proxem

Construction de thésaurus assistée par Machine Learning

Sommaire

- À propos de Proxem

- Initier un thésaurus
  - Les modèles probabilistes de détection de thèmes
  - L’apport des représentations vectorielles de mots

- Compléter un thésaurus préexistant
  - Amélioration du rappel
    - Prise en compte de la synonymie
    - Quelques exemples
  - Amélioration de la précision
    - Prise en compte de l’homonymie
    - Exemples de désambiguïsation

- Conclusion

À propos de Proxem

English_496
I have emailed your complaints team with my concerns and p

English_952
Not good. Dirty. Bad smell

English_958
Great apts,nice and clean.well furnished and great kitchen.ni

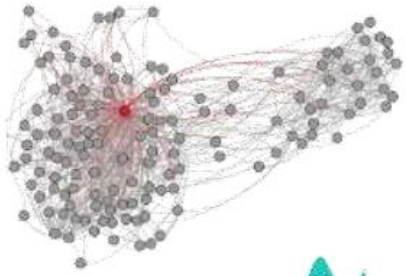
English_883
It smelled from urine in our apartment

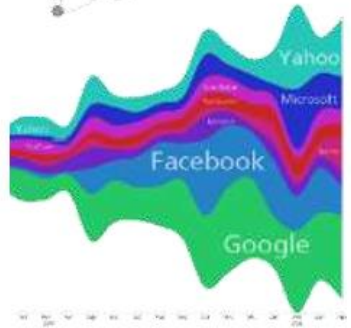
English_1231
Great value, totally relaxing...thank you.....but please get dog

English_966
The floor was verry dirty.

English_1346
Absolutely perfect apart from a slight 'drains' smell whenever

Text mining → Data mining → Dataviz





Initier un thésaurus

« Approche bottom-up »

Détection de thèmes

Modèle : **Allocation de Dirichlet Latente (LDA)***

- Extrait des thèmes d'un corpus de documents de manière non supervisée
- Chaque thème correspond à une distribution de probabilité sur les mots du corpus
- Fourni une classification automatique des documents

Avantages

- **Rapidité** (≈2mn pour apprendre **30 thèmes** sur un corpus de **100k documents, 30k mots de vocabulaire**)
- **Modularité** : possibilité d'incorporer des connaissances (sous forme de « must link » et « cannot link » entre mots du vocabulaire)

* : cf référence en fin de présentation

Exemples

Le corpus

- avis de consommateur dans le domaine de la grande distribution
- Avant preprocessing => 570 Mo, 2M de documents
- Après preprocessing => 100Mo, 700k documents

thème	Vocabulaire associé
Ecologie	écologique, déchets, recyclage, réduire
Accessibilité	véhicules, garer, voie, stationnement, bus
Fourniture scolaires	stylos, collège, écoles, école maternelle, règle
High tech	disque dur, clé usb, casque, asus, windows
Cuisine	poêle, casseroles, induction, casserole
Beauté	visage, tube, shampoing, spray, beauté

0,4330256
Suggestion : Communiquez sur la réduction des emballages , réduisez vos coûts de revient en supprimant les emballages_inutiles et réduisons nos impacts environnementaux ...

0,4329798
Est -ce bien écologique ? Autre inconvénient l' utilisation de cette boite n' est absolument pas pratique lorsqu' il faut en prélever pour les préparations culinaires .

0,432959
Bonjour , je suis auteur compositeur et j' ai écrit une chanson \ " la pollution \ " qui sensibilise les plus jeunes au développement_durable .

0,4329235
En matière de communication notre but est de fédérer des entreprises qui œuvrent pour les énergies_renouvelables , le développement_durable et encouragent les filières d ? excellence .

0,4329221
Je viens_de_découvrir de nouveau produits dont je voulais vous signaler , que je les ai trouvé bon mais le plus c' est qu' ils ne contenaient pas d' huile_de_Palme . Produit que je réachéterai . Continez votre effort sur l' huile_de_palme .

