

Que peuvent les algorithmes de TAL /
Big Data pour l'analyse sociologique des
textes ?

Analyser les discours et caractériser les
locuteurs des plateformes « Grand Débat
National » et « Vrai Débat »



Mr. Smith Goes to Washington – Columbia pictures- 1939

Mathieu Brugidou, Philippe Suignard

5 janvier 2023
Séminaire Social computing
Université de Pau



Un programme de recherche pour réunir les compétences SHS et Data Sciences au service de l'analyse de l'opinion publique sur le web



Mr. Smith Goes to Washington – Columbia pictures- 1939



Mr. Q. Goes to Washington – Capitole- 2021

Un double choc pour les sciences sociales : l'émergence des approches IA et du web. L'exemple de la sociologie de l'opinion publique

La numérisation de l'espace public transforme l'expression et l'analyse de l'opinion publique

Celle-ci se développe à la fois dans un espace public « vertical » (organisé par un système où les acteurs institutionnels produisent des discours et des arguments, des médias qui filtrent leurs expressions et des sondages qui saisissent l'opinion) et dans un espace public « horizontal », numérique où l'accès à l'expression n'est pas filtré *a priori*. (Cf. Cardon, 2010, Benvegnu 2011, Mabi 2014, Kotras, 2018)

Les deux circuits, celui de l'information et de la conversation, déconnectés par Tarde, se superposent dans le numérique avec deux conséquences : avènement d'un régime d'opinion « public-discursif » (Kotras, 2018) et dérégulation du marché informationnel (Cardon, 2018, *in* démocratie à l'heure de la post-vérité, Collège de France)

Le suivi de deux espaces publics est découplé : on ne sait pas corrélérer les enquêtes par sondage et les outils de suivi des expressions sur le web (Boullier, 2015 vs Boyadjian 2016), ni un espace des statuts et un espace des opinions.

Les approches IA questionnent les paradigmes des sciences sociales (Burrows et Savage, 2007, 2009, 2014) : des modèles (sociologiques) pour comprendre (échantillon et espace public) des comportements/représentations ou des algorithmes pour tracer et prévoir des comportements ? (Cf. Bastin, 2019, Cointet et Parasie 2019, Beaudouin, 2019)

A EDF, de nombreux enjeux d'opinion publique liés au numérique et des compétences en data science et en sciences sociales au sein du département SEQUOIA et pourraient être mobilisées pour analyser les réseaux sociaux et les forums pour peu que l'on sache élaborer les bons modèles sociologiques d'analyse de l'opinion publique sur le WEB.

Les sciences sociales à l'épreuve de l'IA et inversement : lever des verrous scientifiques pour l'analyse sociologique du web à EDF R&D



- Constituer des corpus web documentés pour l'analyse sociologique (opinion, controverse, suivi d'acteurs).
- Reconstituer des données manquantes : identifier les propriétés sociales des locuteurs.
- Avoir des outils et des méthodes pour l'analyse sémantique du web orientée opinion et analyse de controverse.

Exemple de l'étude Grand Débat National et Vrai débat.

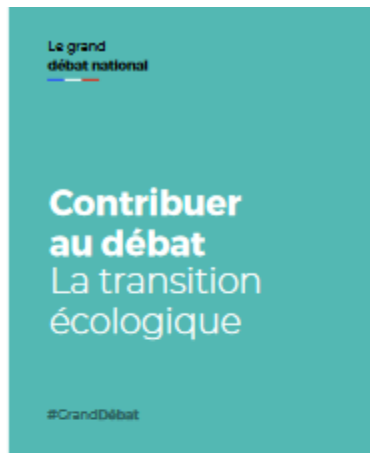
Hypothèses

- ▶ Sociologies politiques très différentes des participants aux deux débats (GDN et VD), cf. enquêtes Cevipof et Pacte.
- ▶ Structuration des discours par le design différent des plateformes
- ▶ Dans quelles mesures peut-on identifier un discours mais aussi public (spéculaire) au cœur du Grand Débat National ?
 - Dans quelles mesures, et selon quelles stratégies d'analyse mais aussi quelles « épistémologies embarquées » (Demazière et al., 2006 : 178), ces méthodes ADT et IA s'ajustent ou non aux formats d'énonciation (questions ouvertes vs échange d'arguments) et aux formats démocratiques (participatif ou délibératif) proposés par ces plateformes ?
 - Quelles solutions méthodologiques permettent de qualifier les propriétés sociales des locuteurs sur lesquelles on n'a que peu d'information directe ?
 - Quelles questions de recherche au croisement de la sociologie politique de l'opinion publique mais aussi des data science cette confrontation ouvre ?

Plan de la présentation

- ▶ Présentation des différents corpus utilisés et méthodes
- ▶ Retrouver un discours Gilets Jaunes dans le Grand débat national ?
Analyser les discours : ADT vs IA
- ▶ Retrouver un public Gilets Jaunes dans le Grand débat National ?
Retrouver les propriétés sociales et politiques des locuteurs à partir des textes, une approche par apprentissage
- ▶ Conclusions et perspectives

Design des plateformes



01. Quel est aujourd'hui pour vous le problème concret le plus important dans le domaine de l'environnement ? (1 seule réponse possible)
- ☐ La pollution de l'air ☐ Les dérèglements climatiques (crues, sécheresses)
☐ L'érosion du littoral ☐ La biodiversité et la disparition de certaines espèces
☐ Autres, précisez :
02. Que faudrait-il faire selon vous pour apporter des réponses à ce problème ?
03. Dites-vous que votre vie quotidienne est aujourd'hui touchée par le changement climatique ?
- ☐ Oui ☐ Non
- Si oui, de quelle manière votre vie quotidienne est-elle touchée par le changement climatique ?

VRAI DEBAT PARTICIPER ! Actualités Processus Sur le terrain ▾ Aidons-Nous ! ▾ À propos ▾ Inscription Connexion

Retour Proposition
Transition Ecologique et Solidaire, Agriculture & Alimentation, Transport

Juliette F - 30 janv.
Non à la voiture électrique !
704 votes · 94 arguments · 5 sources

^ Bénéfices apportés

La voiture électrique n'est pas une solution à la voiture thermique. Prenez 5 minutes pour voir la vidéo dans les sources.
Il existe sûrement d'autres alternatives, dont l'une est le vélo mobile.

Je propose donc un arrêt des subventions pour la voiture électrique et un investissement pour favoriser le développement du vélo mobile et la recherche pour d'autres modes de transports.

