COMPARER LES METHODES DE CLUSTERING TEXTUEL : une étude empirique des algorithmes et des transformations préalables de l'espace des données

Alain Lelu[1], Martine Cadot[1, 2]

[1]Université de Franche-Comté (retraité) – alelu@orange.fr

[2]LORIA – martine.cadot@loria.fr

Abstract

A fair evaluation of text clustering methods needs to clarify the relations between 1) pre-processing, resulting in raw term occurrence vectors, 2) data transformation, and 3) method in the strict sense. We have tried to empirically compare a dozen well-known methods and variants in a protocol crossing three contrasted open-access corpora in a few tens transformed dataspaces. We compared the resulting clusterings to their supposed "ground-truth" classes by means of four usual indices. The results show both a confirmation of well-established implicit combinations, and good performances of unexpected ones, mostly in spectral or kernel dataspaces. The rich material resulting from these some 600 runs includes a wealth of intriguing facts, which needs further research on the specificities of text corpora in relation to methods and dataspaces.

Keywords: evaluation method; method comparison; text clustering; K-Means; Non-negative Matrix Factorization; Latent Dirichlet Allocation ; hierarchical clustering; linkage method; spectral clustering; graph partitioning; kernel clustering.

Résumé

Une évaluation juste des méthodes de clustering de textes se doit de clarifier les relations entre 1) le prétraitement, qui produit des vecteurs d'occurrences brutes de termes, 2) la transformation des données et 3) la méthode au sens strict. Nous avons essayé de comparer empiriquement une douzaine de méthodes et variantes bien connues dans un protocole faisant appel à trois corpus contrastés et d'accès ouvert dans quelques dizaines d'espaces de données transformés. Nous avons comparé les regroupements en résultant à leurs classes supposées de «vérité de terrain» au moyen de quatre indices habituels. Les résultats montrent à la fois une confirmation de combinaisons implicites bien établies, et de bonnes performances de combinaisons inattendues, principalement dans les espaces de données spectraux ou à noyau. Le riche matériel résultant de ces quelque 600 passages comprend une multitude de faits intriguants, qui nécessitent des recherches plus approfondies sur les spécificités des corpus textuels tant par rapport aux méthodes qu'aux espaces de données.

Mots clés : méthode d'évaluation ; comparaison de méthodes ; classification non supervisée de textes ; K-Moyennes ; Factorisation matricielle non-négative ; Allocation latente de Dirichlet; classification ascendante hiérarchique ; classification spectrale ; partition de graphes ; classification non supervisée à noyau.

1. Introduction :

Nous situant dans la problématique de la comparaison de méthodes de clustering spécifiquement sur des textes, les évaluations classiques, plus générales comme [Milligan & Cooper 87] et [Steinley 2006] ne nous sont pas d'un grand secours. Les écoles européennes d'analyse des données textuelles ont longtemps considéré que les méthodes de classification non supervisée ne pouvaient pas faire l'objet d'évaluation quantitatives, faute de classifications

de référence, de “vérité de terrain” indiscutable, compte tenu de la multiplicité de points de vue dont peut faire l'objet une collection de textes - stylistique, thématique (de quoi parlent les textes ?), sémantiques (que disent-ils de ces contenus, en parlent-ils en termes positifs, négatifs ou neutres ?). Les critères d'évaluation se basent plutôt sur l'ergonomie cognitive (représentation en arbre ? en classes strictes ? floues ?), sur la complexité numérique, ou sur la parenté formelle avec telle ou telle théorie de l'information. Et *in fine* sur les “effets de sens” qui ressortent de leur application à des données réelles, chaque domaine d'application ayant souvent ses méthodes préférées.

Le courant anglo-saxon dominant n'a pas eu ces scrupules, et s'est lancé d'emblée dans une démarche de pure ingénierie : quand le but est généralement de montrer qu'une méthode nouvelle est supérieure aux autres plus anciennes, la comparaison des valeurs d'indicateurs numériques s'impose. Un bon nombre d'indicateurs de similarité entre partitions de textes permettent de comparer la classification non-supervisée obtenue sur des ensembles de test, d'accès libre, à leur classification de référence établie à la main - celle-ci peut être issue de mots-clés attribués par des indexeurs, mais le plus souvent du regroupement de plusieurs collections de textes thématiquement homogènes, comme les résumés ou communications de conférences scientifiques de diverses sous-branches d'une branche plus vaste du savoir. Il s'agit donc d'appliquer la méthodologie d'évaluation de classifications supervisées à des classifications non supervisées. Cependant cette « vérité des nombres » peut paraître quelque peu illusoire : la combinatoire étendue induite par le choix des collections de textes et des indicateurs de comparaison limite la crédibilité de telles approches – on peut toujours soupçonner un auteur d'avoir choisi les ensembles de test et les indicateurs permettant de mettre en avant son algorithme. Sans parler des pré-traitements linguistiques d'extraction de termes à partir des textes, difficiles à spécifier exhaustivement.

Le groupe *Cluster benchmarking task force* de l'IFCS a décidé en 2019 de lancer un défi « au deuxième degré », le Neutral cluster benchmarking challenge consistant à faire concourir sur une méthodologie de comparaison d'algorithmes de clustering qui soit la plus neutre possible vis à vis des biais d'auteur, de pré- et post-traitement, dans l'optique d'une transparence et d'une reproductibilité totale des résultats (à partir d'ensembles de tests, de valeurs des paramètres, de codes des algorithmes en accès public). Nous limitant au seul clustering de corpus textuels, nous avons soumis la proposition [Lelu, Cadot, 2019] qui a remporté ce défi. Nous espérons que le présent article, qui résume et complète cette proposition, montrera que l'utilisation d'une méthodologie supervisée est susceptible d'ouvrir de nouvelles perspectives de recherche pour confronter le concept de classification aux questions de catégorisation humaine et algorithmique, et au pré-traitement des données au sens général : dans quel espace issu de la transformation des données les algorithmes sont-ils susceptibles d'approcher au mieux un processus humain de catégorisation ? Cette transformation et cet espace sont-ils universels ? Dépendent-ils des données, et si oui, comment ?

