PARTITIONNEMENT DES DONNÉES À L'AIDE DE CENTROÏDES

Projet réalisé dans le cadre du cours 420-B62-IN PROJET ORACLE

AUTEURS

Jean-Charles Bertrand
Jean-François Lessard
Danick Massicotte

RÉSUMÉ

Le but de ce projet est d'utiliser le partitionnement des données à l'aide de centroïdes (centroïd-based data clustering) afin de faire ressortir des mots sémantiquement semblables d'un ensemble de mots préalablement acquis. Nous utiliserons quelques configurations d'analyse de texte ainsi que différents nombres de centroïdes afin de déterminer si une configuration est plus, ou moins, performante qu'une autre.

INTRODUCTION

Nous sommes maintenant tous familiers avec la puissance et l'efficacité de l'intelligence artificielle à regrouper et catégoriser des données, peu importe leurs provenances, qu'elles soient du domaine médical ou bien financier. En revanche, comment cette intelligence artificielle se débrouille-t-elle dans le domaine lexical? Probablement très bien, s'est-on dit, mais à quel point? Cet article fait état de notre démarche lors de la réalisation de trois tests utilisant le partitionnement de données à l'aide de centroïdes sur notre échantillon de mots. Vous trouverez le résultat de ces tests en trois sections distinctes dans ce document.

MATÉRIEL ET MÉTHODES

Notre banc d'essai se base sur le cumul des mots des textes suivants :

- Les trois mousquetaires (Alexandre Dumas, 1844)
- Le ventre de paris (Émile Zola, 1873)
- Germinal (Émile Zola, 1885)

Tous les textes ont d'abord été traités en cumulant le nombre d'occurrences où paraissent les mots en proximité les uns aux autres dans une phrase. Plusieurs tailles de fenêtres, plus précisément de 5 à 8 mots, ont été utilisées afin de renseigner notre base de données.

Une matrice de cooccurrences contenant les résultats du traitement précédent a ensuite été créée. Cette matrice de mots-vecteurs nous servira au calcul des centroïdes. La méthode de calcul qui a été utilisée dans tous nos tests est celle des moindres-carré (least-square). La quantité de centroïdes générés aléatoirement va de 10 jusqu'à 200 centroïdes calculés.

Est-ce que la taille de fenêtre lors de la création d'un certain nombre de centroïdes viendra aider à la précision du contenu de ceux-ci ?

Un de nos tests était de savoir si la taille de fenêtre avait une incidence directe sur la qualité du contenu des centroïdes. Une fenêtre de taille moindre rapporte-t-elle de meilleurs résultats qu'une fenêtre plus grande? Notre première intuition nous mène à penser que plus la taille de fenêtre d'analyse sera grande, meilleurs seront les résultats.

Voici les configurations de partitionnement de données que nous avons utilisé pour vérifier cette hypothèse :

- Nous avons généré des résultats pour les tailles de fenêtre 5, 6, 7 et 8.
- Le nombre de centroïdes a été fixé à 200 pour chacune des tailles de fenêtres.
- Les centroïdes ont été générés aléatoirement avec la méthode least-square.
- Dix mots sont affichés par centroïdes.

Résultats des données des centroïdes selon différentes tailles de fenêtre.

À la première lecture des résultats, nous remarquons immédiatement que la méthode de partitionnement des données par centroïdes est particulièrement efficace pour regrouper les verbes ayant la même consonance. Toutes les tailles de fenêtre, sans exception, contiennent des centroïdes de ce type.