D'accord Mitigé Pas d'accord

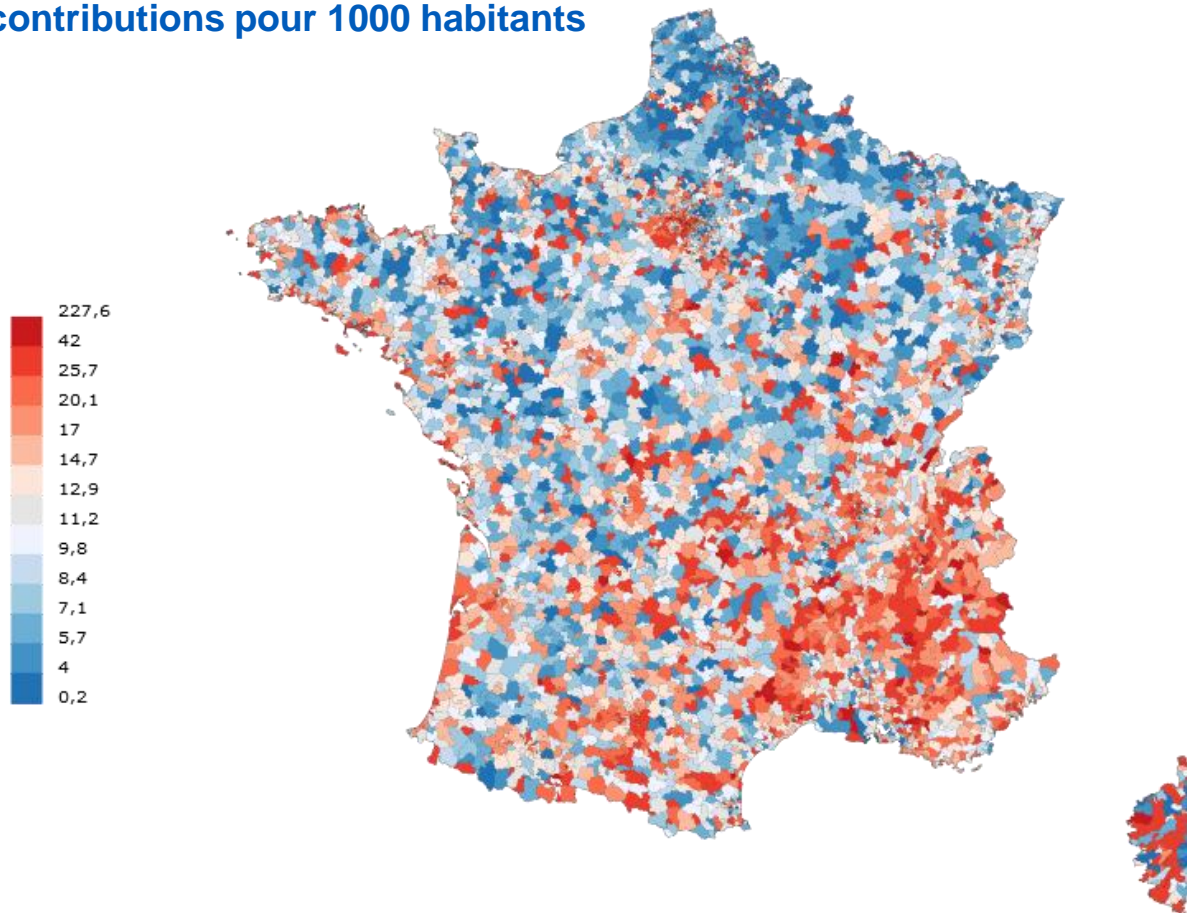
+699
704 votes

Catégorie	Votes
D'accord	434
Mitigé	116
Pas d'accord	154

13 questions ouvertes

Participation au Grand débat National- thème Transition Environnementale

Nombre de contributions pour 1000 habitants



Corpus

► Grand Débat National Transition environnementale et Vrai Débat Gilets Jaunes Transition Environnementale

taille	Vrai débat et arguments		Entendre	
	Vrai débat	Grand Débat National la France		
Nombre de textes:	2 599	6 373	87 552	39 430
Nombre de formes:	17 707	22 380	78 829	34 582
Nombre d'occurrences:	225 039	351 991	21 764 365	1 273 520
Nombre de lemmes:	12 059	15 034	78 829	22 376

