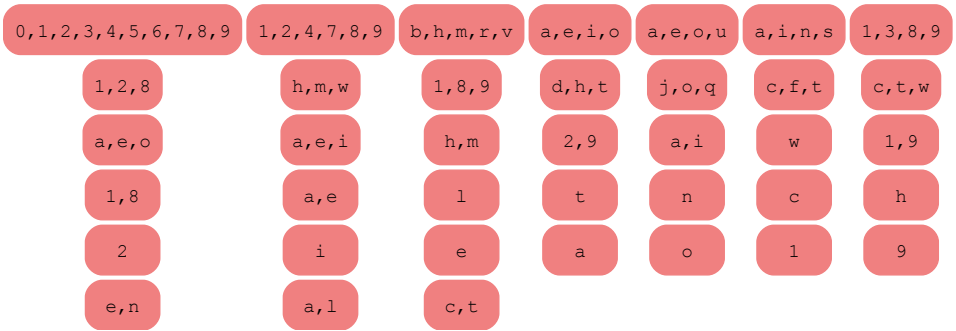


$\star$

$\stackrel{?}{=}$

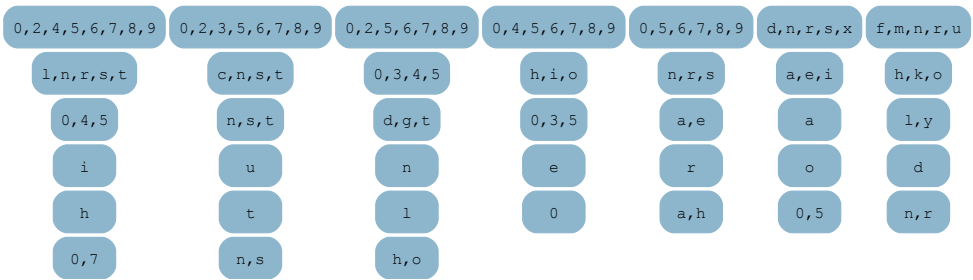
“Eigensets”

$$\mathcal{M}_*\mathcal{M}^* f = f$$

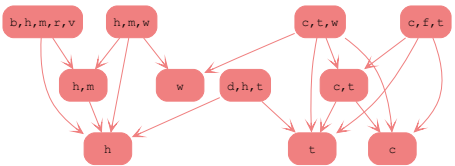


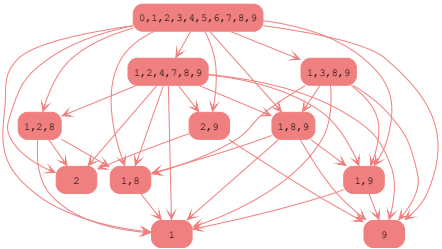
“Eigensets”

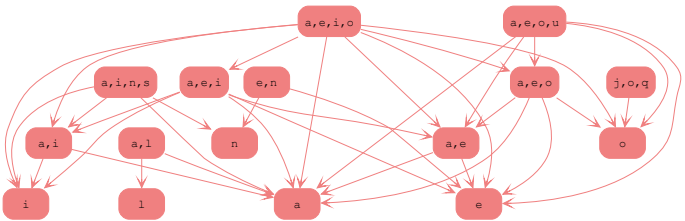
$$M_i^* M_*^i d = d$$



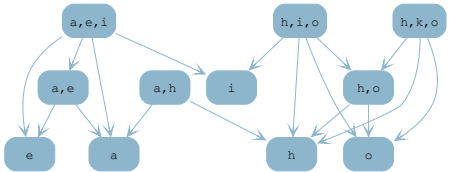
Quelle Structure?