Centroïdes regroupant des verbes se terminant par 'ANT'

Taille de fenêtre	5	6	7	8
Centroïde	criant	mettant	mettant	donnant
	mettant	donnant	criant	mettant
	donnant	criant	donnant	criant
	finir	passant	parlant	passant
	montrant	parlant	arrivant	parlant
	passant	finir	passant	montrant
	arrivant	arrivant	finir	prenant
	sortant	entrant	montrant	finir
	parlant	tirant	allant	jetant
	allant	outre	prenant	levant

Centroïdes regroupant des verbes se terminant par 'AIT' ou par 'AIENT'

Taille de fenêtre	5	6	7	8
Centroïde	sentaient	rêva	guettait	retrouvait
	parlaient	suffisait	suffisait	rencontrait
	virent	ajoutait	préférait	descendait
	voyaient	espérait	désespérait	rêvait
	disaient	hésitait	lâcha	craignait
	aperçurent	posa	espérait	reconnaissait
	seront	prend	secouait	paie
	entraient	préférait	nettement	partait
	aimaient	guettait	touchait	jura
	revenaient	reparut	traversait	expliqua

Centroïdes regroupant des verbes se terminant par 'IEZ' ou par 'EZ'

Taille de fenêtre	5	6	7	8
Centroïde	rappelez	trouvez	trouvez	aviez
	disiez	alliez	alliez	connaissez
	alliez	trouverez	trouverez	aurez
	trouvez	désirez	disiez	serez
	seriez	pourriez	taisez	devez
	taisez	feriez	désirez	étiez
	auriez	saurez	rappelez	comprenez
	ferez	appelez	pourriez	entendez
	désirez	disiez	direz	aimez
	trouverez	taisez	ferez	pouvez

Par contre, les centroïdes obtenus sont très semblables d'une taille de fenêtre à l'autre. La taille de fenêtre ne semble donc pas être un facteur déterminant dans les résultats obtenus.

On peut aussi remarquer que les centroïdes regroupent facilement les mots en anglais.

Centroïdes regroupant des mots en anglais

Taille de fenêtre	5	6	7	8
Centroïde	with	status	license	full
	work	owner	trademark	use
	for	world	about	license
	by	20	works	please
	is	following	ebooks	trademark
	other	right	electronic	distribute
	in	end	with	information
	any	defect	work	will
	ebook	electronically	is	have
	foundation	which	literary	about

Rien ici ne nous mène à croire que la taille de fenêtre joue un rôle clé dans la formation des centroïdes contenant des mots en anglais. Peu importe la taille de fenêtre, il semble que tous les mots en anglais se regroupent en un ou plusieurs centroïdes.

Là où nous avons obtenu des résultats particulièrement intéressants, est en trouvant un regroupement de mots désignant des parties du corps dans les résultats de la fenêtre de taille 5.

Centroïdes regroupant des mots désignant une partie du corps

Taille de fenêtre	5	6	7	8
Centroïde	doigts	ouvriers	camarades	camarades
	cheveux	enfants	enfants	mousquetaires
	oreilles	camarades	ouvriers	enfants
	dents	jambes	petits	pieds
	jambes	petits	pieds	gens
	lèvres	chevaux	gardes	gardes
	genoux	pieds	mousquetaires	épaules
	pieds	gardes	gens	choses
		lèvres	choses	maheu
		mousquetaires	épaules	femmes

Seule la fenêtre de taille 5 a su regrouper autant de mots désignant une partie du corps en un seul centroïdes. Chose intrigante, certaines parties du corps (mots) ne figurent tout simplement pas dans aucun centroïde des fenêtres de taille 6 à 8 (ex. doigts, cheveux).

Un phénomène encore plus intéressant se produit avec des mots désignant des chiffres et des nombres.

Centroïdes regroupant des mots désignant des chiffres et des nombres

Taille de fenêtre	5	6	7	8
Centroïdes	vingt	six	six	n.d.
	cent	vingt	vingt	n.d.
	ans	cent	cent	n.d.
	mille	mille	mille	n.d.
	francs	cinq	cinq	n.d.
	cinq	francs	huit	n.d.
	dix	huit	francs	n.d.
	huit	dix	dix	n.d.
	jours	ans	ans	n.d.
	ou	jours	jours	n.d.

Les tailles de fenêtre 5 à 7 nous donnent d'excellents résultats alors que la taille de fenêtre 8 ne réussit à regrouper aucun des mots figurants dans les centroïdes de taille de fenêtre plus petite. On s'explique difficilement cette situation vu la présence de tous ces mots dans notre entraînement à l'aide d'une fenêtre de taille 8.