***Pour la reconnaissance de l'Age, du genre et du soutien aux Gilets jaunes**

► Entendre la France

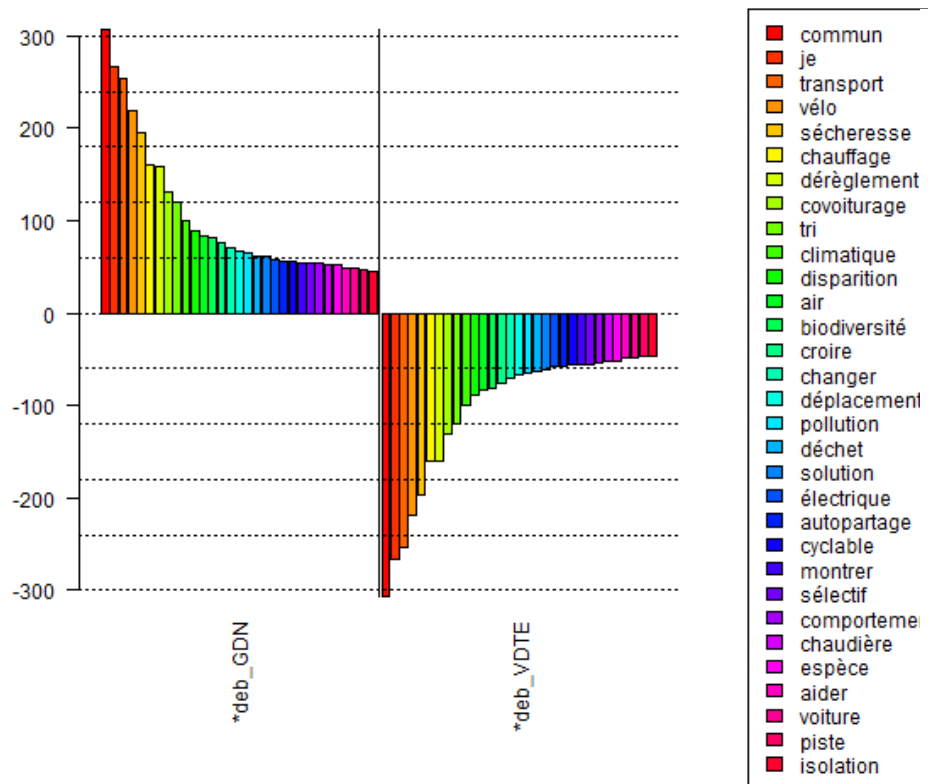


**RETROUVER UN DISCOURS
GILETS JAUNES DANS LE GRAND
DÉBAT NATIONAL ?**

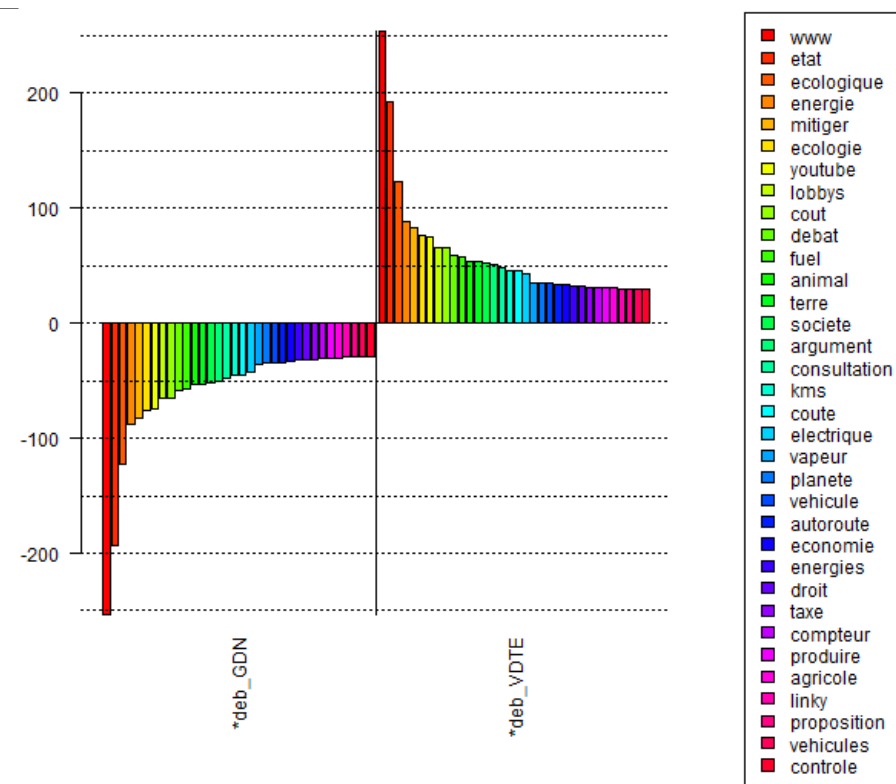
**ANALYSER LES DISCOURS- ADT
ET IA**

Spécificités: Grand Débat National et Vrai Débat

Grand Débat National



Vrai Débat



Spécificités: Grand Débat National et Vrai Débat (environnement)

Grand Débat National

1- Problèmes

- ◆ **Dérèglement**, climatique, transition, sécheresse, érosion, déforestation (mondial, global, planète)
- ◆ **Pollution**, pollueur, pesticide, fossile (diesel, charbon)
- ◆ **Biodiversité**

2- solutions

- ◆ **Verbes/action** : changer, arrêter, interdire, réduire, diminuer, développer, encourager, inciter, privilégier, promouvoir, sensibiliser
- ◆ **Prise de conscience**, comportement, éduquer, responsable, éducation, habitude, effort, courage, citoyen
- ◆ **Économique**, croissance, investir, entreprise, industrie

◆ **Modalisation** : croire, falloir



Vrai Débat

Web, you tube, internet

1- Problèmes

- ◆ **Radar**, vitesse, route autoroute
- ◆ Carbone
- ◆ **Alimentation**, alimentaire, fruit, fruitier, légumes, étiquetage, label paysans, producteurs, distributeurs, PAC
- ◆ **Animal** abattoir, mort, souffrance
- ◆ **Sécurité**, accident, santé
- ◆ **Prix**, euro, montant, facture, gratuité, vendre, vente, spéculation, marché
- ◆ Linky, compteur, onde, Bure, EDF
- ◆ **Macron**

2- solutions

- ◆ **Verbes/action** : refuser, suspendre, exiger, permettre
- Obligatoire; interdiction, suppression, création
- ◆ **Privatisation**, privé,, nationaliser, public, coopératif
- ◆ Ligne, sncf, réseau, barrage
- ◆ **Modalisation** : pouvoir

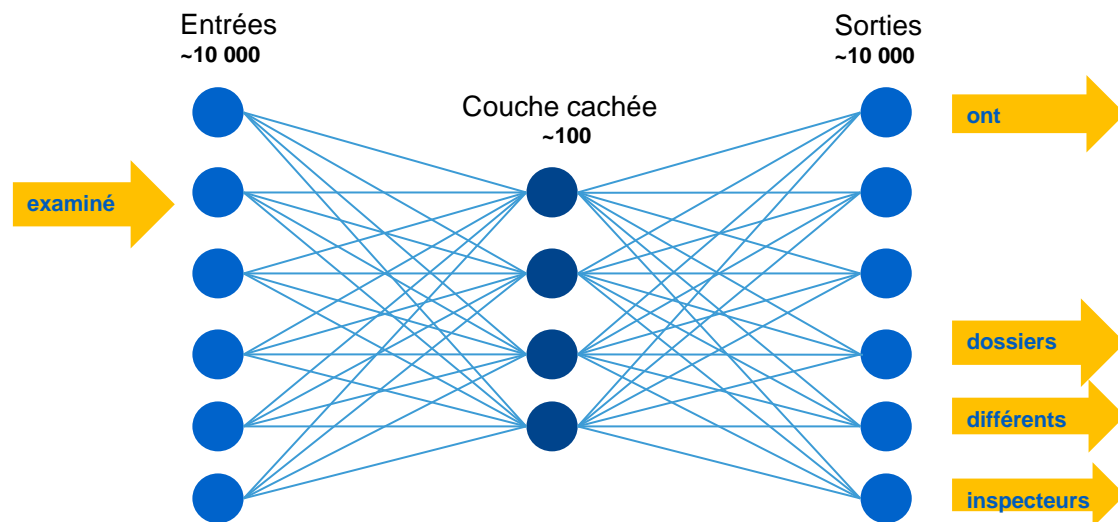
Embeddings de mots - Word2Vec

► Algorithme proposé en 2013 - 2014

- Th. Mikolov – Google
- Rupture par rapport à l'approche « Sac de Mot » (TF-IDF...)

► Principe de l'algorithme

- Recherche à entrainer un réseau de neurones à **prévoir le contexte** d'utilisation d'un mot
- Méthode totalement non supervisée
- Nécessite des corpus « importants »
- Exemple avec : « ... inspecteurs ont **examiné** différents dossiers ... »



