Qui se ressemblent s'assemblent

Des méthodes algorithmiques de regroupement de données permettent d'analyser des textes, de les classifier, et parfois même d'en identifier les auteurs.

Christian Genest
Université McGill
Jean-François Plante
HEC Montréal
Ostap Okhrin
Technische Universität Dresden

Grâce aux progrès technologiques, on peut dorénavant recueillir des masses de données à coût modique. Pour qu'elles soient informatives et utiles à la prise de décision, ces données doivent toutefois être traitées et analysées. Les méthodes statistiques jouent un rôle de tout premier plan dans ce processus de valorisation. Que ce soit dans la gestion de l'état, la recherche fondamentale ou les affaires, les informations issues de l'analyse statistique soutiennent une prise de décision objective et... éclairée. Elles permettent aussi parfois de découvrir des phénomènes insoupçonnés.

Nous allons explorer ici une toute petite partie des rouages mathématiques de cette vaste entreprise en décrivant une technique de regroupement hiérarchique souvent utilisée, entre autres, pour l'analyse exploratoire de corpus textuels. Dans ce contexte, chaque observation se présente sous la forme d'un vecteur dont les composantes nous renseignent sur la fréquence relative de certains mots ou groupes de mots de même racine dans une production écrite, qu'il s'agisse d'un livre, d'un discours ou d'un gazouillis. Comme nous le verrons, le choix des mots peut s'avérer fort révélateur !

Considérons à titre d'exemple les 12 ouvrages de Tolkien cités dans l'encadré en bas de page. À l'aide du progiciel de forage de données tm (pour « text mining ») intégré au gratuiciel R, nous avons compté le nombre de mots dans chacune de ces œuvres (en version originale anglaise). Nous avons ensuite déterminé, pour chaque ouvrage $i \in \{1, \ldots, 12\}$, la proportion g_i de mots appartenant à l'ensemble $G = \{get, gets, got, getting\}$. Nous avons aussi calculé la proportion s_i de mots de l'ouvrage i provenant de l'ensemble $S = \{say, says, said, saying\}$. Pour illustrer la méthode de calcul de g_i et de s_i, appliquons-la au texte suivant, qui compte 37 mots :

Twitter is a great tool for learning effective communication. You must say a lot in 140 characters and you typically get feedback. If you are not getting any, or people misunderstand what you were saying, try again[1].

Puisque *get*, *getting*, *say* et *saying* y apparaissent chacun une fois, on trouve $g_i = s_i = 2/37$.

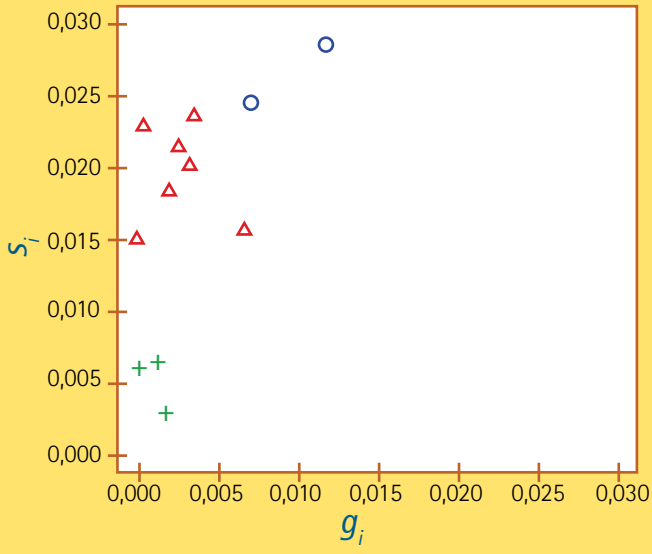
Titres des ouvrages analysés

1. Le Hobbit (1937)
2. La Fraternité de l'Anneau (1954)
3. Les Deux Tours (1954)
4. Le Retour du Roi (1955)
5. Les Aventures de Tom Bombadil (1962)
6. Smith de Grand Wootton (1967)
7. Feuille, de Niggle (1945)
8. Le Fermier Gilles de Ham (1949)
9. Le Retour de Beorhtnoth, fils de Beorhthelm (1953)
10. Le Silmarillion (1977)
11. Contes et Légendes Inachevés (1980)
12. Les Enfants de Húrin (2007)