L'évaluation des méthodes de clustering de textes est l'un des principaux problèmes de la délimitation bibliométrique des domaines scientifiques. En tant que co-auteurs de [Zitt et al. 2019], nous avons essayé de tester dix-sept méthodes de clustering sur un ensemble de test accessible au public, celui de Reuters [Lewis et al. 2004], qui présente, entre autres difficultés, des classes de références, constituées à la main, de tailles fortement déséquilibrées (la «vérité de terrain » ciblée). Notre rapport est accessible en ligne [Cadot et al. 2018] en tant que matériel supplémentaire au chapitre du livre susmentionné. Un résultat inattendu a été que les méthodes de clustering anciennes, en particulier le clustering hiérarchique de Ward, ont donné de meilleurs résultats que de nombreuses méthodes plus récentes. Est-ce le cas pour tous les

types de corpus ? Surtout, nous avons compris que, pour des comparaisons justes ainsi que pour plus de clarté conceptuelle, nous devions clairement séparer les transformations des données brutes par comptage de mots (par exemple en représentation vectorielle pondérée “tf-idf” de Salton, ou espace spectral laplacien, etc.) des algorithmes au sens strict, plutôt que d'utiliser des combinaisons implicites acceptées de longue date. Par exemple, aucun argument rationnel n'interdit d'utiliser la factorisation matricielle non négative (NMF) dans un espace spectral. Enfin, des recommandations inattendues peuvent provenir de combinaisons non classiques. Cette clarification est l'un de nos fils conducteurs dans la présente recherche.

Bien que restreignant notre intérêt au clustering de textes, il est manifeste que les textes à traiter sont de nombreux types : résumés ou textes complets d'articles scientifiques, qui sont notre principal intérêt scientifique, ou textes journalistiques ou juridiques, ou issus du caractère social des communications sur Internet, comme les contributions aux discussions de forum ou les réseaux sociaux. Nous avons décidé de baser notre étude sur trois corpus de tests typiques et contrastés: une base de données scientifiques en texte intégral, un corpus de dépêches d'agence de presse, et un forum de discussion sur Internet. Il est clair que la chaîne complète de prétraitement du texte sort de l'objectif de notre recherche, nous devions donc reposer sur un même processus linguistique, ou faiblement linguistique, d'extraction de lemmes, de racines ou de termes, et sur la même élimination de mots peu fréquents ou trop fréquents. Nous explorons aussi l'influence de la troncature du vocabulaire en considérant des quantiles choisis de distribution du vocabulaire. Ceci alors que les études comparatives habituelles se contentent de mentionner un seuil d'occurrence absolu, point final... Toutes ces spécifications nous ont conduits aux choix que nous exposons dans la section méthodologie.

Bien sûr, nos choix de méthodes et de types d'espaces de données à considérer sont inévitablement arbitraires : nous avons essayé de prendre en compte les algorithmes, ou les familles de méthodes les plus courants, tels que les K-Means, la classification ascendante hiérarchique, le clustering spectral, le clustering de graphes, le clustering à noyau. Et nous avons ajouté deux méthodes plus spécifiques, à savoir la factorisation matricielle non négative (NMF) et l'allocation latente de Dirichlet (LDA), d’où une douzaine de méthodes et variantes.

Concernant les espaces de données, nous avons choisi d'ajouter au simple espace vectoriel d'occurrences de termes les espaces transformés par les schémas de pondération “tf-idf” de Salton et d'Okapi, par la métrique du chi-deux, par la décomposition spectrale laplacienne, par l'analyse des correspondances et enfin par l'expansion par noyau polynomial d'ordre 2. Ces transformations sont résumées dans la section suivante.

Compte tenu de la combinatoire des trois éléments principaux - types de texte, espaces de données et algorithmes - notre étude ne peut être qu'une exploration, fortement contrainte par les ressources disponibles. Cependant, quelques conclusions intéressantes seront tirées de cette exploration. Dans la conclusion finale, nous traiterons de ce qui peut être poursuivi et approfondi dans notre perspective, compte tenu des résultats.

Terminons cette introduction en disant combien nous sommes redevables à la remarquable initiative de l'équipe brésilienne LABIC [Rossi et al. 2013] qui ont prétraité de manière homogène [stemmer LABIC] une quarantaine de collections de textes et mis à disposition les matrices documents × termes en ligne sur leur site [données LABIC].

2. Methodologie

En ce qui concerne la question du biais d'auteur, notons que, bien qu'étant auteurs de quelques algorithmes de clustering (K-means Axiales, Analyse en composantes locales, Germen), nous avons exclu ces algorithmes de notre étude.

Un autre impératif fondamental que nous nous sommes fixé est la transparence et la reproductibilité : en plus du lien direct vers les matrices documents × termes que nous avons fourni ci-dessus, notre site HAL de matériel complémentaire [Lelu, Cadot 2019] comporte les liens vers les codes que nous avons utilisés. Bien que beaucoup d'algorithmes soient théoriquement insensibles à l'ordre des vecteurs d'entrée, dans la pratique, nous avons constaté que des effets d'ex-aequo, entre autres, pouvaient affecter les résultats. C'est pourquoi nous avons tiré au hasard l'ordre des vecteurs de données, par précaution.