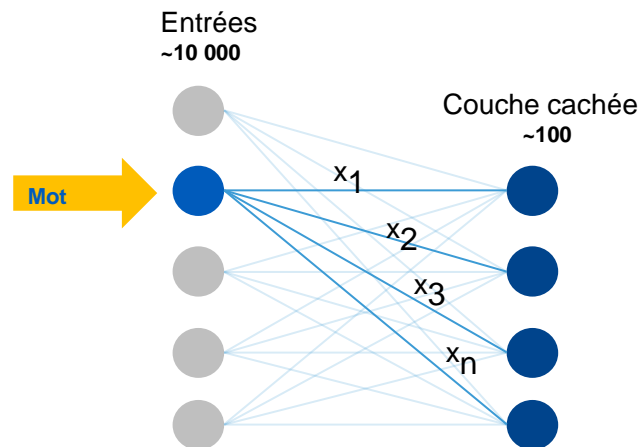
Embeddings de mots - Word2Vec

► Intérêt

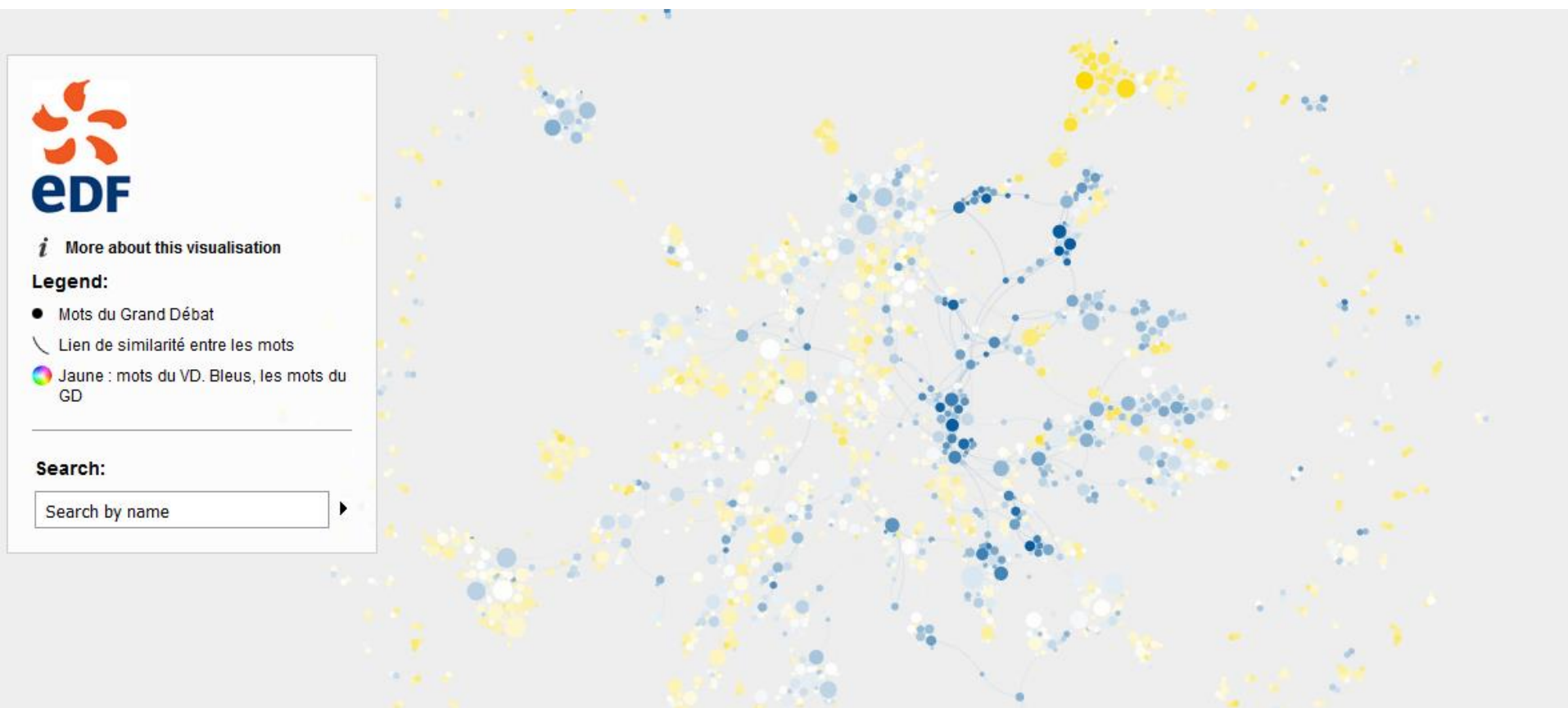
- Permet de transformer un **mot** en **vecteur** (la fameuse couche cachée)
- Passer d'une représentation **discontinue** (les mots) à une représentation **continue**
- Ici, « mot » est transformé en un **vecteur** x_1, x_2, \dots, x_n
- Méthode de plongement (« **word embeddings** ») ou **vecteurs mots**
- Permet de faire des sommes de mots, soustraction de mots, moyennes de mots...

Facture	0,21	-0,33	0,81	-0,23	...	0,45
Index	0,67	-0,01	0,54	0,33	...	0,33

- Calcul de **similarité** entre mots (avec cosinus par exemple)
 - Cosinus (facture, index) = 0,321
 - Cosinus (facture, sociologie) = 0,07
- Permet de trouver des **champs/univers linguistiques**
- Deep learning => **shallow learning**



Classer/explorer avec l'IA : Les mots du VD dans le GD



Les points : les mots du Grand Débat

Les liens : similarité entre les mots

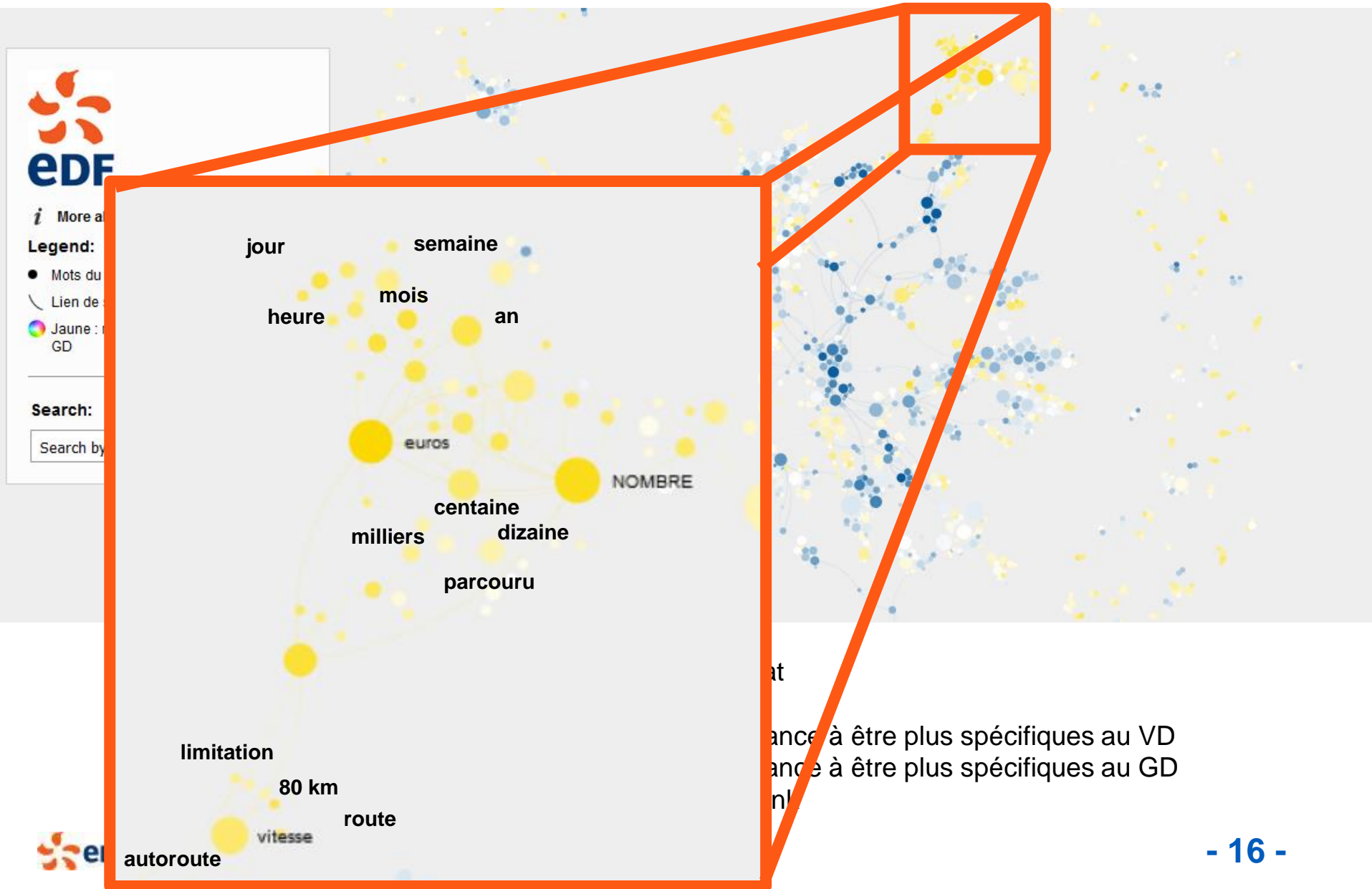
Couleur jaune : les mots ayant tendance à être plus spécifiques au VD