1. *Twitter est un excellent outil pour apprendre à communiquer de façon efficace. Vous devez dire beaucoup en 140 caractères et on réagit généralement à vos propos. Si personne ne réagit, ou si les gens comprennent mal votre message, essayez à nouveau.*

Le processus est le même pour chacun des 12 livres de Tolkien, mais il serait long et fastidieux à réaliser à la main. Grâce au progiciel tm, il est pourtant complété en une fraction de seconde. Au vu des résultats, illustrés à la figure 1, il semble se détacher au moins deux groupes de points, sinon trois. En effet, les trois ouvrages identifiés par une croix verte (dont l'identité est précisée plus loin) présentent tous une très faible ordonnée. Autrement dit, les mots de l'ensemble S y sont relativement peu utilisés. À l'inverse, on observe une haute fréquence des mots des ensembles G et S dans les deux ouvrages identifiés par des cercles. Y a-t-il une raison fondamentale pour laquelle ces groupes de livres se distinguent par leur usage relatif de mots aussi fréquents et anodins que *get*, *say* et leurs variantes ? Peut-être que oui, mais nous n'avons aucune raison de le soupçonner a priori. Si nous avons choisi cet exemple, c'est seulement pour pouvoir illustrer dans un cas simple une méthode qui permet d'opérer automatiquement les regroupements et d'objectiver, dans une certaine mesure, les résultats. Nous verrons ensuite comment cette approche peut s'avérer très révélatrice lorsqu'elle est appliquée non pas seulement à deux groupes de mots, mais à plusieurs groupes simultanément. Comme il ne sera alors plus possible de visualiser le nuage de points, il importe de bien comprendre le principe de la méthode, si on veut pouvoir se fier aux résultats !

Figure 1 :
Fréquences relatives des groupes de mots G et S observées dans douze œuvres de J.J.R. Tolkien



La regroupement hiérarchique

Les algorithmes de regroupement visent à systématiser la formation de classes à partir d'un ensemble de vecteurs :

$$x_1, \ldots, x_n \in \mathbb{R}^d.$$

Il en existe de toutes sortes, mais ils s'inspirent tous du dicton « Qui se ressemblent s'assemblent ». L'objectif est de regrouper les points en classes homogènes et bien différenciées. Pour ce faire, on doit d'abord se doter d'une mesure de similarité ou de proximité entre deux vecteurs :

$$a = (a_1, \ldots, a_d),\ b = (b_1, \ldots, b_d) \in \mathbb{R}^d.$$

Deux mesures fréquemment employées sont la *distance de Manhattan*, définie par :

$$d(a,b) = \sum_{i=1}^{d} |a_i - b_i|$$

et le carré de la distance euclidienne, soit :

$$d(a,b) = \sum_{i=1}^{d} (a_i - b_i)^2.$$

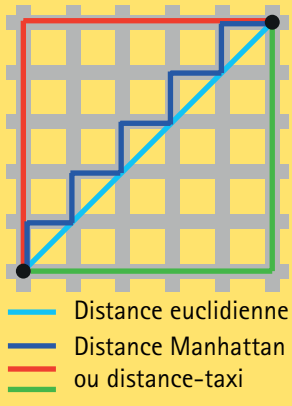
Aux fins de regroupement, il est important que $d(a,a) = 0$ et $d(a,b) = d(b,a)$ mais il n'est pas nécessaire que soit vérifiée l'inégalité du triangle, à savoir :

$$d(a,c) \leq d(a,b) + d(b,c),$$
pour tous $a, b, c \in \mathbb{R}^d$.

Dans l'approche agglomérative hiérarchique sur laquelle nous allons nous pencher, chaque singleton $\{x_i\}$ constitue au départ une classe en soi. On compare ensuite les paires (x_i, x_j) deux à deux afin d'identifier celle pour laquelle la valeur de $d(x_i, x_j)$ est la plus petite.

La distance de Manhattan

Appelée aussi *taxi-distance*, la *mesure de Manhattan* est la distance parcourue par un taxi lorsqu'il se déplace entre deux points dans une ville où les rues sont agencées selon un quadrillage.