2.1. Choix des corpus

Les trois corpus de test “prototypiques" mentionnés dans l'introduction sont, tout d'abord, le sous-ensemble "ModApté Split" de Reuters [Apté et al. 1994] limité aux huit classes les plus importantes ("Re8" dans la présente étude, 7674 documents, 8901 termes), deuxièmement, la collection ACM constituée des actes de quarante conférences dans différents domaines de l'informatique (3493 articles, 60 768 termes) , troisièmement, la collection "20 Newsgroups" ("Ng20") composée de 18 808 messages publiés dans vingt groupes Usenet (45 434 termes). La taille des classes de référence est fortement déséquilibrée dans le cas de Re8 (deux d'entre elles constituent 81% des documents), à peu près égale dans le cas d'ACM et de Ng20.

Il est à noter que seules les étiquettes de classe de Reuters sont issues d'une indexation manuelle directe. Les deux autres proviennent de la concaténation de sous-corpus de tailles comparables. ACM et Ng20 pourraient donc être considérés comme des "données semi-réelles", peu représentatives des corpus réels, ceux-ci non annotés par définition. Notons également que le coefficient "Silhouette" moyen [Kogan, 2007], qui mesure les contrastes inter-classes, est plus élevé (0,89) dans le cas de Re8 que dans Ng20 (0,80) et ACM (0,75).

2.2. Tronquer les vocabulaires

La taille des vocabulaires étant déséquilibrée (Re8: environ 8900 termes, ACM: 60 800 termes, Ng20: 45 000 termes) mais étendue (les hapax, c'est-à-dire les termes d'occurrence totale un, sont inclus dans ce décompte), nous avons décidé une procédure commune, à savoir une troncature du vocabulaire par seuils, indépendante de sa taille : en plus de l'option de base de conserver la matrice documents × mots avec l'ensemble du vocabulaire, nous avons construit deux matrices dérivées par corpus en ne conservant 1) que le troisième quartile de la distribution des termes (25% du total des occurrences) pour tronquer significativement le vocabulaire, et 2) le septième "octile" (12,5%) pour une très forte troncature.

2.3. Choix des methodes de clustering

Pour une première approche, nous n'avons pas pris en considération les algorithmes à deux paramètres (DBscan, Affinity Propagation, Smart Local Moving Algorithm), ni ceux à un seul paramètre qui montraient des résultats décevants sur le corpus de Reuters (Density Peaks, Independent Component Analysis, Fuzzy c-means, K-Means ++). Nous avons sélectionné :

- Premièrement, la méthode *K-means* simple ("KM") – cf. l'algorithme [Mac Queen 1967] (dans le cas adaptatif) et [Forgy 1965] (dans le cas itératif habituel) - initialisée en tirant au

hasard des vecteurs de données (nous avons eu une expérience peu convaincante de l'initialisation de type K-means++ dans le contexte du clustering de texte). Nous avons réalisé 20 passages élémentaires, ou «répliques», par passage, en sélectionnant la meilleure en termes d'optimum local de la fonction objectif des K-means (somme des carrés des distances euclidiennes intra-cluster) – ce qui ne garantit rien en termes d'optimum des critères externes de proximité à la partition de référence, mais donne des gages de stabilité et reproductibilité des résultats.

- La *classification ascendante hiérarchique* avec deux variantes de lien : lien moyen ("HCa") et Ward ("HCw") [Ward 1963]. Initialement de complexité temporelle $O(\#docs)^3$, les contributions plus récentes [Murtagh 1984] et [Müllner 2013] ont abaissé cette contrainte à $O(\#docs)^2$.

- La c*lassification spectrale* [Meila, Shi 2000] : nous avons utilisé la combinaison "standard" K-means dans l'espace spectral Laplacien, mais avons également exploré (avec succès, voir plus loin) de nombreuses autres combinaisons.

- Les *méthodes de partition de graphes* : nous avons choisi les deux plus largement reconnues, à savoir Louvain [Blondel et al. 2008] et InfoMap [Rosvall, Bergstrom 2007]. A noter que ces méthodes, contrairement à toutes les autres testées, n'ont pas besoin de fixer le nombre souhaité de clusters, d'où un avantage opérationnel majeur lorsqu'aucune idée du "vrai nombre de clusters" n'est connue à l'avance - le clustering hiérarchique étant dans une position intermédiaire, car en un seul passage il laisse à l'utilisateur le choix du niveau de coupure de l'arbre, donc du nombre de clusters.

- La *Factorisation non négative matricielle* ("NMF") [Lee, Seung 1999] : cette décomposition factorielle se transforme en une méthode de clustering lorsque l'étiquette d'un document est déterminée par le numéro d'axe de sa projection maximale. Comme cette méthode converge vers des optima locaux de sa fonction objectif, nous avons mis en œuvre la même stratégie «20 répliques» que pour les K-means. Il est à noter que la représentation graduée des documents qui en résulte dans les clusters est également intéressante dans la mesure où des valeurs aberrantes, voire des germes de clusters nouveaux, peuvent être identifiés.

- L'*Allocation Latente de Dirichlet* ("LDA") [Blei et al. 2003] est connue et respectée pour ses bases théoriques.

- Le *Clustering à noyau* [Girolami 2002] : grâce au "kernel trick", une matrice de similitude documents × documents ("Gram matrix") est construite sans nécessité d'étendre explicitement l'espace de données brut par une fonction noyau. Ici, nous avons utilisé un noyau polynomial d'ordre 2, ce qui revient à prendre en compte l'intégralité des couples de 2 termes ("2-itemsets"), quelle que soit leur position dans le texte, dans chaque document lors de la comparaison des vecteurs-documents - ceci en plus des "1-itemsets" habituels. Dans ce cas, l'espace de données brut documents × descripteurs n'est pas constitué de vecteurs d'occurrence numériques, mais de vecteurs binaires d'existence d'itemsets.