Couleur bleue : les mots ayant tendance à être plus spécifiques au GD

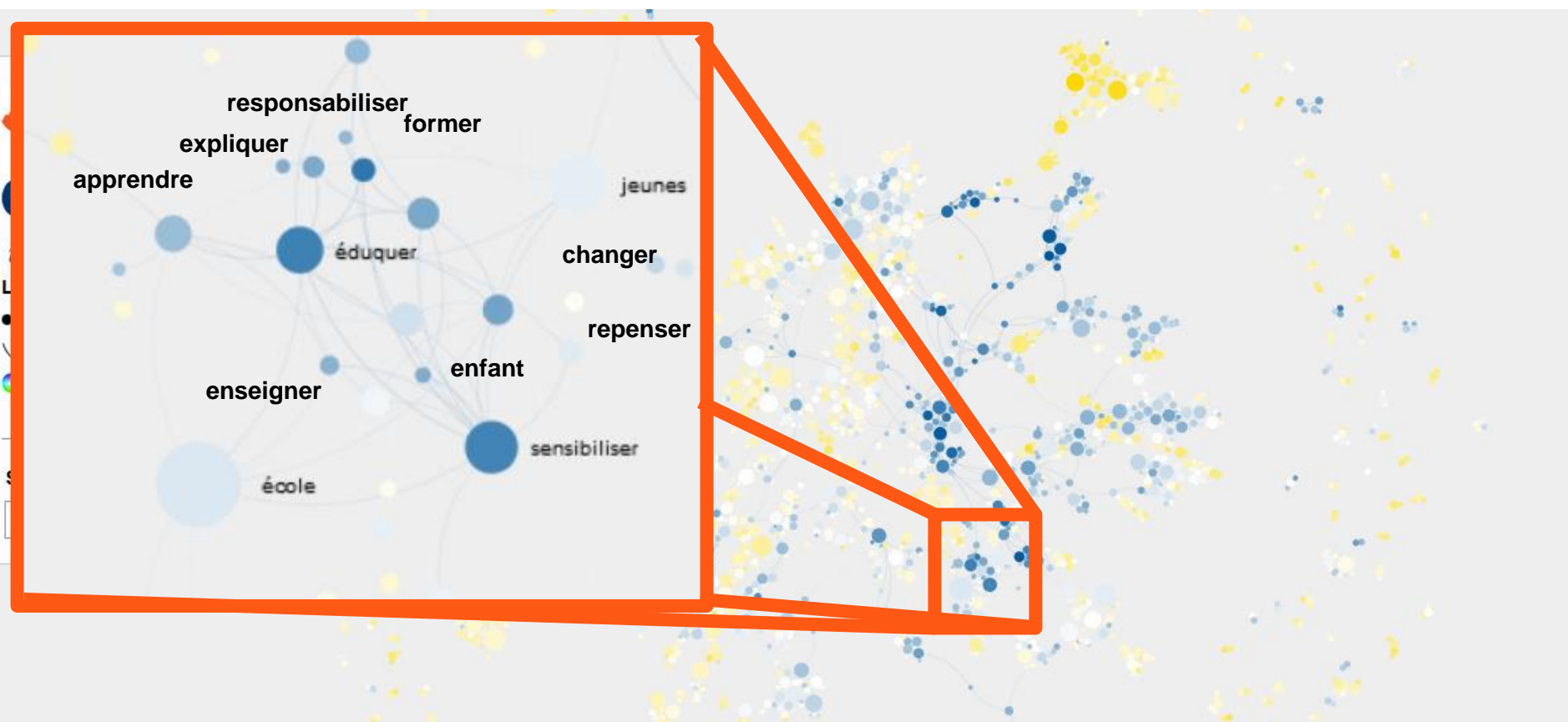
Taille des mots : algorithme PageRank

Visualisation : Gephi

Classer/explorer avec l'IA : Les mots du VD dans le GD



Classer/explorer avec l'IA : Les mots du VD dans le GD



Les points : les mots du Grand Débat

Les liens : similarité entre les mots

Couleur jaune : les mots ayant tendance à être plus spécifiques au VD

Couleur bleue : les mots ayant tendance à être plus spécifiques au GD

Taille des mots : algorithme PageRank

Visualisation : Gephi

Classer/explorer avec l'IA : Les mots du VD dans le GD

Grand Débat National

◆ **Chimique**, pesticide, herbicide

Dangereux, polluants

◆ **Cargos, bateaux**

Lourds, kérosène, charbon

Taxer, sanctionner, lourdement, durement

Forts, drastiques, contraignant

Mesures, actions concrètes

Arrêter, stopper, abandonner, interdire, limiter


Promouvoir, favoriser, développer, privilégier, encourager

◆ **Pollueurs**, industries, multinationales

◆ **Citoyens**, comportements, habitudes

Modifier, changer, revoir, radicalement

Éduquer, informer, responsabiliser, sensibiliser, apprendre, école, gestes ,

 **Niveau**, international, européen, mondial, local
Inde, Usa, Etats unis

Vrai Débat

◆ **Nombre**, Km, millions, milliards, euros, années

◆ Limitation, vitesse, route, autoroute

◆ **Lyon**, Paris, Marseille, Bordeaux, Toulouse

◆ Essence, éthanol, hydrogène, hybride

◆ **Produits**, venant, importés, importations
fabriqués, achetés

◆ **Parking**, péage, train, vélo, RER, métro, trajet,
travail, domicile

◆ **Taxe**, TVA, surtaxé, écotaxe

Coût, prix, tarif,

Montant, prêt, crédit, impôt

Prix, aide,

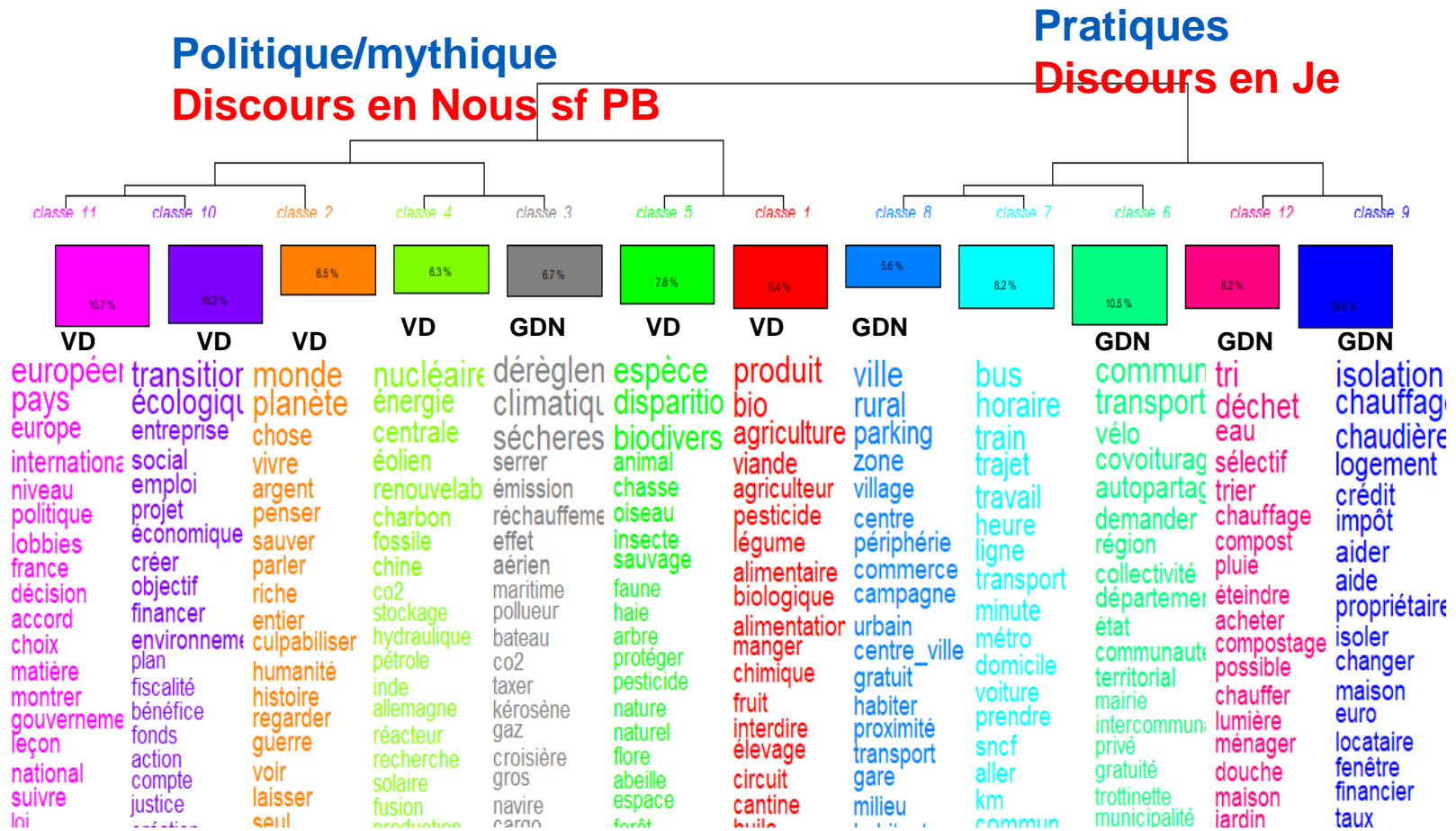
Faible, bas, élevé,

Revenu salaire

◆ **Cantine**, scolaire, végétarien, budget

◆ **Céréales**, lait, légumes, fruits

Corpus Global TE, approche hiérarchique : Grand Débat et Vrai débat (avec arguments)



Une comparaison des types de mots

	Grand débat National	VD_Proposition	VD_Arguments
* Verbes :			
Factif	64,40%	59.5%	49.0%
Statif	21,40%	24.6%	32.3%
Déclaratif	14,10%	15.8%	18.4%
Performatif	0,10%	0.1%	0.4%
* Connecteurs :	Grand débat National	VD_Proposition	VD_Arguments
Condition	3,10%	4.0%	6.4%
Cause	6,10%	8.0%	11.8%
But	2,40%	2.7%	1.4%
Addition	64,10%	56.9%	45.1%
Disjonction	8,50%	10.9%	8.0%