On doit donc faire $\binom{n}{2} = n(n-1)/2$ comparaisons. Si par exemple x_1 et x_2 sont les deux points les plus proches, on fusionne les classes $\{x_1\}$ et $\{x_2\}$ et on dénote $\Delta_1 = d(x_1, x_2)$. La valeur Δ_1 est donc le degré de proximité auquel un premier regroupement se produit. Si $n = 2$, c'est dire qu'il n'y avait au départ que deux vecteurs à regrouper. Comme c'est maintenant fait, la tâche est complétée. Sinon, il reste $n-2$ singletons et une classe constituée de deux vecteurs.

Avant de pouvoir procéder à un nouveau regroupement, il faut alors se demander comment on devrait mesurer la proximité entre un point (un singleton) et la nouvelle classe qui comporte deux éléments. Il se peut en effet qu'à l'étape 2, on choisisse encore de regrouper deux singletons, mais il se peut aussi qu'on préfère fusionner un singleton à la classe comportant déjà deux éléments, parce que ce sont ceux qui se ressemblent le plus. À la fin de cette seconde étape, il y aurait donc deux cas de figure possibles : soit on a deux classes de deux membres et $n-4$ singletons, soit on a une classe de trois membres et $n-3$ singletons. On entend ensuite poursuivre la démarche de façon itérative, de sorte qu'au bout de $n-1$ étapes, tous les points se retrouvent dans la même classe.

De façon plus générale, il faut donc en fait s'entendre sur la façon de calculer le degré de proximité entre deux classes Y et Z de dimensions n_Y et n_Z arbitraires, car on sera éventuellement confronté à ce problème à force de répéter la procédure d'agglomération. Notons

$$Y = \{y_1, \ldots, y_{n_Y}\} \text{ et } Z = \{z_1, \ldots, z_{n_Z}\}$$

ces deux ensembles de vecteurs dans $\mathbb{R}^d$. L'approche la plus usitée consiste à poser la définition ci-dessous :

$$d(Y,Z) = \frac{n_Y n_Z}{n_Y + n_Z}\left\{\frac{2}{n_Y n_Z}\sum_{i=1}^{n_Y}\sum_{j=1}^{n_Z} d(y_i, z_j) - \frac{1}{n_Y^2}\sum_{i=1}^{n_Y}\sum_{j=1}^{n_Y} d(y_i, y_j) - \frac{1}{n_Z^2}\sum_{i=1}^{n_Z}\sum_{j=1}^{n_Z} d(z_i, z_j)\right\}. \quad (1)$$

On note d'abord que dans le cas où $Y = \{a\}$ et $Z = \{b\}$, la formule (1) se réduit à $d(a, b)$, ce qui justifie l'emploi du même symbole pour désigner la proximité entre deux classes. Dans sa forme générale, la formule comporte trois termes : le premier mesure la proximité entre les éléments du groupe Y et ceux du groupe Z ; les deux autres termes mesurent le degré de similarité au sein de chacun des deux groupes. Il n'est pas évident a priori qu'on a toujours $d(Y,Z) \geq 0$ mais c'est le cas pour la distance de Manhattan et le carré de la distance euclidienne.

Par construction, la fusion des groupes Y et Z minimisant $d(Y,Z)$ rencontre l'objectif visé : regrouper les observations les plus semblables. Au terme de l'étape ℓ de l'algorithme, les observations initiales sont regroupées en $n-\ell$ classes, disons $C_1, \ldots, C_{n-\ell}$. On doit alors mesurer l'écart entre toutes les paires possibles de classes pour déterminer lesquelles seront fusionnées. C'est ici que la formule (1) révèle son plus grand atout : elle permet de calculer la proximité entre $C_i \cup C_j$ et C_k en fonction des écarts déjà connus entre ces classes. Au prix de calculs algébriques assez longs mais simples, on voit que :

$$d(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k} d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j).$$

où n_i, n_j et n_k sont les tailles respectives des classes C_i, C_j et C_k. Cette formule récursive est à l'origine de l'algorithme de Lance et Williams, qui permet de réduire considérablement le temps de calcul. Ceci s'avère crucial lorsqu'on veut analyser non pas 12 livres mais plutôt les gazouillis de millions d'abonnés de Twitter pour identifier des communautés ou des groupes d'intérêt. La complexité des calculs nécessaires à la classification de n individus par l'algorithme de Lance-Williams est de l'ordre de n^2, tout comme la capacité de stockage requise.