Les données recueillies nous portent à croire que la taille de fenêtre tend à être moins efficace plus elle est de grande taille. Les résultats que nous avons obtenus avec la taille de fenêtre de 5 mots sont tout aussi bons, sinon meilleurs, que ceux des fenêtres de plus grandes tailles. Notre hypothèse de départ étant que plus une fenêtre d'analyse est grande, meilleurs sont les résultats, ne semble pas tenir la route. Le contraire serait plutôt valable.

Il serait intéressant de voir si nous obtiendrions des résultats semblables en utilisant un calcul de centroïdes de type City-Block plutôt que de type Least-Square, comme c'est le cas ici.

Des centroïdes de départ aléatoires donneraient-ils des résultats comparables d'un test à un autre?

Lors de réflexions au sujet des centroïdes de départs aléatoires, la question de reproductibilité des résultats est survenue : est-ce que deux séries de centroïdes aléatoires donneraient des résultats similaires si l'on garde les mêmes paramètres de tests, soit la même librairie de mots, la taille de fenêtre de mots lors de l'entraînement de l'algorithme et la cooccurrence des mots ainsi que le nombre de clusters voulus?

L'hypothèse derrière cette question est que les mots avec un sens ou contexte similaire se retrouvent sensiblement dans une même zone théorique et plus les mots sont similaires, plus cette zone sera dense donc aura un plus gros « attrait » lorsqu'un centroïde se retrouvera près. Donc, considérant que les paramètres outre les centroïdes de départ ne changent pas et que les mots restent les mêmes d'une série de tests à l'autre, les mêmes zones fortes devraient éventuellement tirer un centroïde vers leur centre, qu'il soit généré aléatoirement ou non au début du test. Certains centroïdes devraient donc être relativement similaires entre deux séries de centroïdes.

Résultats des tests de centroïdes aléatoires : série de clusters 1 vs. série de clusters 2

Pour débuter les tests de centroïdes aléatoires et les résultats qu'ils génèrent, nous avons testé nos paramètres sans facteur aléatoire : le test s'est effectué sur notre librairie de mots tirés des textes « Les Trois Mousquetaires » d'Alexandre Dumas, « Germinal » d'Émile Zola et « Le Ventre de Paris » d'Émile Zola avec une taille de fenêtre de cinq mots lors de l'entraînement.

Nous avons aussi généré aléatoirement vingt clusters pour le premier test, qui ont été sauvegardés pour référence future. Lorsque l'algorithme fut terminé, un deuxième test a été lancé avec les centroïdes initiaux sauvegardés, les mêmes mots des mêmes textes et la même taille de fenêtre. Sans aucune surprise, les résultats générés par les deux tests sont exactement les mêmes : les mots sont placés dans le même nombre d'itérations, dans les mêmes clusters aux mêmes positions avec les mêmes scores dans les deux cas. On peut donc conclure qu'il n'y a pas d'éléments qui affecteraient les résultats hors des paramètres mentionnés ci-haut.