Opposition	8,80%	9.8%	18.5%
Comparaison	4,90%	5.2%	5.4%
Temps	2,20%	2.4%	3.5%
* Modalisations :	Grand débat National	VD_Proposition	VD_Arguments
Temps	10,2%	11.6%	10.5%
Lieu	8,1%	9.0%	7.5%
Manière	12,1%	12.1%	9.9%
Affirmation	3,8%	4.1%	7.5%
Doute	0,4%	0.3%	0.5%
Négation	16,7%	17.4%	22.4%
Intensité	48,7%	45.5%	41.7%

		Grand Débat National	Vrai Débat
Taille		3033858	333897
			dont 75519 mots (2356 arguments) pour
			et 46085 mots (1418 arguments) contre
* Adjectifs :	Grand débat National	VD_Proposition	VD_Arguments
Objectif	54,4%	51.3%	41.5%
Subjectif	36,2%	33.1%	43.6%
Numérique	9,3%	15.6%	14.9%
* Pronoms :	Grand débat National	VD_Proposition	VD_Arguments
Je	10,7%	10.3%	20.1%
Tu	0,6%	0.3%	1.2%
Il	32,6%	30.1%	25.2%
Nous	13,5%	13.5%	8.3%
Vous	2,3%	2.5%	8.5%
Ils	7,9%	9.7%	6.8%
On	13,9%	15.7%	17.1%

Arguments pour vs contre = Vous 6.2% **11.7%**
- 20 -

**RETROUVER UN PUBLIC GILETS
JAUNES DANS LE GRAND DÉBAT
NATIONAL ?**

CARACTÉRISER LES LOCUTEURS

Caractériser les locuteurs

► Corpus « Entendre La France »



► Prédire les variables suivantes

- Sexe : homme/femme
- Age : 4 catégories
- Position vis-à-vis des gilets jaunes : 2 catégories



► Méthodes

- Machine learning/IA : apprentissage sur 70% du corpus et test sur 30%
- Différentes méthodes :
 - Approche bayésienne basée sur les mots
 - Régression logistique
 - Embeddings de mots, puis embeddings de documents
 - MLM (Masked Language Model) de type BERT
- Concaténation de toutes les réponses pour un répondant dans un document
- But : prédire la catégorie du document

Variables des répondants



► Age

18-24 ans	6628	jeune	6628
25-34 ans	2053	jeune actif	2053
35-44 ans	578	actif	1118
45-54 ans	540		
55-64 ans	502	sénior	838
65-74 ans	292		
75+ ans	44		
	10637		10637

► Sexe

Homme	7671
Femme	4727

► Soutien aux gilets jaunes

Soutient	2825	Soutient	3537
Ne soutient pas	2741	Ne soutient pas	2741
Participe activement	712		

Résultats

► Prédiction « Soutien Gilets Jaunes » sur « Entendre la France »

	Précision	Rappel	F-Mesure
Naive Bayes	0,664	0,667	0,665
Regression	0,648	0,645	0,646
Embeddings Docov	0,645	0,648	0,641
BERT	0,680	0,609	0,589

► Méthode finale

- Les 3 premières méthodes
- + mécanisme de vote

Résultats sur « Grand Débat »

► Prédiction « Soutien Gilets Jaunes » sur « Entendre la France »

■ Test sur le corpus « Entendre la France » (~1000 documents)

