Constitution de lexiques: extraction de termes et de collocations, variations de termes

REGUIG Ghiles

Sommaire

Extraction et exploitation de termes complexes dans un texte

Différentes formes de termes complexes

Mots composés

“pomme de terre”, “savoir-faire”,...

-› Indissociables, syntaxe fixe

-› Signification pas toujours compositionnelle

Expressions idiomatiques

“poser un lapin”, “casser sa pipe”

-› Syntaxe changeante

-› Signification non compositionnelle

Collocations

“caresser l’espoir”, “exercer une profession”

-› hautement variables (composantes non contigues, utilisation de synonymes)
-› signification en fonction du contexte

Extraction de termes

Approche syntaxique

Patron : Dét. Indéf. masc. sing. + Npr(s) + Adj. Qual. masc. sing.

Da-ms-i	(Np)+	Afpms
Segments repérés :		
un	Don Juan	têtu
un	Tom Jones	flou
un	Disneyland	bas de gamme
un	Bauhaus	new-age
un	Thésée	craintif

source : https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2004-1-page-25.htm

Approche statistique

Concordance | Concordance Plot | File View | Clusters/N-Grams | Collocates | Word List | Keyword List

Total No. of Cluster Types 20 **Total No. of Cluster Tokens** 138

Rank	Freq	Range	Cluster
1	17	1	de l'immigration
2	7	1	communiqués contre l'immigration
3	7	1	contre l'immigration
4	7	1	dans la liste immigration
5	7	1	français d'immigration
6	7	1	la liste immigration
7	7	1	office français d'immigration
8	7	1	ofii (office français d'immigration
9	7	1	premier dans la liste immigration
10	7	1	roms : ça suffit ! o immigration
11	7	1	référendum sur l'immigration

source : https://immigrationclandestineblog.files.wordpress.com/2017/01/clustersfra1.png?w=625

Approche hybride

-› Extraction basée sur la syntaxe/mesure statistique

-› Sélection basée sur un score de probabilité (cohésion du terme)

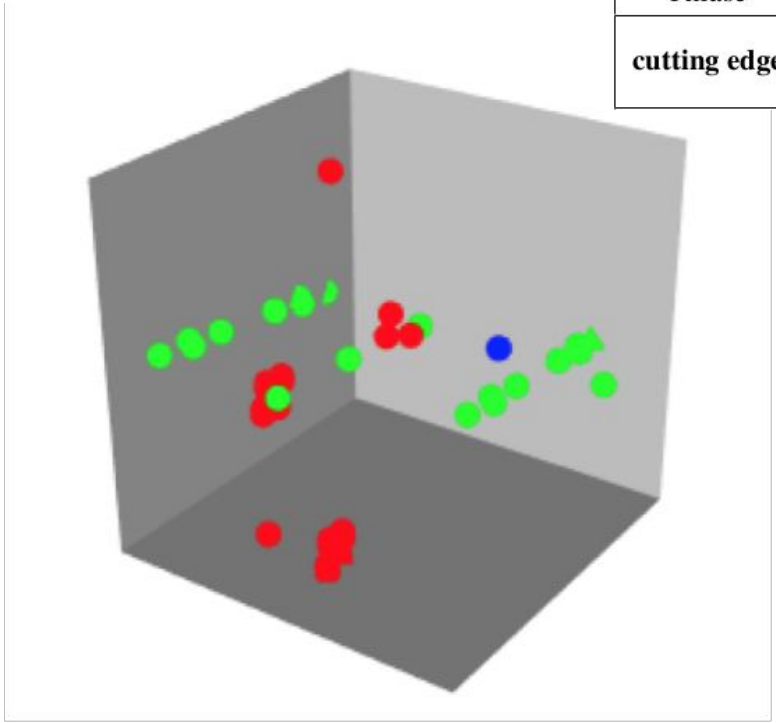
Représentation numérique et compositionnalité

Représentation vectorielle de termes complexes

1. *Représentation compositionnelle* : moyenne des représentations des parties
2. *Représentation en tant qu'entité* : chaque terme complexe est transformé en entité. *Exemple : New York -› New_York*
3. *Représentation contextuelle :* représentation du terme selon sa projection orthogonale en fonction du sous-espace de son contexte

Représentation contextuelle

Phrase	Compositional Context	Non-compositional Context
cutting edge	the flat part of a tool or weapon that (usually) has a **cutting edge**. Edge - a sharp side.	while creating successful film and TV productions, a **cutting edge** artworks collection.



The compositional context embeddings of cutting edge are denoted by green points, and the non-compositional context embeddings by red points. The embedding of phrase cutting edge is denoted by the blue point. Note that the phrase embedding is very close to the space of the compositional context while being farther from the space of its non-compositional context.

source : https://arxiv.org/pdf/1611.09799.pdf

Classification et prédiction d'offres d'emploi

Outils utilisés

Corpus

Constitution d'un corpus d'offre d'emploi par webscrapping sur www.monster.fr

-› Pas de "validation manuelle" des annonces.