Série 1

Groupe: 1 roi --> 12571.428571428572 soir --> 12595.714285714288 monde --> 16312.0 jour --> 18742.57142857143 grand --> 33908.28571428572 cardinal --> 66080.85714285714 jeune --> 85209.14285714284	Groupe: 8 attendre --> 1516.2912341407152 ketty --> 1651.4677047289504 pourtant --> 1837.2716262975778 entendre --> 1938.3500576701267 soit --> 2034.4088811995384 chaval --> 2169.722606689735 savoir --> 2500.3108419838527 va --> 2515.8794694348326 attendait --> 2642.8990772779703 sera --> 2652.7814302191455	Groupe: 18 monsieur --> 21371.9136 non --> 24235.5936 oui --> 24821.033599999995 donc --> 25221.753600000007 cela --> 29994.393600000007 porthos --> 40383.5136 aramis --> 42215.9136 moi --> 43528.873599999984 tu --> 70987.67360000001 dieu --> 80542.23359999999
Groupe: 1 roi --> 12571.428571428572 soir --> 12595.714285714288 monde --> 16312.0 jour --> 18742.57142857143 grand --> 33908.28571428572 cardinal --> 66080.85714285714 jeune --> 85209.14285714284	Groupe: 8 attendre --> 1516.2912341407152 ketty --> 1651.4677047289504 pourtant --> 1837.2716262975778 entendre --> 1938.3500576701267 soit --> 2034.4088811995384 chaval --> 2169.722606689735 savoir --> 2500.3108419838527 va --> 2515.8794694348326 attendait --> 2642.8990772779703 sera --> 2652.7814302191455	Groupe: 18 monsieur --> 21371.9136 non --> 24235.5936 oui --> 24821.033599999995 donc --> 25221.753600000007 cela --> 29994.393600000007 porthos --> 40383.5136 aramis --> 42215.9136 moi --> 43528.873599999984 tu --> 70987.67360000001 dieu --> 80542.23359999999

Série 2

Ensuite nous avons procédé aux tests avec centroïdes aléatoires pour les deux séries de clusters avec les paramètres de vingt clusters à générer, les mêmes mots que précédemment et une taille de fenêtre de cinq mots encore une fois. Voici une comparaison entre trois clusters plutôt intéressants :

Série 1

Groupe: 1 près --> 10123.611111111113 nouveau --> 10500.877777777776 voir --> 11117.877777777776 leurs --> 13367.54444444445 coups --> 13797.544444444446 venait --> 14230.21111111111 côté --> 14855.944444444445 ou --> 18979.544444444444 nom --> 19382.344444444443 suite --> 20796.27777777777	Groupe: 5 usages --> 2.4966386430197467 coque --> 2.551870173595922 bienfaits --> 2.561944404773017 continuité --> 2.602594811277082 vigilance --> 2.6249525348543172 guichetiers --> 2.654114782998538 abrutissement --> 2.685486292365806 embouchure --> 2.685486292365806 sommets --> 2.7025417890077286 approuvant --> 2.7290529236842924	Groupe: 7 toujours --> 5569.748721694668 enfin --> 5628.127100073048 alors --> 5832.289262235207 faisait --> 7381.4784514243975 maintenant --> 7810.775748721697 car --> 8080.397370343317 allait --> 8858.1811541271 disait --> 10101.910883856834 aussi --> 11216.721694667642 très --> 11565.640613586558
Groupe: 14 nouveau --> 7201.097777777777 près --> 8917.964444444446 coups --> 10112.097777777777 voir --> 10844.497777777775 leurs --> 11526.497777777779 côté --> 12937.031111111111 besoin --> 12964.964444444446 venait --> 14837.964444444442 nom --> 14979.831111111109 cet --> 16006.964444444444	Groupe: 1 usages --> 2.5655006156084887 coque --> 2.602518785076989 bienfaits --> 2.643357637174119 continuité --> 2.6597441204480194 vigilance --> 2.6952340168650686 abrutissement --> 2.746685876263948 embouchure --> 2.746685876263948 guichetiers --> 2.747365108731259 sommets --> 2.8035715954013227 approuvant --> 2.827599443932483	Groupe: 0 enfin --> 4151.282933454052 toujours --> 4746.516976007244 maintenant --> 6066.602082390221 car --> 6285.751018560435 alors --> 6515.963784517882 faisait --> 7349.112720688095 déjà --> 8406.346763241285 allait --> 8722.516976007244 aussi --> 9144.942507922138 florent --> 9546.516976007244

Série 2

Comme on peut le constater avec cet échantillon, certains clusters ont en effet été reproduits à plus ou moins deux mots de différences, malgré qu'ils ne soient pas exactement identiques. On voit que les mots peuvent être dans un ordre différent, ou que certains mots de la première série ont été remplacés par d'autres mots dans la deuxième série. On peut aussi remarquer que les scores attribués aux mots sont différents entre les séries, même si beaucoup d'entre eux sont similaires d'un test à l'autre. De plus, invisibles à l'échantillon ci-haut, beaucoup d'autres clusters n'ont pas de correspondance dans l'autre série.