2.4. Choix des espaces de données

Nous renvoyons à notre texte [Lelu, Cadot, 2019] pour les formalismes mathématiques. En plus de l'espace vectoriel des simples occurrences de termes, nous avons construit:

- L'espace vectoriel de Salton, pondéré par le schéma tf-idf classique.

- L'espace vectoriel Okapi (également dit BM25) [Robertson et al. 1994], avec un système de pondération statistiquement mieux fondé.

- La métrique du chi-deux, qui équivaut à un espace vectoriel euclidien avec des vecteurs transformés comme suggéré dans [Legendre, Gallagher 2001].

- L'espace spectral laplacien [Von Luxburg 2007].

- L'espace spectral issu de l'analyse factorielle des correspondances [Benzecri 1973] [Greenacre 1984] [Lebart et al. 1998]. A noter que les distances euclidiennes dans cet espace factoriel complet sont égales aux distances du khi-deux [Benzécri 1973]. Par conséquent, tronquer cet espace revient à considérer des distances du khi-deux "partielles", a priori plus pertinentes que les distances du khi-deux. Nous vérifierons ce point plus loin.

- Espace noyau [Girolami 2002] : étant données les valeurs très contrastées de la matrice de Gram, la distance du cosinus est bien adaptée à cet espace de données.

Ces six transformations d'une matrice documents × termes conviennent aux méthodes de clustering KM, NMF, LDA et spectrales. D'autres méthodes, telles que le clustering hiérarchique, les méthodes de graphes et les méthodes à noyau, nécessitent une matrice de similarité (ou dissimilarité) documents × documents[1]. Nous avons dérivé les matrices de distance ou de cosinus pour tous les espaces de données mentionnés ci-dessus. En fonction de chaque combinaison espace de données + méthode, nous avons utilisé soit la distance euclidienne, soit la distance dite «cosinus» (en fait : 1-cosinus, c'est-à-dire la moitié de la corde au carré entre vecteurs-documents normalisés [Legendre, Gallagher 2001]).

2.5 Choix des mesures d'évaluation

Nous avons choisi les quatre indices de proximité entre partitions en clusters vs. en classes de référence les plus courants rencontrés dans la littérature d'évaluation, c'est-à-dire 1) l'information mutuelle normalisée (NMI) [Cover, Thomas 1991] et l'indice de Rand ajusté (ARI) [Rand 1971], qui se calculent quel que soit le nombre de clusters, 2), la moyenne des F-scores locaux cluster vs. classe [Van Rijsbergen 1979] et le score global de pureté (c'est-à-dire 1 - taux d'erreur global) qui nécessitent le même nombre de clusters et de classes, et les mêmes étiquettes. Nous avons pour cela aligné les k* classes “a priori” avec les k* clusters les plus "proches", au sens des F-scores locaux, au moyen de l'ordre issu du premier facteur non trivial de l'analyse de correspondance de la matrice clusters×classes des F-scores.

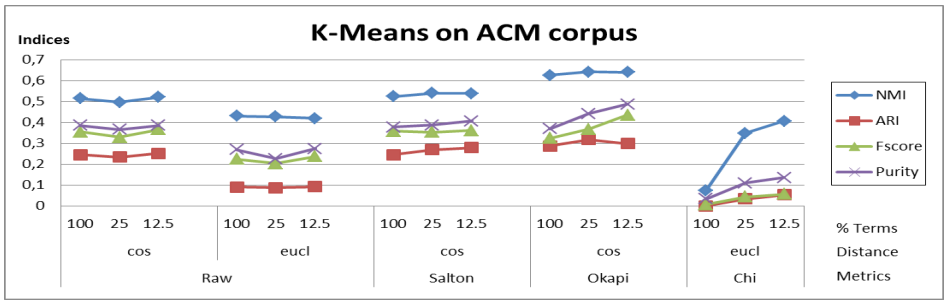
2.6 Codes utilisés et efficacité informatique

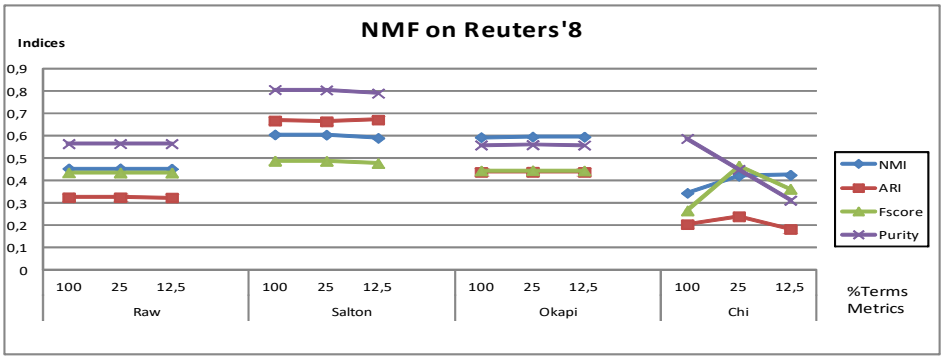
L'efficacité informatique n'étant pas notre objectif, nous avons programmé les transformations de données, les méthodes et le post-traitement dans un environnement Octave, sur un ordinateur Intel 6 cœurs I7, 3,33 GHz, 48 Go de RAM. Les codes des méthodes ont été dérivés de codes Matlab® existants (les liens vers les codes originaux sont disponibles sur notre site des matériaux supplémentaires [Lelu, Cadot 2019]). Leur degré d'optimisation du temps de calcul varie considérablement : par exemple dans le cas du jeu de test Ng20 de 19 000 documents, depuis 2 minutes pour vingt passages du code standard "litekmeans.m" (qui lui-même comporte 20 répliques élémentaires de l'algorithme), à 6 h pour un passage de la méthode Louvain, … et 24 h pour la classification hiérarchique avec lien moyen.