Soutient	56,3%	jeune	62,3%	Homme	61,9%
Ne soutient pas	43,7%	jeune actif	19,3%	Femme	38,1%
		actif	10,5%		
		sénior	7,9%		

■ Résultats sur « Le Grand Débat » (87 552 documents)

Soutient	36,0%	jeune	39,0%	Homme	76,0%
Ne soutient pas	64,0%	jeune actif	40,8%	Femme	24,0%
		actif	16,5%		
		sénior	3,7%		

➤ Production de 3 variables étoilées pour Iramuteq

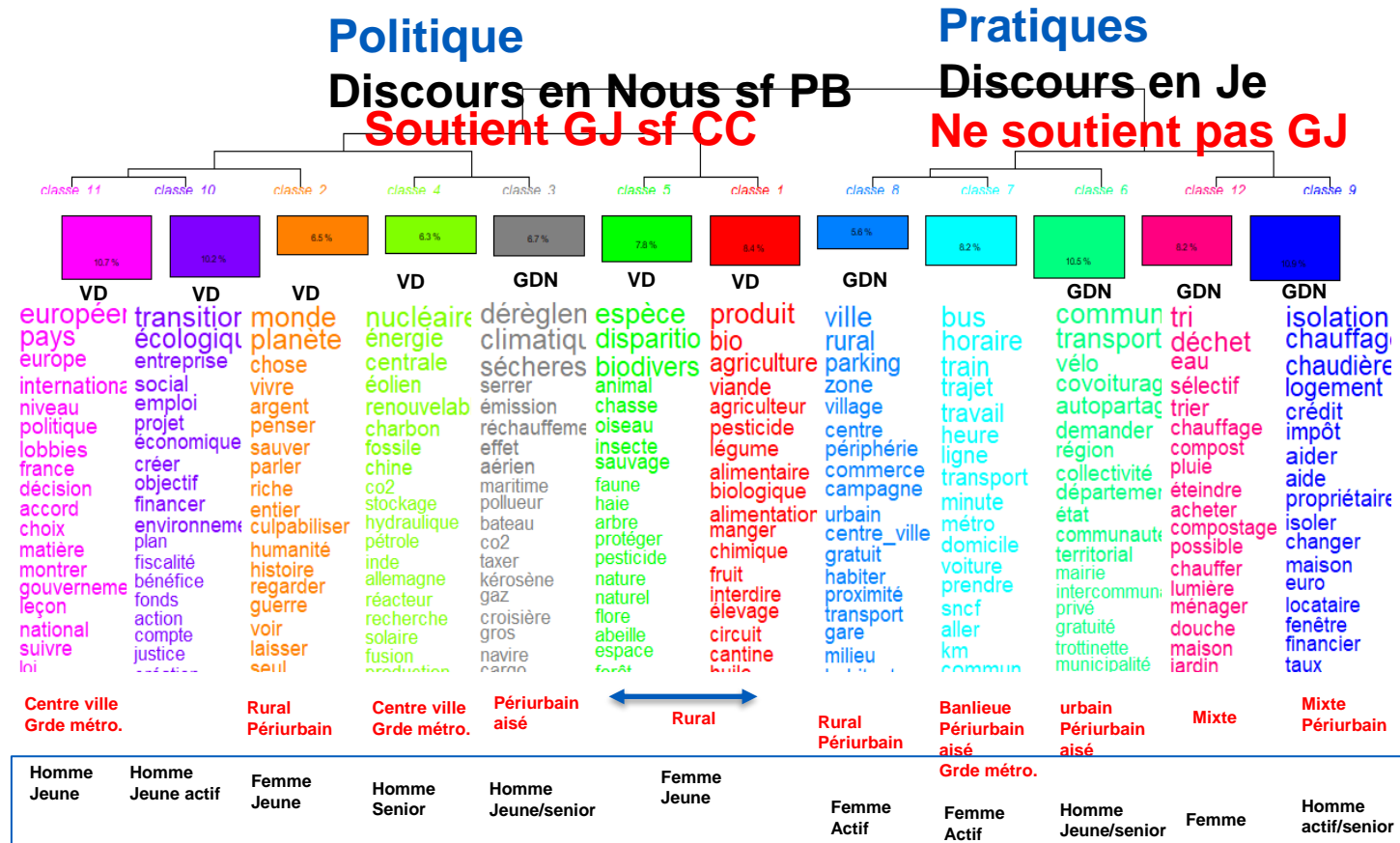
Les mots caractéristique des classes

Sexe	
Homme	Femme
investissement	enfants
taxe	déchets
taxation	parents
impôt	école
quota	vais
plan	consomme
progressif	peux
lourdement	vrac
fiscale	recycler
calcul	achète
annulation	mange
sortir	essaye
désastreux	courses
excessive	mes
financée	prends

Age			
actif	sénior	jeune actif	jeune
composants	fonctionnaires	production	faudrait
réduite	cour	polluante	éducation
impôt	impôts	produisent	sensibiliser
cour	composants	achat	pense
taxe	impôt	produit	école
ancien	impôts	produits	végétarien
voitures	servent	limitant	sensibilisation
vs	massives	hygiène	devrais
isolation	dépenses	pêche	écologie
ogm	hauts	toxiques	importance
salaire	retraite	agriculture	commencer
chinois	supplémentaires	aviation	réserve
croisière	mettez	optimisation	simple
compagnies	ministres	respectueuse	jeunes
ethanol	augmentation	carbone	sujet

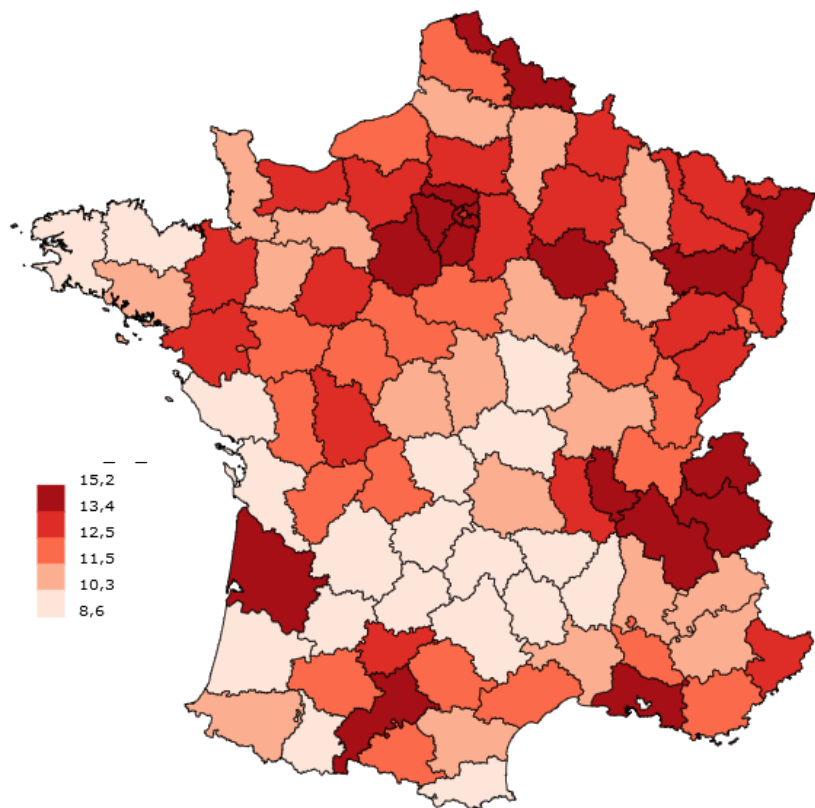
Gilets jaunes	
Soutient	Ne Soutient pas
monsanto	école
peuple	éducation
grosses	apprentissage
polluant	sensibilisation
groupes	information
paquebots	isolation
compter	jeunes
arrête	informer
entreprises	informations
arrêter	pied
mines	journées
lobbys	abordables
avions	accès
commencent	citoyennes
puissants	compostage