Représentation graphique

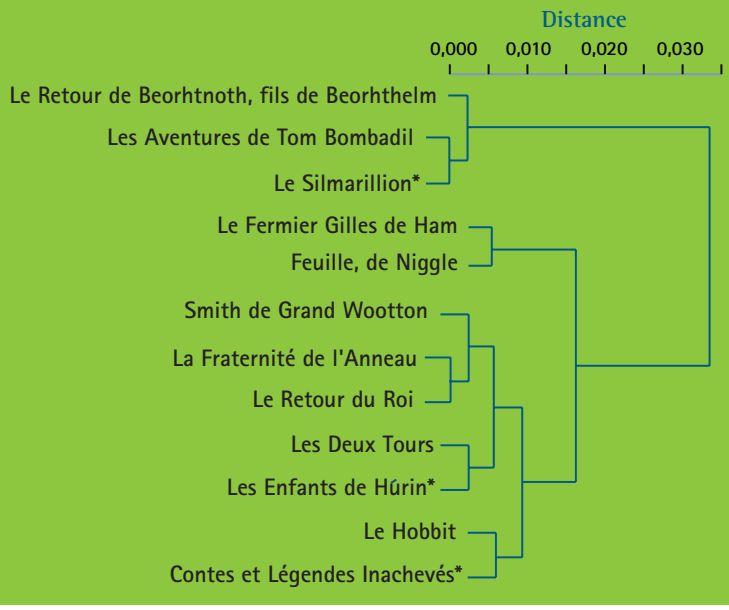
On a déjà vu plus haut comment est défini le premier seuil de regroupement, dénoté Δ_1. Les seuils des $n-2$ regroupements suivants sont définis de la même manière au moyen de la formule (1). On peut montrer qu'on a nécessairement $\Delta_1 < \cdots < \Delta_{n-1}$ lorsque pour tous groupes A, B, C mutuellement disjoints, on a $d(A,C) \leq \max\{d(A,B), d(B,C)\}$. Cette propriété dite de monotonie est vérifiée, entre autres, pour la distance de Manhattan et le carré de la distance euclidienne.

Une fois muni des définitions de classes et des seuils $\Delta_1, \ldots, \Delta_{n-1}$ auxquels se produisent les regroupements, on peut construire une représentation graphique des résultats à l'aide d'un dendrogramme.

Un dendrogramme est une figure arborescente qui permet de visualiser le processus agglomératif de formation des classes. La figure 2 illustre le résultat de cette opération pour les données de la figure 1. Chacune des $n = 12$ lignes horizontales apparaissant à gauche du diagramme correspond à l'un des ouvrages de Tolkien mentionnés plus haut. Les $n - 1 = 11$ lignes verticales sont tracées aux hauteurs $\Delta_1 < \cdots < \Delta_{n-1}$ auxquelles deux ouvrages ou deux groupes d'ouvrages ont été successivement fusionnés.

La figure 2 permet de retracer le processus d'agrégation des points de la figure 1 dont la proximité est mesurée par le carré de la distance euclidienne : le résultat pourrait être différent si on employait la distance de Manhattan. Les trois premiers ouvrages (à partir du haut) se distinguent : ils sont relativement distants des neuf autres, auxquels leur classe n'est réunie qu'à la toute dernière étape, à une distance d'environ 0,035. Ils correspondent aux trois points qui avaient été

Figure 2 : Dendrogramme de regroupement des points de la figure 1 par le carré de la distance euclidienne