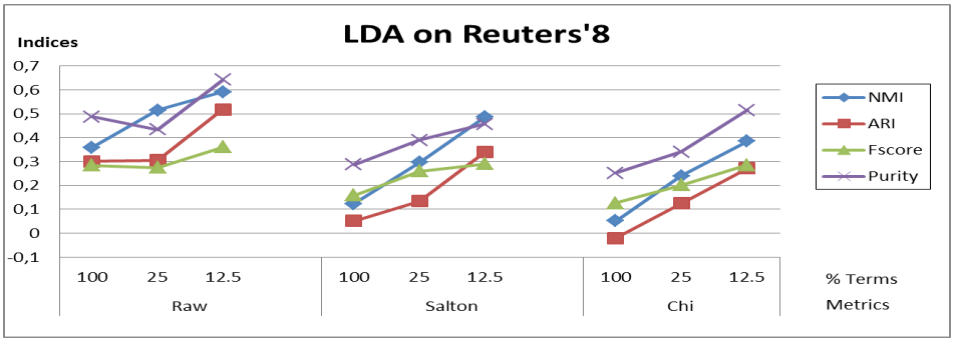
3. Résultats de l'évaluation pour chaque méthode

Nous renvoyons à [Lelu, Cadot, 2019] pour l'intégralité des 592 résultats de passages, sous forme de graphiques. Nous ne montrerons ici que quelques extraits caractéristiques.

[1] Dans certains cas, le recalcul de la matrice des distances chaque fois que nécessaire peut optimiser l’exécution.







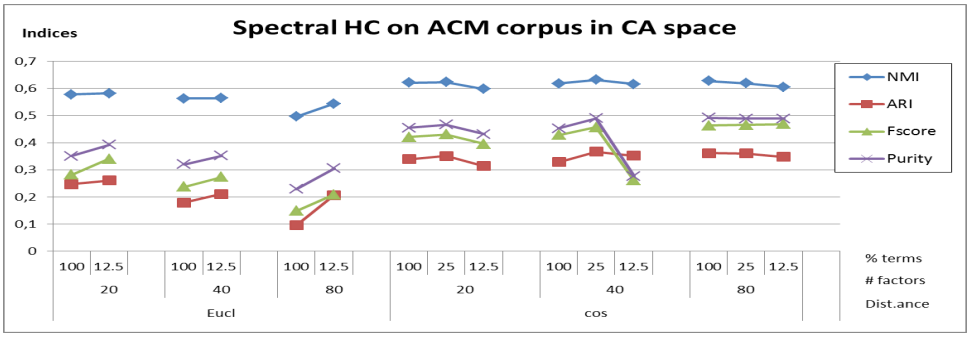
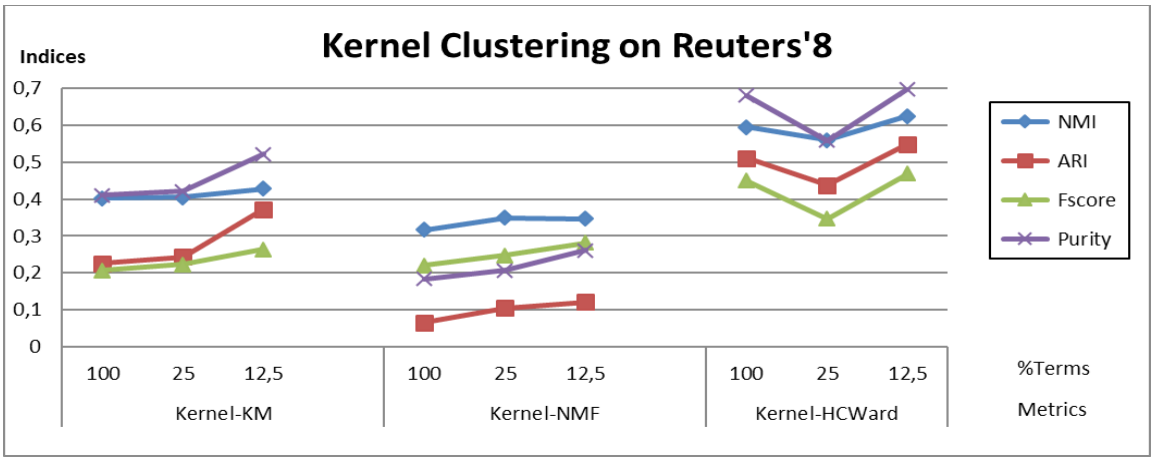


Figure 1 : exemples de K-Moyennes, de Factorisation matricielle non-négative, d'Allocation latente de Dirichlet, et de Classification hiérarchique (lien de Ward) dans l'espace AFC. Effets du seuillage du vocabulaire (% des termes retenus, les plus fréquents), de la métrique utilisée, et dans le cas du clustering spectral, du nombre de facteurs retenus.

Beaucoup de constats intriguent. Un parmi d'autres : l'effet de la troncature du vocabulaire ou des cosinus peut augmenter ou diminuer les performances, voire être nul, selon la méthode utilisée, ou même selon le paramétrage utilisé pour une même méthode !