Corpus Global TE : Grand Débat et Vrai débat (avec arguments), la question des pseudo-propriétés

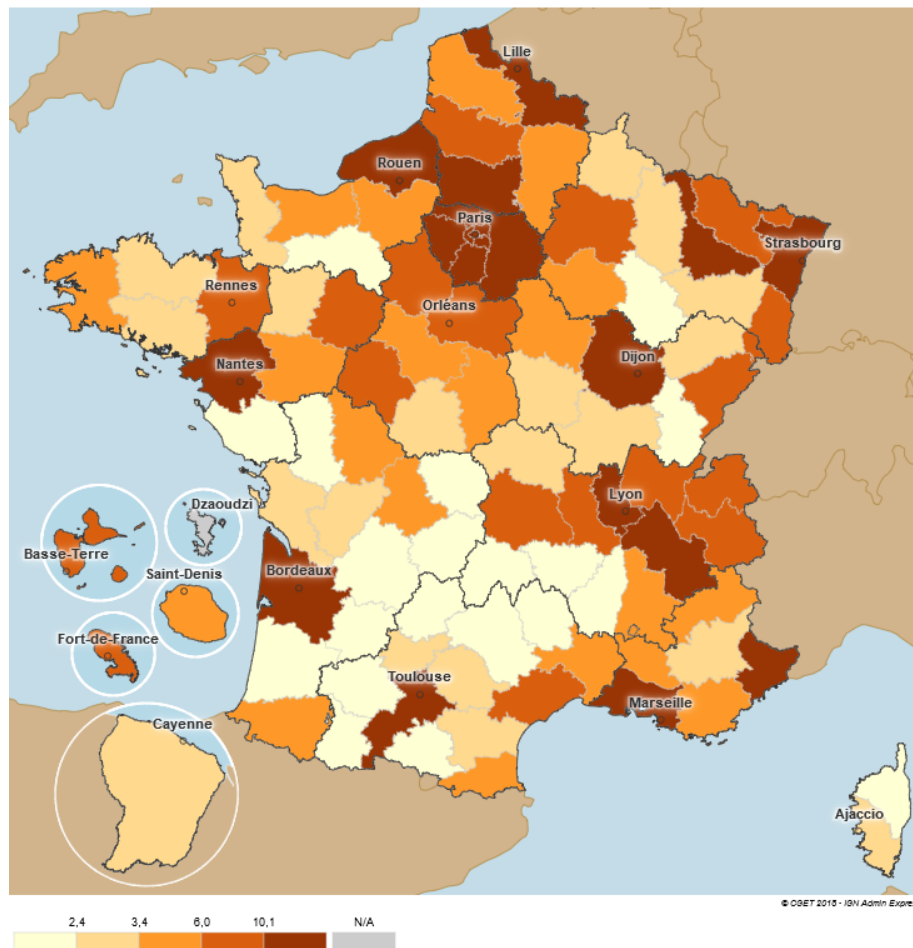


Comparaison des cartes de la classe sur la mobilité et part des déplacements en transports en commun

Part des verbatims traitant des déplacements (%)



Part des déplacements domicile-travail en transports en commun, 2016 (%) - Source : Insee, RP 2006-2011-2016





Mr. Smith Goes to Washington – Columbia pictures- 1939

CONCLUSION ET PERSPECTIVES

Conclusion

I- Sur les discours :

1- cadrages

- Cadrages GDN : biodiversité, dérèglements climatique, Taxer les transports
- Cadrages VD : Agriculture intensive/alimentation, croissance/décroissance/humanité/démographie, débats scientifiques/lobbys/politique
- Cadrages plutôt communs : action politique/citoyen, transport en commun, énergie

2- Formes de discours : Discours plus argumentatif VD. GDN, discours orienté sur l'action.

II- Effets sensibles de design des plateformes :

- GDN cadrages questions ouvertes, énoncés orientés problème/solution
- Formes argumentatives du VD.

III- Sociologie des participants

- Discours de ceux qui soutiennent ou ne soutiennent pas les Gilets Jaunes différents (grâce à Entendre la France) : montre des effets indépendants des effets du design des plateformes. Corrobore l'analyse VD vs GDN, plus politisé, dénonciation des effets du pouvoir/pollueurs.
- Dans le GDN, avons-nous identifié un public Gilets Jaunes ou public soutenant les Gilets jaunes (ou les 2) ?
- Problèmes posés par la reconstitution de données manquantes (propriétés sociales et politiques des locuteurs) : dans quelles mesures peut-on extrapoler du corpus entendre la France au GDN ?, quel est le statut de ces affectations probabilistes ?

IV- Algorithmes

- Alceste/ Iramuteq : mis en évidence des cadrages/hiérarchisation
- IA : analyse fine des thématisations par des champs lexicaux. Approches plus paradigmatique que syntagmatique. Efficacité en termes d'indexation/classification sur le modèle du traitement des questions ouvertes que l'on referme (cf. exemple de Proxem).
- Approches des formes de discours/énoncés par des analyses textométriques/lexicométriques classiques

Perspectives et références

Discours

- Argumentation
 - Sujet qui se développe en ce moment (Workshop internationaux « Argument Mining »)
 - Détecter les arguments mis en œuvre dans les verbatims (pg avec l'INALCO)

Approche sociologique

- Détecter des compétences sociolinguistiques
- Variable socio-démographique ou d'attitude vue comme des probabilités plutôt qu'un état

Algorithmes

- Caractériser les épistémologies embarquées entre Iramuteq/algos IA
- Explicabilité de l'IA : « machine à poser des questions » pour les sociologues

Références

Un discours et un public « Gilets Jaunes » au coeur du Grand Débat National ? Combinaison des approches IA et textométriques pour l'analyse de discours des plateformes « Grand Débat National » et « Vrai débat »

Mathieu Brugidou, Philippe Suignard, Caroline Escoffier, Lou Charaudeau

JADT 2020 : 15es Journées internationales d'Analyse statistique des Données Textuelles, Université de Toulouse - Jean Jaurès, Jun 2020, Toulouse, France. pp.1-12

Que peuvent les algorithmes de plongement de mots pour l'analyse sociologique des textes ? Analyser les discours et caractériser les locuteurs des plateformes « Grand Débat National » et « Vrai Débat »

Suignard Philippe, Caroline Escoffier, Lou Charaudeau, Mathieu Brugidou

Statistique et Société, 2021, Gilets jaunes et Grand Débat National : outils, données et analyses., 9 (1-2), pp.133-145