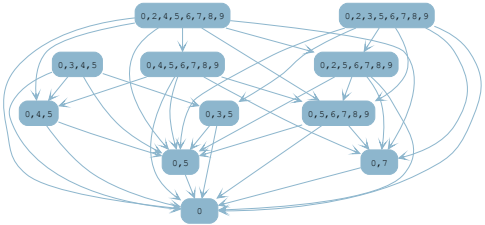


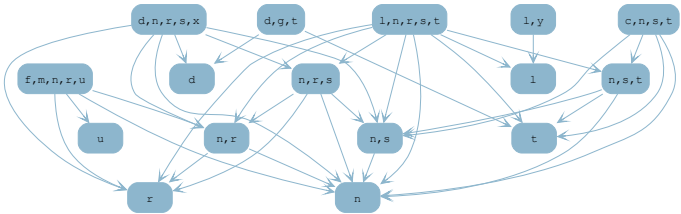




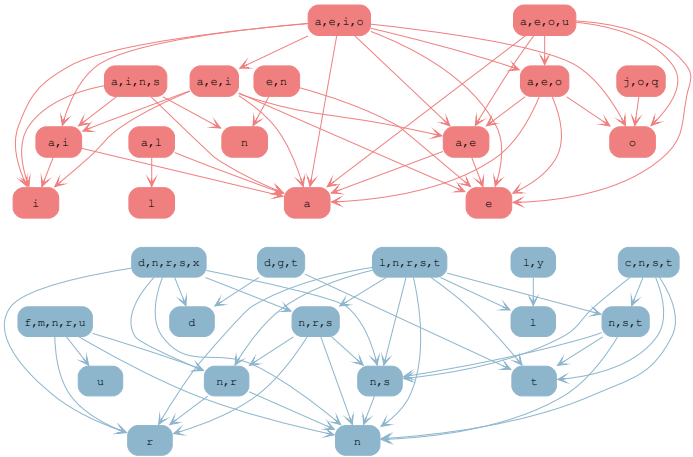
Quelle Structure?



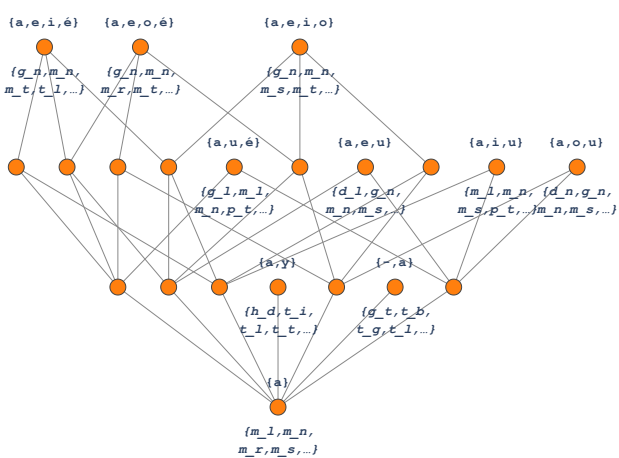


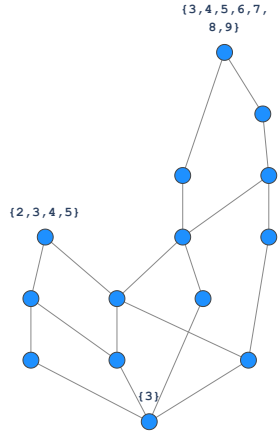


Quelle Structure?