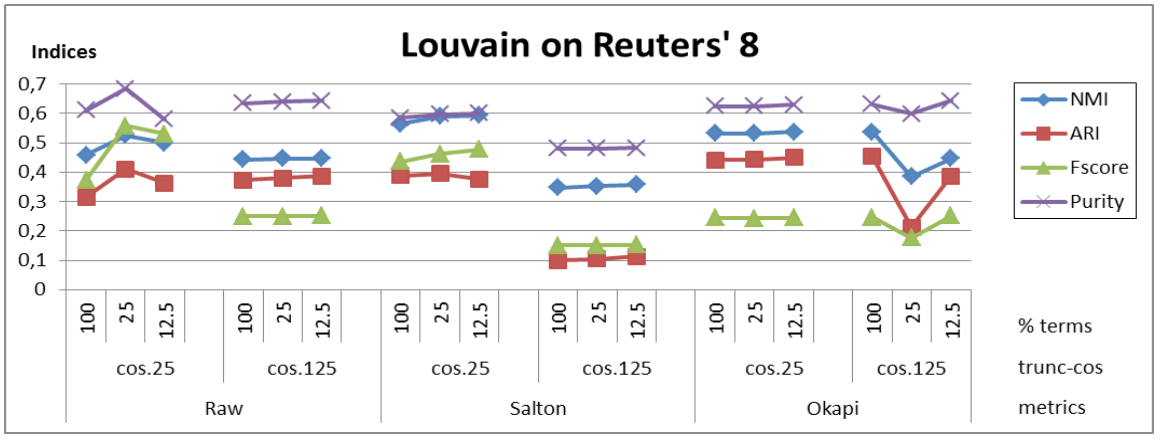


Figure 2 : exemples de Classification à noyau polynomial et de méthode de Louvain de partition de graphes. Effets de la métrique, du seuillage du vocabulaire, et dans le cas de la partition de graphe, du seuillage des cosinus inter-documents.

4. - Observations principales et synthèse

Concentrons-nous d'abord sur les outils de mesure: nous pouvons observer que dans le cas des deux corpus "équilibrés", les quatre indices d'évaluation se comportent de manière très parallèle et ordonnée - voir le 1er graphique de la fig.1 En revanche, ce parallélisme et ce classement régulier se détériorent dans le corpus déséquilibré de Reuters8, et dans une moindre mesure lorsque des méthodes hiérarchiques sont utilisées. Les scores F et de pureté peuvent (voir le 2e graphique de la fig.1) ou pas (voir le 4e graphique de la fig.1) présenter un comportement quelque peu contradictoire ou non monotone. Une étude approfondie pourrait peut-être expliquer ces divergences intéressantes, mais est clairement hors de nos objectifs actuels. Nous avons donc choisi l'indice NMI, le plus stable, comme mesure de référence pour classer les passages de chaque corpus (ACM: 246 passages, Re8: 237 passages, Ng20: 109 passages, soit un total de 592 passages).

4.1. «Top 3» des combinaisons méthode-espace des données gagnantes pour chaque corpus

Tableau 1 : en ligne 1 chaque corpus avec ses nombres de classes, documents et termes ; lignes suivantes : les 3 meilleures combinaisons méthode/espace des données pour chacun, une par colonne.

Corpus	*ACM (k=40)*	*3493 docs*	*60768 terms*	*Reuters'8 (k=8)*	*7674 docs*	*8901 terms*	*NewsGroups (k=20)*	*18808 docs*	*45434 terms*
Top methods	**1: HC-Ward**	**2: HC-Ward**	**3: K-Means**	**1: K-Means**	**2: HC-averag**	**3: HC-Ward**	**1:NMF**	**2: HC-Ward**	**3: K-Means**
cos thresh./eucl	cos 12.5% to 100%	cos 100%	eucl	cos 100%	cos 100%	cos 100%	eucl	cos 100%	cos 100%
Dataspace type	**Spectral Lapl.**	**Standard**	**Spectral Lapl.**	**Standard**	**Spectral CA**	**Kernel**	**Standard**	**Spectral CA**	**Standard**
#factor or order	40; 80 (=k; 2k)	40 (=k)	80 (=2k)	8 (=k)	8 (=k)	Poly order-2	20 (=k)	40 (=2k)	20 (=k)
% vocabulary	100%; 12.5%	100.%	12.5%	100%; 12.5%	12.5%	12.5%	100.%	100.%	12.5%; 100%; 25%
Metrics	Salton	Okapi	Okapi	Okapi; Salton	Raw	Raw	Okapi	Raw	Salton
NMI	.6980 to .6714	.6739	.6712	.6461 to .6314	.629	.625	.6252	.6220	.621
Computing time	77s	75s	6s	10s	24h	800s	107s	4h	150s

Ces combinaisons optimales dépendent clairement des corpus. Une grande variété de transformations d'espace des données (vocabulaire tronqué ou non, espace de données de Salton, Okapi ou brut, espace noyau ou spectral laplacien, ...) et de méthodes (HC-Ward, K-Means, NMF...) sont présentes, surtout quand on étend la sélection aux “top 50”. On peut noter que quatre méthodes sur neuf peuvent être considérées comme "classiques" : K-Means dans l'espace pondéré Salton ou Okapi, Clustering Hierarchique standard dans l'espace pondéré Okapi sans seuillage du vocabulaire, et Factorisation matricielle non-negative pondérée également Okapi sans seuillage.

4.2. Limites

Il est intéressant de se rendre compte de ce que signifient dans le monde réel les valeurs d'indice maximales, en retournant aux données. Le tableau de correspondance entre les classes et la meilleure partition en clusters pour le corpus Ng20 (voir figure 3) montre une structure en "serpent", plus qu'une diagonale : certains clusters correspondent totalement ou partiellement à plusieurs classes, et inversement. Les classes N° 3 et N° 9 sont dispersées au milieu de nombreux clusters. Le NMI global de 0,62 et la pureté de 0,57 ne sont donc pas si bons et cela est confirmé par le F-score moyen de 0,53.