Catégorie	Nombre d'annonces
Architecture, Creation et spectacle	142
Edition et Ecriture	315
Ingénierie	399
Marketing	272
Ressources Humaines	390
Services Administratifs	392
BTP et second oeuvre	371
Informatique et Technologies	399
Logistique et Réparation	288
Recherche et Analyses	173
Sécurité	24
Commercial/Vente	392
Gestion de projet/Programme	392
Juridique	191
Qualité/Inspection	397
Santé	398
Stratégie et Management	392
Comptabilité et Finance	153
Installation et Réparation	394
Production et Opérations	395
Restauration et Hôtellerie	296
Services Clientèle	388
Total	6975

Table 1: Répartition des annonces selon les catégories.

Représentation word2vec

Pré-traitements :

- passage du corpus en minuscule
- élimination des ponctuations
- découpage en phrases

Construction de différents espaces de représentation

UnigramAnnonces

Appris sur le corpus en ne prenant en compte que les unigrammes.

BigramAnnonces

Appris sur le corpus en prenant en compte les unigrammes et les bigrammes.

TrigramAnnonces

Appris sur le corpus en prenant en compte unigrammes, bigrammes et trigrammes.

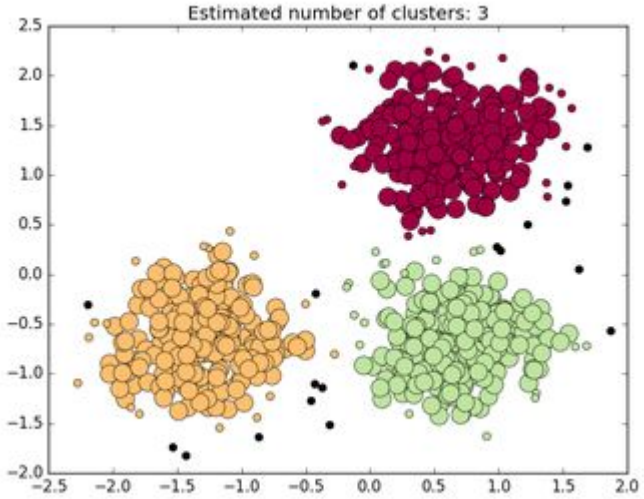
UnigramWiki

Modèle récupéré sur internet, entraîné sur wikipédia Fr.

Modèle	Taille du Vocabulaire
UnigramAnnonces	14 132
BigramAnnonces	19 474
TrigramAnnonces	20 659
UnigramWiki	475 475

Table 2: Taille du vocabulaire en fonction du modèle *word2vec*

Evaluation de modèle : DBSCAN



source : http://scikit-learn.org/stable/_images/sphx_glr_plot_dbscan_thumb.png

Représentation fréquentielle/présentielle

Fréquentielle :

Chaque document est représenté par la moyenne des représentations de mots pleins, pondérés par leur nombre d'occurences.

Présentielle :

Chaque document est représenté par la moyenne des mots pleins présents dans le document.

Modèle	epsilon									
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
LSA										
Nombre de clusters	56	64	74	87	102	115	124	149	162	151
Taille du cluster de bruit	6393	6273	6146	5988	5745	5494	5206	4835	4330	3634
Pureté moyenne des clusters	0.973	0.967	0.941	0.939	0.907	0.891	0.854	0.831	0.817	0.801
Variance de la pureté des clusters	0.012	0.012	0.024	0.022	0.026	0.033	0.045	0.050	0.053	0.055
UnigramAnnonces										
Nombre de clusters	75	110	26	14	7	5	3	2	2	2
Taille du cluster de bruit	6176	3825	857	246	103	46	27	13	10	8
Pureté moyenne des clusters	0.942	0.845	0.906	0.885	0.815	0.748	0.495	0.057	0.057	0.057
Variance de la pureté des clusters	0.018	0.046	0.048	0.062	0.116	0.160	0.192	0.0	0.0	0.0
BigramAnnonces										
Nombre de clusters	71	90	23	13	9	5	3	3	2	2
Taille du cluster de bruit	6191	3230	625	180	67	26	12	10	7	6
Purete moyenne des clusters	0.943	0.860	0.905	0.856	0.874	0.748	0.529	0.529	0.057	0.057
Variance de la purete des clusters	0.022	0.045	0.055	0.096	0.096	0.160	0.222	0.222	0.222	0.0
TrigramAnnonces										
Nombre de clusters	71	83	26	15	8	4	3	3	3	2
Taille du cluster de bruit	6179	3196	634	193	69	29	12	10	7	5
Purete moyenne des clusters	0.934	0.855	0.916	0.914	0.856	0.664	0.529	0.529	0.529	0.057
Variance de la purete des clusters	0.024	0.048	0.049	0.059	0.107	0.184	0.222	0.222	0.222	0.0
UnigramWiki										
Nombre de clusters	44	56	67	95	82	56	36	28	16	13
Taille du cluster de bruit	6550	6447	6295	5868	4838	3414	2243	1444	984	686
Purete moyenne des clusters	0.991	0.990	0.959	0.886	0.916	0.891	0.897	0.871	0.882	0.862
Variance de la purete des clusters	0.002	0.002	0.011	0.036	0.029	0.041	0.055	0.064	0.072	0.090

Table 3: Résultats de l'algorithme DBSCAN sur chaque modèle avec représentation fréquentielle

Démonstration

Conclusion