0,431649
L' Association " " Pour la planète " " aimerait établir un partenariat avec vous . L' Association propose de mettre votre logo sur les tracts , affiches ... ainsi que sur le site de l' Association , en échange d' une aide de votre part envers l' Association qui servira_à_financer une campagne de prévention du public ou bien une manifestation (ramassage de déchets dans la nature ...) .

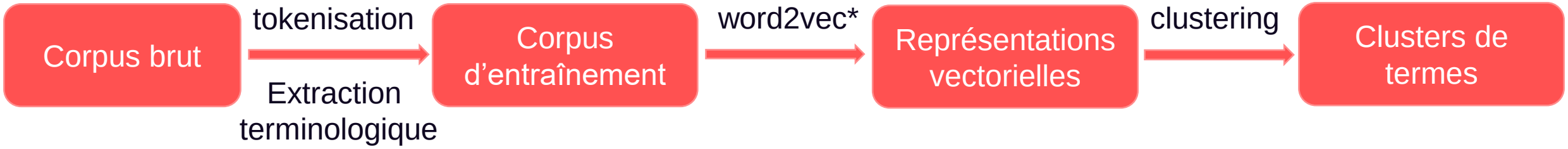
0,4306817
Cette cliente voudrait savoir si nous vendons du KONJAC ou SHIRATAKI c' est un produit pour réduire la faim

0,4305406
je suis mecontente encore une fois le container qui recycle les bouteilles plastiquesest en panne faute de maintenance . j aime la politique en matiere de recyclage mais pour ce faire il faudrait que le systeme a la hauteur des engagements en matiere d ecologie . a quoi bon stocker les bouteilles pour 1 euro si en arrivant sur place le recyclage n est pas possible .

L’apport des vecteurs de mots

« You shall know a word by the company it keeps » Firth (1957)

L’apport des vecteurs de mots



- Gel douche, shampoing 2 en 1, liquide vaisselle, le petit marseillais, …
  - shampoing, gel douche, …
  - liquide vaisselle, éponge, …
  - coloration cheveux, châtain foncé, châtain clair, …
  - lotion capillaire, mousse à raser, après rasage, crème hydratante , …
- Ordinateur portable, pc portable, téléphone portable, …
  - ordinateur portable, pc portable, appareil photo, …
  - lave linge, sèche linge, …
  - téléphone portable, smartphone, téléphone, …
  - centrale vapeur, cafetière, friteuse, …
- Quinoa nature, farine de châtaigne, chili con carne, …
  - Sucre semoule, quinoa nature, sucre poudre, sucre spécial confiture, …
  - Sorbet citron, thé vert menthe, sirop de violette, sirop menthe, sirop d’agave, ...
  - Fourme d’ambert, reblochon, bresse bleu, …

* : cf référence en fin de présentation

Compléter un thésaurus préexistant

Approche « top-down »

Objectifs

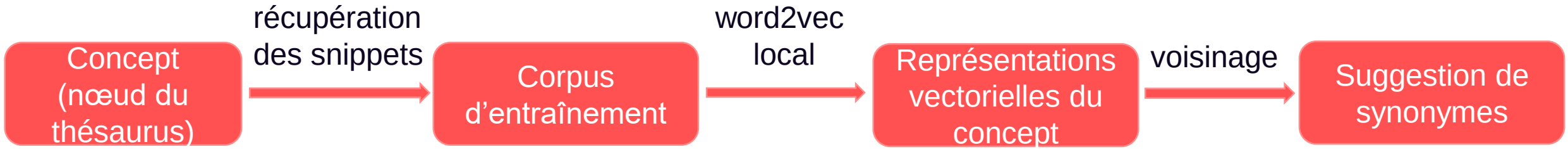
- Compléter un thésaurus obtenu avec les méthodes précédentes
- Adapter un thésaurus « général » à un corpus particulier

Amélioration de la précision par la résolution des ambiguïtés

Amélioration du rappel par la prise en compte des synonymes

Amélioration du rappel

- Adaptation des représentations vectorielles pour l'apprentissage de vecteurs de concepts **comparable aux vecteurs de mots**.