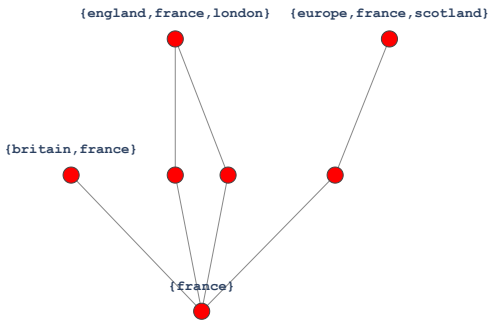


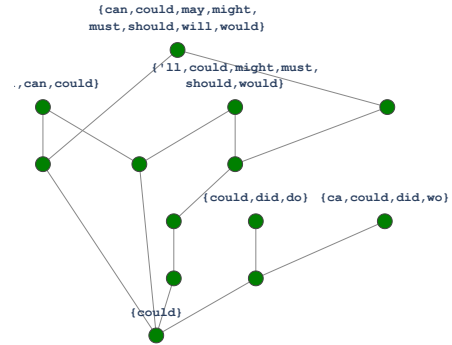
Concepts formels



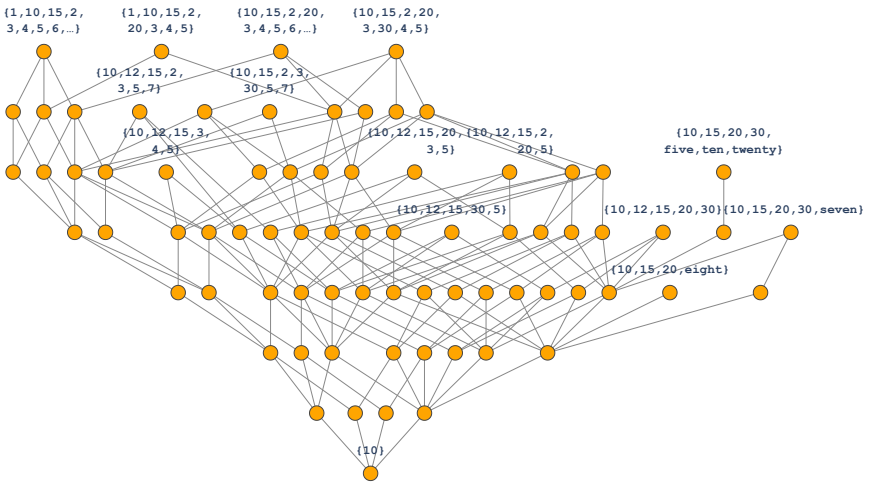


Concepts formels (mots)





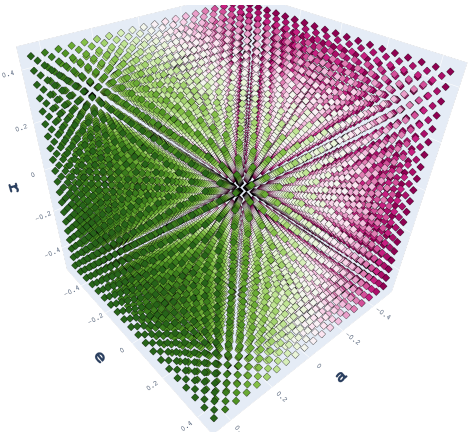
Concepts formels (mots)

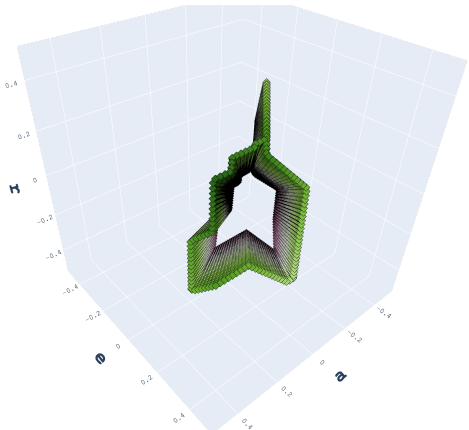


Chladni

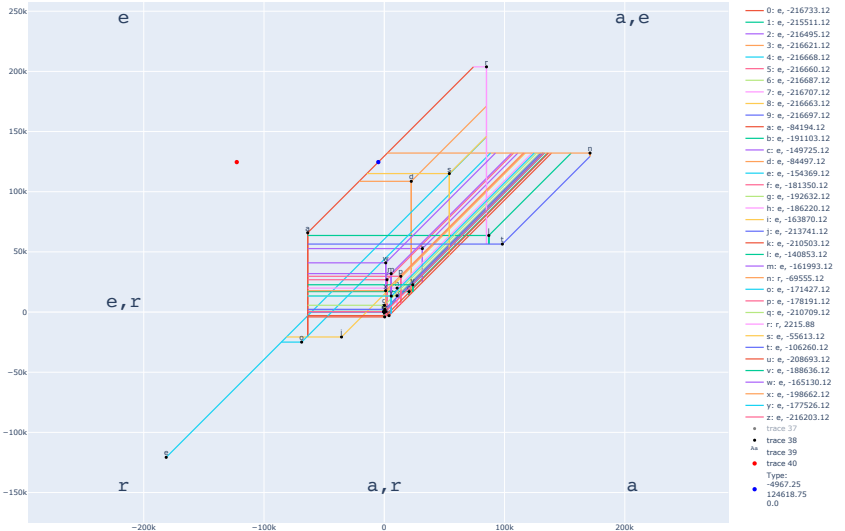
Noyeau (nucleus)

$$\bar{\mathbb{R}}\{\mathtt{a,e,r}\} \xrightarrow{\mathcal{M}_*\mathcal{M}^*} \bar{\mathbb{R}}\{\mathtt{a,e,r}\}$$