Les résultats des tests effectués pour l'hypothèse que les mots se ressemblant plus attireraient inévitablement un centroïde et formeraient les mêmes clusters à chaque test semblent moyennement conclusifs: on remarque qu'en effet certains clusters se forment à nouveau d'un test à un autre, mais seulement quelques-uns et la plupart semblent ne pas se répéter dans les deux séries de clusters, ce qui peut potentiellement être dû à de nombreux mots n'ayant pas vraiment de sens commun, donc plusieurs zones à faible densité n'attirant pas de clusters fortement. Les résultats peuvent aussi être en lien avec un nombre relativement faible de clusters à générer. On pourrait essayer les mêmes tests avec deux cents clusters au lieu de vingt.

Manipulation des données de départ des centroïdes : piste d'intérêt pour bonifier les résultats?

Dans le but d'innover et de trouver une méthode entraînant des résultats supérieurs nous nous sommes intéressés à la genèse des centroïdes au temps zéro. L'intérêt étant que la détermination de la position de départ créant les clusters aurait une importance significative. (hypothèse à vérifier) . Une des hypothèses était qu'une certaine injection de connaissance pour moduler (amoindrir) l'alea dans la détermination des valeurs d'un centroïde pourrait mieux positionner celui-ci au départ plutôt que de courir le risque que l'alea le projette dans une zone vide.

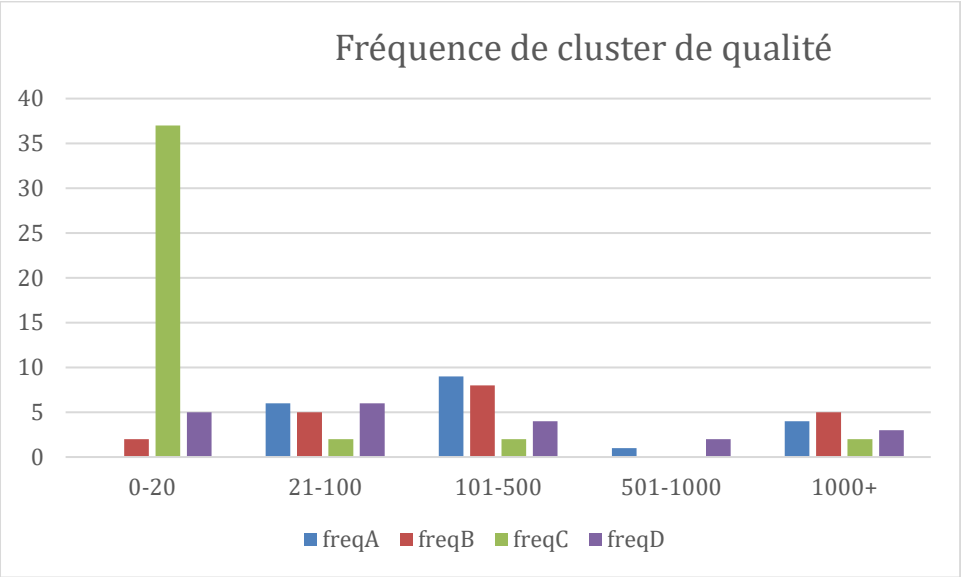
En particulier nous avons testé une méthode (méthode A) générant des valeurs aléatoires pour chaque position du vecteur/centroïde à sa création. (les valeurs possibles pour chaque 'colonnes' étaient de 0 au maximum rencontré dans n'importe quelle colonne).