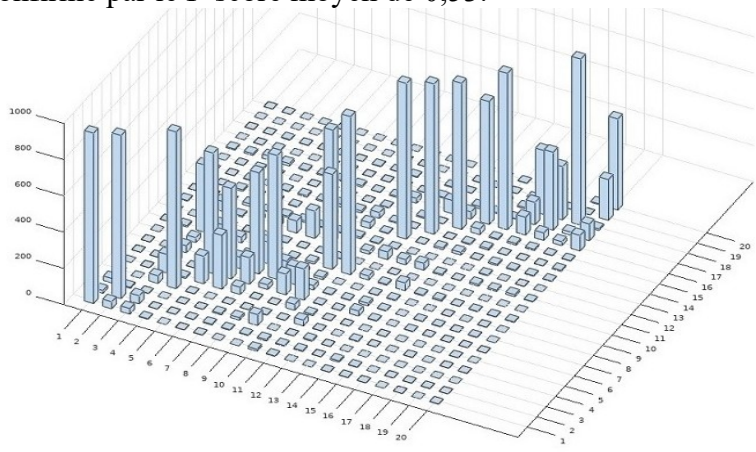


Figure 3 - Croisement des clusters et des classes sur le meilleur résultat NMI du corpus 20 NewsGroups. A gauche: classes; à droite: clusters. Verticalement : nombre de documents.

Les raisons de la divergence entre les catégories humaines et les clusters pourraient être approfondies avec des experts du domaine d'application, en examinant par exemple le cas des classes 3 et 9. Ce processus peut-il converger vers une "vérité de terrain" consensuelle, ou

diverger, montrant les limites de la seule information textuelle, ou les limites d'une extraction de termes faiblement linguistique ? Pour revenir aux figures de la section précédente, la régularité visuelle de nombreuses observations non intuitives montre que des phénomènes et des interactions non clairement élucidés sont à l'œuvre : par exemple quel rôle pourrait être joué par l'inégalité de taille et le nombre des clusters, par la taille du vocabulaire, par le type et le style des textes - en fonction des multiples points de vue classificatoires selon lesquels on peut les comparer -, par les contrastes de densité dans les espaces de données, etc. ? Cela ouvre un champ de recherche inexploré à notre connaissance.

4.1. Points communs

Ceci étant dit, des points communs apparaissent, au niveau local de chaque corpus comme au niveau global. En examinant le “top 50” des passages de chaque corpus (classés selon les valeurs décroissantes de NMI), quelques comportements communs émergent.

Points communs partiels:

- Pour le corpus ACM : la combinaison HC-Ward dans toutes les variantes de l'espace Laplacien Okapi domine massivement, les KMeans spectrales dans les mêmes espaces s'intercalant 9 fois, HC-Ward standard et NMF standard le faisant resp. 2 fois et 3 fois.

- Pour le corpus Re8 : les K-Means standard pondérées Okapi ou Salton dominent, suivies de spectral HC-average dans l'espace factoriel AFC, puis de spectral K-Means dans l'espace factoriel AFC, parfois Laplacien ; 3 HC-Ward à noyau s'intercalent, ainsi que 6 NMF standard et 3 Louvain dans l'espace Salton.

- Pour le corpus Ng20 : les K-Means standard occupent le haut du tableau, ainsi que NMF Okapi et spectral HC-Ward dans l'espace AFC. Suivent les K-Means spectrales pondérées Salton (parfois Okapi) dans l'espace AFC et HC-Ward spectral dans le même espace. HC-Ward à noyau ferme la marche.

Points communs globaux

Pour nous faire une idée sur les comparaisons inter-corpus, nous avons construit un indicateur relatif de performance en rapportant le NMI d'une combinaison donnée “ espace de données + méthode” pour un certain corpus au NMI maximum constaté sur ce corpus. Disposant de trois valeurs par combinaison, on peut situer – à titre heuristique plutôt que rigoureux – les performances globales d'une combinaison en calculant la moyenne et l'écart maximum entre ces trois valeurs. Les lignes qui suivent résument ce processus pour les trois combinaisons qui nous ont paru réaliser le meilleur compromis entre performances et indépendance par rapport au corpus (c'est à dire écart faible), et que nous recommandons :

1) *NMF standard sur données pondérées Okapi, avec vocabulaire non tronqué* :

NMI rel. : 95,4%, écart : 7,9%

2) *Clustering hiérarchique (lien Ward) spectral dans l'espace des 2k* premiers facteurs AFC, avec une troncature forte du vocabulaire* (12,5% du vocabulaire d'origine) :

NMI rel. : 92,0%, écart : 9,1%

3) *K-Means standard sur données pondérées Okapi, et vocabulaire tronqué aussi à 12,5%* :
NMI rel. : 91,1%, écart : 18,2%

Le principal problème pour suivre la recommandation N°2 est de construire le ou les espaces spectraux pour des données de grande taille. Mais dans de nombreux langages informatiques, il existe des procédures efficaces de décomposition aux valeurs singulières de matrices clairsemées (“sparse”), appropriées lorsque le problème consiste à extraire un nombre limité de valeurs propres et de vecteurs propres principaux à partir de très grands tableaux de données, ce qui est le cas dans la présente étude. Sinon, des coprocesseurs graphiques parallèles peuvent être dédiés à cette tâche.