Exemples

Concept test « **fruits et légumes** » contenant les occurrences de termes suivants : aubergine(s), courgette(s), fraise(s), framboise(s)

Voisinage après apprentissage du vecteur de concept

- Tomates séchées
- Tomate cerise
- Pomme de terre
- Asperges blanches
- Asperges vertes
- Châtaignes
- Salade verte
- Courgette bio

Amélioration de la précision

- Adaptation du modèle « Adaptive Skip-gram »* qui permet d'apprendre plusieurs vecteurs par mots modélisant ainsi l'ambiguïté.

* : cf référence en fin de présentation

Exemples

couleur	télécom	fruit	jus de fruit
Pr : 39,672	Pr : 28,654	Pr : 20,221	Pr : 11,115
orange#3	orange#2	orange#5	orange#1
intense#3	sfr#3	oranges#1	joker#1
bleu#3	free#2	oignons#1	citron#1
framboise#2	bouygues#1	orange_joker#1	multifruits#1
cannelle#1	bouygues#2	orange_tropicana#1	ss_pulpe#1
menthe#1	sfr#4	bio#5	tropical#1
amandes_noisettes#1	abonnement#1	jus#4	multivitamine#1
blanc#3	numericable#1	pruneaux#1	tropicana#1
glacé#1	livebox#1	amande#1	reveil_fruite#1
vert#2	virgine#1	oignon#1	agrumes#1
rouge#2	mobil#2	abricots#3	x20cl#1
cassis#1	mobile#4	jus#1	orangina#1
ananas#3	telecom#1	epice#1	brique#1
arôme#1	samsung#5	endive#1	classic#1
allégée#1	prépayée#1	multi_vitaminé#1	oasis#1

Exemples : désambiguïsation en contexte

phrase	Couleur	Telecom	Fruit	Jus de fruit
orange	39.6	28.6	20.2	11.6
Pourriez vous me dire où je pourrais acheter des chaussons « pattes de tigre », **orange** et noir rayés ?	**97.5**	0.3	2.1	0.1
En juillet, SFR a écrasé ma ligne téléphonique professionnelle (**orange**) fixe et mobile	2.1	**96.9**	1	0
J'ai pris un filet d'**orange** du brésil, 2kg.	1.1	0.2	**98.7**	0
Je constate que le jus d'**orange** bio n'est plus présent dans les rayons.	0.6	2.2	19.8	**77.4**