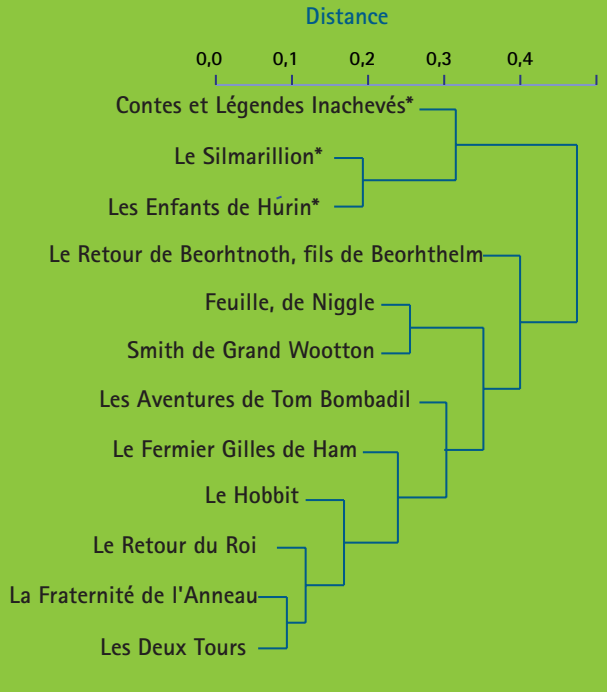
identifiés par des croix vertes dans la figure 1. Parmi les neuf autres livres se dégage une paire formée des 4e et 5e titres, qui se fusionne aux sept autres à une distance d'environ 0,017. Vous avez deviné : il s'agit des points identifiés par des cercles sur la figure 1. L'algorithme permet donc d'objectiver ce que nous avions perçu à l'œil. Toutefois, le résultat n'est pas intéressant en soi, à moins d'avoir des raisons de penser que les ouvrages se distinguent par l'emploi qu'on y fait des mots *get* ou *say* et de leurs variantes grammaticales. Maintenant que nous maîtrisons le principe, nous allons décrire une application de grande envergure qui sera beaucoup plus instructive.

Une analyse plus ambitieuse

En faisant appel au progiciel tm, nous avons calculé la fréquence de tous les mots apparaissant dans les 12 ouvrages de Tolkien et nous avons sélectionné les 250 plus fréquents d'entre eux. À cette fin, nous avons toutefois exclu tous les noms de personnages (Bilbo, Frodo, Gandalf, etc.), ainsi qu'une série de marqueurs de relation et autres « mots vides » qui sont considérés comme non informatifs par les spécialistes de l'analyse de données textuelles. Pour l'anglais, la commande stopwords de tm identifie 174 tels mots, dont *is*, *are*, *be* et *me*, *myself*, *and*, *I*,... La liste correspondante pour le français est légèrement plus courte ; elle en compte 164.

Les singletons qui constituent nos 12 classes de départ ne sont donc plus des paires de fréquences relatives dans $[0,1]^2$ mais plutôt des vecteurs de longueur 250 dans $[0,1]^{250}$. Du coup, il n'est plus possible de visualiser ce nuage de points pour y déceler des groupes.

Figure 3 : Dendrogramme de regroupement des douze ouvrages de J.R.R. Tolkien par la distance de Manhattan appliquée aux vecteurs des fréquences relatives des 250 mots les plus fréquents



En revanche, l'algorithme de Lance-Williams s'applique tout aussi facilement qu'avant. Il conduit au dendrogramme de la figure 3, qui s'appuie cette fois sur la distance de Manhattan, généralement préférée pour ce type d'application.