Structure interne du noyeau



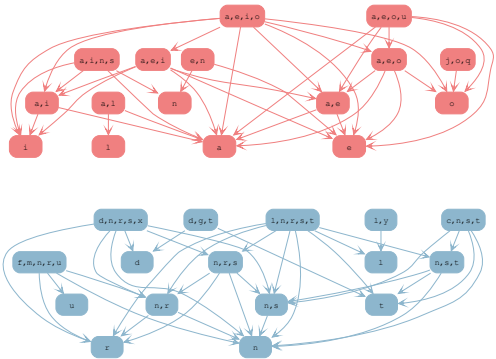
Théorie des types computationnels

Definition (Polaire/Orthogonal - Girard, 2006)

[É]tant donnée une fonction binaire $a, b \rightsquigarrow \langle a|b\rangle \colon A \times B \to C$ et un sous-ensemble $P \subset C$ (le « pôle »), on peut définir le *polaire* $X^\perp \subset B$ d'un sous-ensemble $X \subset A$ (resp. $Y^\perp \subset A$ d'un sous-ensemble $Y \subset B$) par :

$$X^\perp := \{y \in B \colon \forall x \in X, \langle a|b\rangle \in P\}$$
$$Y^\perp := \{x \in A \colon \forall y \in Y, \langle a|b\rangle \in P\}$$

- L'application « polaire » est décroissante: $X \subset X' \Rightarrow X'^\perp \subset X^\perp$.
- L'ensemble $\mathrm{Pol}(A) \subset \mathcal{P}(A)$ des ensembles *polaires*, i.e., de la forme $Y^\perp$, est stable par intersections arbitraires. En particulier, A est polaire et $X^{\perp\perp}$ est le plus petit ensemble polaire contenant X.
- En conséquence, $X^{\perp\perp\perp} = X^\perp$.



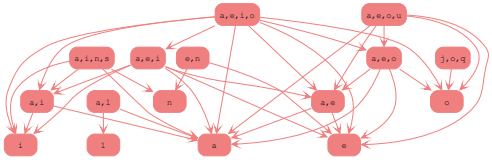
Théorie des types computationnels

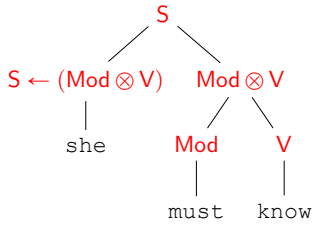
Definition (Polaire/Orthogonal - Girard, 2006)

[É]tant donnée une fonction binaire $a, b \rightsquigarrow \langle a|b\rangle \colon A \times B \to C$ et un sous-ensemble $P \subset C$ (le « pôle »), on peut définir le *polaire* $X^\perp \subset B$ d'un sous-ensemble $X \subset A$ (resp. $Y^\perp \subset A$ d'un sous-ensemble $Y \subset B$) par :

$$X^\perp := \{y \in B \colon \forall x \in X, \langle a|b\rangle \in P\}$$
$$Y^\perp := \{x \in A \colon \forall y \in Y, \langle a|b\rangle \in P\}$$

- L'application « polaire » est décroissante: $X \subset X' \Rightarrow X'^\perp \subset X^\perp$.
- L'ensemble $\mathrm{Pol}(A) \subset \mathcal{P}(A)$ des ensembles *polaires*, i.e., de la forme $Y^\perp$, est stable par intersections arbitraires. En particulier, A est polaire et $X^{\perp\perp}$ est le plus petit ensemble polaire contenant X.
- En conséquence, $X^{\perp\perp\perp} = X^\perp$.