Exemples

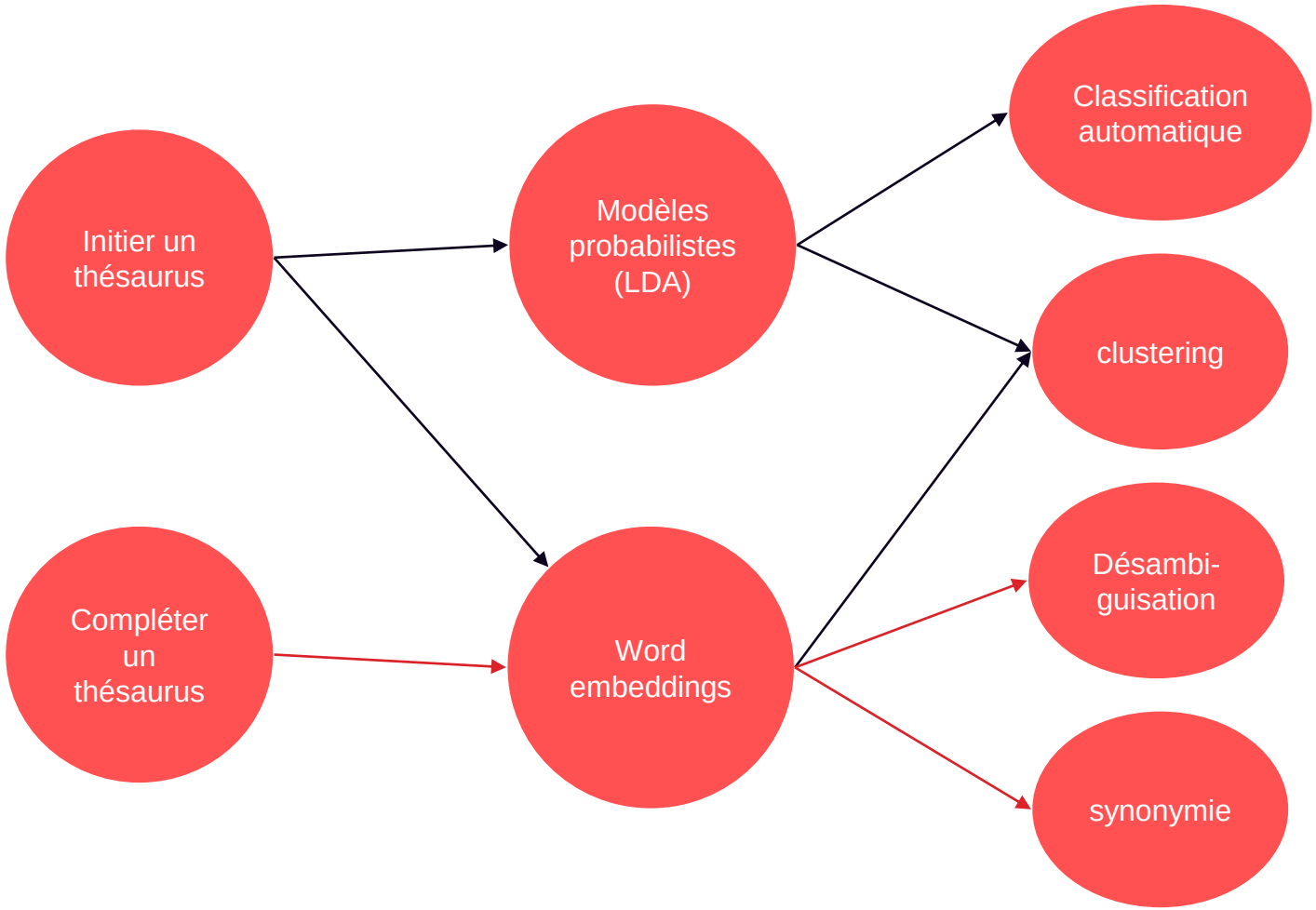
Processus de désambiguïsation

- Récupération des fenêtres de mots autour des annotation du concept
- Apprentissage de plusieurs vecteurs
- Inférence sur les snippets
- Statistiques sur les mots de contexte

Transformation en requête avec ajout d'inhibiteurs et d'activateurs

Sens	mots de contexte les plus fréquents
Couleur	article, intense, très, épices
Telecom	contrat, mobile, opérateur, souscrit, téléphonie, compatible, samsung
Fruit	Kilo, kg, filets, kgs, déguster, 2.49€
Jus de fruit	x1litre, unitaire, quantité, pur jus, briques

Conclusion



- Des gros progrès ces dernière années dans le domaine.
- Des outils précieux pour l'analyse de gros volumes.
- Première idée rapide sur le contenu d'un corpus de document
- Il reste nécessaire de connaître les limites des modèles…
- … et de toujours associer une validation humaine après chaque suggestion

References

- David M. Blei, Michael I. Jordan, Andrew Y. Ng. 2003. *Latent Dirichlet Allocation.*
- David Andrzejewski, Xiaojin Zhu, Mark Craven. 2009. *Incorporating domain knowledge int topic modelling via Dirichlet Forest priors*
- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space.*
- Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, Dmitry Vetrov. 2015. *Breaking sticks and ambiguities with Adaptive Skip-gram*

Merci pour votre attention. Questions?

Proxem – 105 rue La Fayette – 75010 Paris