5. Conclusions et perspectives

Nous espérons avoir apporté quelques éclaircissements sur le problème de l'évaluation des procédures de clustering de texte, en considérant séparément les algorithmes et les espaces de données dans lesquels ils opèrent. Nous avons réalisé quelque six cent exécutions d'une douzaine d'algorithmes et de variantes, dans quelques dizaines d'espaces de données différents, sur trois corpus de test prototypiques et d'accès public. Nous avons mis au jour une variété inattendue de combinaisons optimales de méthodes et d'espaces de données, dont nous recommandons prudemment les trois meilleures. La variété des transformations et des paramètrages possibles nécessite un effort à poursuivre considérable pour améliorer notre compréhension et notre maîtrise des processus de catégorisation artificielle vs. humaine. Nous espérons que cette étude empirique contribuera à un tel défi. Dans une première étape modeste, nous explorerons l'influence du prétraitement linguistique : choix ou élimination des catégories de mots, comparaison entre la prise en compte des expressions multi-mots et l'utilisation du kernel clustering, c’est à dire l’expansion exhaustive des unitermes par noyau polynomial.

Références

Apté C., Damerau F. and Weiss S.M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, vol.(12.3) : 233-251.

Benzécri J.-P. (1973). *L’analyse des correspondances, Analyse des données*, *Vol.2*, Dunod, Paris

Blei D.M., Ng A.Y. and Jordan M.I. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research,* vol.(3): 993-1022

Blondel V.D., Guillaume J.-L., Lambiotte R. and Lefebvre E. (2008). Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, vol.(10): P10008

Cadot M., Lelu A. and Zitt M. (2018). Benchmarking seventeen clustering methods on a text dataset. (Research Report) LORIA. 〈hal–01532894v6〉

Choi S.-S., Cha S.-H., and Tappert C.C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, vol.*(* 8.1): 43-48.

Cover T.M. and Thomas J-A. (1991). Entropy, relative entropy and mutual information. *Elements of information theory*, vol.(2): 1-55.

Forgy E. (1965): Cluster analysis of multivariate data: efficiency vs. interpretability of classifications, Abstract. *Biometrics*, vol.(21): 768-769.

Girolami M. (2002). Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, vol.(13.3): 780-784

Greenacre M.J. (1984). *Theory and applications of correspondence analysis.* Academic Press

Kogan J. (2007). *Introduction to Clustering Large and High-Dimensional Data*. Cambridge University Press.

[LABIC data] http://sites.labic.icmc.usp.br/text_collections/; accédé le 25/3/2020

[LABIC stemmer] http://sites.labic.icmc.usp.br/tpt/; accédé le 25/3/2020

Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data.* Kluwer Academic Publisher, Dordrecht, Boston,

Lee D.D. and Seung H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* vol.(401.6755) : 788-791.

Legendre P. and Gallagher E.D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia* vol(129): 271. https://doi.org/10.1007/s004420100716

Lelu A. and Cadot M. (2019). Evaluation of text clustering methods and their dataspace embeddings: an exploration. <hal-02116493v4>

Lewis D.D., Yang Y., Rose T.G. and Li F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, vol.(5-Apr): 361-397. http://www.davidlewis.com/resources/testcollections/rcv1/

Mac Queen J. (1967): Some methods for Classification and Analysis of Multivariate Observations, *Proc. of 5th Berkeley Symp. Math. Stat. Proba.*, pp. 281-297.

Meila M. and Shi J. (2000). Learning Segmentation by Random Walks. In Todd K.L., Thomas G.D. and Volker T. editors, *Proc. of NIPS'00 (Denver, CO).*

Milligan, G. W., & Cooper, M. C. (1987). Methodology Review: Clustering Methods. *Applied Psychological Measurement*, 11(4), 329–354.

Daniel Müllner (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python, *Journal of Statistical Software* vol.(53.9): 1-18.

Murtagh F.. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, Vol.(26.4): 354-359.

Murtagh F. (1984). Complexities of Hierarchic Clustering Algorithms: the state of the art. *Computational Statistics Quarterly*, vol.(1): 101-113.

Murtagh F. and Legendre P.. (2014). Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative. *Algorithm Journal of Classification*, vol.(31.3): 274-295.

Rand W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, vol.(66.336): 846-850.

Robertson S., Walker S., Hancock-Beaulieu M. and Gatford, M. (1994). Okapi and TREC-2. In Harman D.K. (Ed.) *Proc. of the Third Text REtrieval Conference*: 21-34.

Rossi, Marcacini, Oliveira and Rezende (2013). Benchmarking Text Collections for Classification and Clustering Tasks No 395 ICMC TECHNICAL REPORT. São Carlos, SP, Brazil.

Rosvall M., Bergstrom C.T. (2007). An information-theoretic framework for resolving community structure in complex networks, *Proc. of the Nat. Academy of Sciences*, vol(104.18): 7327-7331.

Steinley, D. (2006). Profiling local optima in K-means clustering: Developing a diagnostic technique. *Psychological Methods*, 11(2), 178–192

Van Mechelen I., Boulesteix A.-L., Dangl R., Dean N., Guyon I., Hennig C., Leisch F. and Steinley D. (2018). Benchmarking in cluster analysis: A white paper. arXiv:1809.10496v2 [stat.OT].

Van Rijsbergen C.J. (1979). *Information Retrieval (2nd ed.)*. Butterworth-Heinemann.

Von Luxburg U. (2007). A tutorial on spectral clustering. *Statistics and computing* vol.(17.4): 395-416.

Ward J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*. Vol.(58.301): 236-244.

Zitt M., Lelu A., Cadot M. and Cabanac G. (2019). Bibliometric delineation of scientific fields. In Glänzel W., Moed H.F., Schmoch U. and Thelwall M., editors, *Handbook of Science and Technology Indicators*, Springer International Publishing.