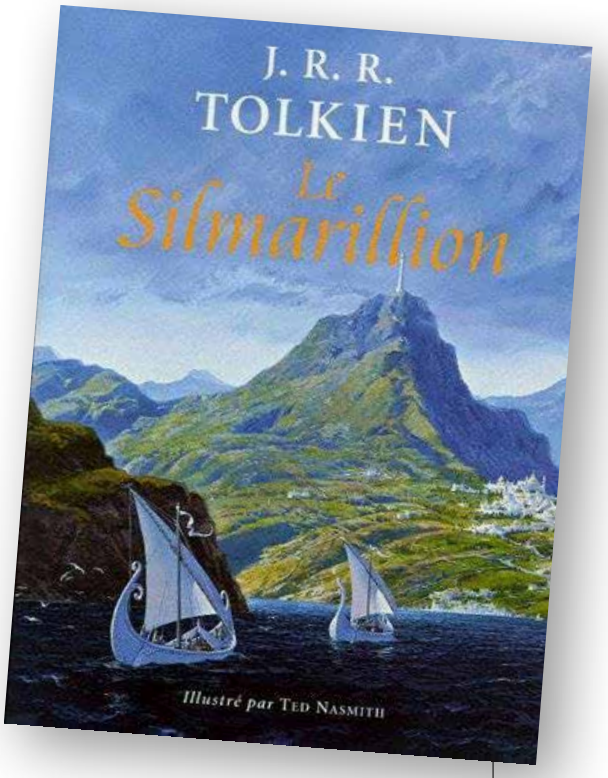
Comme précédemment, trois ouvrages se détachent du lot et constituent très nettement un groupe à part. Il s'agit de *Contes et Légendes Inachevés*, *Le Silmarillion* et *Les Enfants de Húrin*. Qu'ont-ils donc de particulier ? Les amateurs de littérature fantastique auront spontanément trouvé la réponse : ces trois ouvrages, publiés à titre posthume, ont été complétés par un des fils de J.R.R. Tolkien en sa qualité d'exécuteur littéraire. Avec l'aide de l'écrivain canadien Guy Gavriel Kaye, Christopher Tolkien a consacré plusieurs années à l'édition de ces trois ouvrages et autres textes inédits de son père. Il appert que les styles du père et du fils (ou plus exactement leurs champs lexicaux) se distinguent par un je-ne-sais-quoi que l'analyse de données textuelles a démasqué en un tournemain.

Perspectives

Le traitement statistique de corpus d'écrits issus de grands fonds documentaires revêt un intérêt linguistique indéniable. Toutefois, l'engouement actuel pour l'analyse textuelle prend sa source ailleurs. À l'aire des téléphones intelligents et d'internet, les gens sont de plus en plus connectés et ils génèrent des masses de données textuelles. Sur les réseaux sociaux, la capacité à analyser ces textes permet de classer les usagers plutôt que les œuvres. On peut ainsi associer des individus à des groupes d'intérêt, afin de leur offrir des produits et services qui correspondent à leurs préférences. C'est d'ailleurs ainsi, hélas ! que les conspirationnistes sont alimentés en contenus qui renforcent leurs convictions. Les méthodes de regroupement permettent aussi de traiter des données non textuelles. Elles sont utilisées entre autres

par certains fournisseurs de téléphonie mobile pour segmenter leur clientèle en classes de service auxquelles elles peuvent alors offrir des tarifs et des promotions taillés sur mesure. Bien qu'elles visent en principe à «améliorer l'expérience usager», ces stratégies de marketing peuvent néanmoins induire certaines formes d'inéquité puisque c'est à leur insu que certains clients n'ont pas accès à certains forfaits.

Par ailleurs, l'approche décrite ici s'appuie exclusivement sur la fréquence des mots, sans tenir compte par exemple de la façon dont ils sont assemblés. Même si c'est là l'approche la plus communément utilisée en analyse textuelle, sa capacité à décrypter le sens des écrits s'avère limitée. Plusieurs axes de recherche en informatique, statistique et linguistique tentent de déverrouiller la signification des mots par une analyse sémantique des textes, mais il ne semble pas encore y avoir de consensus sur l'approche à adopter pour résoudre ce problème.



J.R.R. Tolkien (1892–1973)

John Ronald Reuel Tolkien est un écrivain, poète, philologue et professeur d'université, né le 3 janvier 1892 à Bloemfontein (Afrique du Sud) et mort le 2 septembre 1973 à Bournemouth (Angleterre). Après des études à Birmingham et à Oxford, il fait carrière comme professeur d'anglais et de littérature anglaise à l'Université de Leeds, puis à Oxford. Sa production littéraire est marquée par son amour des langues et sa grande érudition. Il acquiert une notoriété publique grâce surtout à ses romans *Le Hobbit* (1937) et la trilogie du *Seigneur des Anneaux* (1954-55). Il prend sa retraite universitaire en 1959 mais continue de travailler sur sa mythologie jusqu'à sa mort, sans parvenir à parachever *Silmarillion*. Il décède à l'âge de 81 ans. Entre autres réalisations d'une vie intellectuelle très riche, il a publié des poèmes et participé à la traduction de la Bible de Jérusalem.