Une autre méthode (méthode B) cherchait la valeur maximale rencontrée dans une colonne donnée et utilisait cette valeur comme plafond pour son tirage aléatoire.
Malheureusement, nous n'avons pas pu épuiser un nombre plus vaste de méthodes de variation de l'aléa. Une avenue possiblement intéressante que nous avons considérée est de pondérer les probabilités d'une valeur aléatoire en fonction de la représentation statistique d'une valeur. Il faut toutefois demeurer sceptique de la qualité du résultat obtenu, car de sa nature même la méthode aléatoire assigne des valeurs qui ne sont jamais couplées, or la nature même de la cueillette des cooccurrences génère des couplages entre colonnes. Conséquemment, la 'signature' des vecteurs aléatoires sera nécessairement trop distincte de la signature d'un mot un peu commun. Un vaste problème est issu du fait de valeurs nulles extrêmement sous représentées dans nos centroïdes aléatoires.

Conséquemment, nous avons tenté la méthode D consistant à assigner 0 dans 90% des cas, une valeur faible dans 8% des cas et une valeur aléatoire avec un plafond représentatif des valeurs locales.

Finalement, une autre méthode (méthode C) commence avec un centroïde avec des valeurs aléatoires et tous les mots avec un score least square dépassant un certain seuil crée un nouveau centroïde. Les points suivants sont évalués pour ces 2 centroïdes et si le seuil est franchi dans les 2 cas, à nouveau on crée un nouveau centroïde. Cette méthode n'est pas du tout pratique et l'algorithme actuel bien qu'il tente de pallier au problème d'excès massif de centroïdes qu'un seuil bas génère, il n'est pas possible d'assigner a priori le nombre de centroïdes désiré. Conséquemment, le seuil a été modifié pour être altéré dynamiquement en fonction du nombre de centroïdes en place. Plus il y en a, plus on élargit le seuil pour réduire la production de nouveaux centroïdes. Un seuil fixe générait trop de centroïdes « très distants » voir valeurs éloignées. Conséquemment, ces centroïdes étaient peu peuplés et donc la qualité du résultat était très médiocre.

Résultats d'évaluation de l'hypothèse de gestion d'injection de modes d'atténuation de création aléatoire de centroïdes



Ce tableau dénombrant la fréquence qu'un cluster généré était de qualité pour chaque type de méthodes évaluées (avec 20 clusters (ABD), fenêtres de taille 7) n'est pas très concluant de toute évidence. Certes, le modèle avec seuil (méthode C) montre des clusters vides/sans intérêts avec une plus grande fréquence. Sinon, doit-on l'admettre, les améliorations tentées au niveau des créations aléatoires ne sont pas concluantes. Probablement que l'effet est négligeable. Comme discuté, la relation de couplage de colonnes pour les mots réels est moins probable d'être générée par du hasard, même contrôlé. Il est tout de même sûrement prudent de dire qu'une quête d'améliorer nos résultats finaux devrait probablement passé par d'autres moyens que l'injection de savoir pour réduire l'alea excessif des méthodes ABD. La méthode C n'est pas vraiment intéressante non plus. La difficulté de gérer le nombre de centroïdes est également un problème évident. La solution implantée du seuil mobile ressemble davantage à une bricole qu'une véritable solution gagnante. L'approche est sûrement à éviter.

Discussion : Des méthodes alternatives modifiant les facteurs de rassemblement seraient sûrement beaucoup plus porteuses de variation dans les résultats finaux. L'altération mineure aux centroïdes semble être une piste plutôt sèche. Des modes de clustering basé sur la densité permettraient de réduire le 'bruit' des points qui devraient ne pas se retrouver dans un regroupement. Nos données semblent indiquer que ces points sont nombreux et donc cette approche serait sûrement à tester et évaluer.

Conclusion

On rétrospective, on peut affirmer que notre programme est efficace pour une utilisation basique du partitionnement des données à l'aide de centroïdes, mais qu'il serait bénéfique d'optimiser les méthodes de calculs et de bonifier la taille du corpus d'entraînement afin d'avoir des résultats plus flamboyants.

On a tout de même pu apprécier le potentiel immense d'un outil d'intelligence artificielle tout en étant conscient que ce projet n'était qu'une initiation à un domaine infiniment complexe qui n'est est encore qu'à son âge de pierre.