(Gastaldi and Pellissier, 2021)

Plan

Introduction

LLMs comme des objets formels

La structure des ‘embeddings’
- L’algèbre derrière les embeddings
- La structure derrière l’algèbre
- Les categories derrière la structure

Conclusion

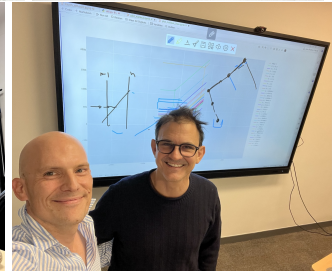
Conclusion: Pour un formalisme critique

- Il est urgent d'aborder la dimension épistémologique du projet critique pour pouvoir aborder l'IA critiquement

- Cela nécessite de développer une approche critique au sein des sciences formelles où la formalisation n'est pas supposée conduire à une naturalisation.
  - Le nouveau rôle des données au sein des sciences formelles est crucial en ce sens

- Un formalisme critique sera incomplet s'il reste déconnecté de la dimension politique (voire esthétique) associée aux données.
  - Nous avons besoin d'une nouvelle alliance entre les sciences formelles et les sciences humaines et sociales.

Collaborations



J. Terilla (CUNY), T.-D. Bradley (SandboxAQ), L. Pellissier (Paris-Est Créteil), Th. Seiller (CNRS), S. Jarvis (CUNY)

Papiers de référence

- Gastaldi, J. L. (2021).Why Can Computers Understand Natural Language? *Philosophy & Technology*, *34*(1), 149–214. https://doi.org/10.1007/s13347-020-00393-9

- Gastaldi, J. L., & Pellissier, L. (2021).The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*. https://doi.org/10.1080/03080188.2021.1890484

- Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024).The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*. https://api.semanticscholar.org/CorpusID:263613625

Références I

Ali, S. M., Dick, S., Dillon, S., Jones, M. L., Penn, J., & Staley, R. (2023).Histories of artificial intelligence: A genealogy of power. *BJHS Themes*, *8*, 1–18. https://doi.org/10.1017/bjt.2023.15

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2024).Leace: Perfect linear concept erasure in closed form. *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021).On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020).Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bradley, T.-D., Gastaldi, J. L., & Terilla, J. (2024).The structure of meaning in language: Parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*. https://api.semanticscholar.org/CorpusID:263613625

Church, A. (1936).An unsolvable problem of elementary number theory. *American Journal of Mathematics*, *58*(2), 345–363.

Gastaldi, J. L. (2021).Why Can Computers Understand Natural Language? *Philosophy & Technology*, *34*(1), 149–214. https://doi.org/10.1007/s13347-020-00393-9

Gastaldi, J. L., & Pellissier, L. (2021).The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*. https://doi.org/10.1080/03080188.2021.1890484

Références II

Girard, J.-Y. (2006). *Le point aveugle: Cours de logique. vers la perfection*. Editions Hermann.

Gödel, K. (1986 (1934)). On undecidable propositions of formal mathematical systems. In *Collected works* (pp. 346–371). Clarendon Press Oxford University Press.

Goldberg, Y., & Levy, O. (2014).Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, *abs/1402.3722*.

Kirschenbaum, M. (2023). *Again theory: A forum on language, meaning, and intent in the time of stochastic parrot*. https://critinq.wordpress.com/2023/06/26/again-theory-a-forum-on-language-meaning-and-intent-in-the-time-of-stochastic-parrots/

Levy, O., & Goldberg, Y. (2014).Neural word embedding as implicit matrix factorization. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2177–2185.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Le, Q., & Strohmann, T. (2013). *Learning representations of text using neural networks. NIPS deep learning workshop 2013 slides*.

Nietzsche, F. (1873). *De la vérité et du mensonge au sens extra-moral* (P. Wotling, Trans.). Flammarion.

Turing, A. (1937).On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230–265. https://doi.org/10.1112/plms/s2-42.1.230

Underwood, T. (2023, October 15). *The empirical triumph of theory* [Accessed: 2023-10-15]. https://critinq.wordpress.com/2023/06/29/the-empirical-triumph-of-theory/

Séminaire HiPhiS

Univ. de Montpellier, Univ. Paul Valéry, IRES, CNRS

Montpellier, France

Épistémologie de l'apprentissage machine

Pour un formalisme critique

Juan Luis Gastaldi

www.jlgastaldi.com

ETH zürich

15 Avril, 2025