Analyse temporelle de données textuelles provenant de DBLP

Clément Malvy, Rayan Moissonnier, et Noé Lebreton
{clement.malvy,rayan.moissonnier,n.lebreton}@univ-lyon2.fr
Encadrant : Fadila Bentayeb

Résumé : Depuis plusieurs années, nous perfectionnons des méthodes d'analyse de données textuelles. Cependant, il y a peu de travaux concernant l'analyse dans le temps de ces données. Cet article explore donc des approches permettant ces analyses. Dans cette optique, une méthode largement quantitative fut adoptée, consistant en une analyse factorielle des données, soutenue par une approche plus mathématique de comparaison de l'évolution des mots à des droites ou des lois. Les résultats obtenus offrent de nouvelles perspectives et nous confortent quant à la pertinence d'analyses temporelles, puisque nous pouvons détecter des ressemblances entre les années mais également des événements comme l'émergence et disparition de termes.

Mots clés : Données textuelles, évolution temporelle, ACP, fréquences d'apparition

1 Introduction

L'informatique est un domaine imposant qui évolue rapidement, il est important de prendre conscience que les mots utilisés dans ce domaine peuvent varier rapidement. Dans notre sujet, nous avons cherché à détecter leur fréquence d'utilisation mais aussi les nouveaux mots qui apparaissent au cours du temps. Pour cela nous avons utilisé les données d'un site web qui recense les publications d'articles scientifiques liés à l'informatique. DBLP répertorie des revues, des conférences, des ateliers et des monographies ; ce site comptait en 2019 plus de 4.4 millions de publications écrites par 2.2 millions d'auteurs. Nous avons fait le choix de ces données pour plusieurs raisons, le premier étant que nous avons déjà travaillé sur ces données mais pas d'un point de vue temporel. La deuxième raison concerne la qualité des articles de DBLP, sur ce site il n'est pas possible de publier n'importe quel type d'article. Pour résumer notre choix, les mots-clés provenant du titre des articles sont un bon axe pour analyser le monde de la recherche.

Dans un sujet comme le nôtre il y a plusieurs axes de recherche comme la détection de thème via des méthodes de classification ou encore la prédiction des mots-clés présents dans le futur. Pour cet article nous avons fait le choix de nous orienter vers l'étude des ressemblances entre les mots et les années ainsi que la détection d'évolution des mots d'un point de vue temporel.

2 Etat de l'art

La détection de ressemblance entre les individus et les variables est un sujet assez vaste. Les articles qui ont motivé notre travail traitent de sujets divers comme l'analyse de données textuelles, la fréquence des mots, mais aussi la labellisation des données. Dans cette partie de notre article, nous allons faire le parallèle avec des travaux connexes.

Les auteurs des articles [6] et [2] traitent de sujet similaire, l'article [6] analyse le discours en matière d'informatique scolaire au Québec, ils utilisent plus particulièrement des méthodes factorielles appliquées à des données textuelles tout comme l'article [2] qui lui traite la réponse à la question "Aimez-vous les maths ?". Ces deux articles utilisent des ACP pour répondre à leurs problématiques, cela nous a donc permis d'orienter nos recherches pour expliquer la ressemblance entre les années, mais aussi l'éventuel lien entre les groupes d'années et les mots.

Les articles [1] et [4] traitent des sujets liés à la fréquence des mots, nous avons retrouvé un point pour lequel nous avons porté un grand intérêt. L'article [4] est une étude sur le suivi du sens des mots, elle met en avant la fréquence des termes des archives du Times pour l'utilisation du mot Saint-Pétersbourg, en calculant à la fois des fréquences, mais aussi en affichant des graphiques avec une dimension temporelle. Quant à l'article [1] qui traite du sujet de la statistique sur la fréquence des mots dans la linguistique de corpus, a de grandes similitudes avec l'article [1] dans leurs grandes lignes. Il traite également du sujet de la fréquence des mots et présente des graphiques avec une dimension temporelle, cela nous a inspiré pour la représentation graphique des évolutions temporelles significatives des mots de DBLP.

L'article [5] est certainement celui ayant répondu à un de nos plus grand problème, malheureusement il n'est pas disponible à la lecture, mais il nous a permis de réfléchir à un point important. Cet article traite de l'évolution du langage, le résumé met en avant l'importance de la prise en compte de l'effet d'échelle des données lorsque nous appliquons des statistiques avec une vision temporelle.

L'article [3] traitant des statistiques des fréquences de mots sur les données de Google-ngram, Wikipedia anglais et une collection d'articles scientifiques. L'auteur utilise plusieurs approches, l'analyse statistique de textes et des modèles stochastiques simples. Ces méthodes lui permettent d'expliquer l'apparition des lois d'échelles dans la fréquence des mots. L'auteur met en place un test d'adéquation avec une loi pour l'intégralité d'un corpus, nous avons fait le choix de réduire l'axe de travail à des mots pour tester l'adéquation de la répartition de la fréquence à une droite. Nous avons pour objectif de faire varier en fonction du type d'évolution que nous souhaitons observer.

3 Solution proposée

Afin de répondre à la problématique de l'article, nous avons développé dans cette partie nos 3 différents axes de travail.

3.1 Prétraitement des données

Dans le cadre de notre article, nous avons fait le choix de travailler sur un jeu de données issu du site web DBLP, il est important de préciser que ces fichiers sont fournis par un groupe d'étudiants du master 1 Informatique de l'université Lyon 2. Afin d'obtenir un jeu de données répondant à nos attentes, nous avons appliqué plusieurs filtres comme par exemple la suppression des publications de type "www" qui ne nous apportent aucune information nécessaire.

Suite à cette étape, nous nous sommes rendu compte que les publications de notre jeu de données étaient en plusieurs langues. L'objectif que nous voulions atteindre pour commencer nos analyses était d'obtenir un corpus de mots ayant une seule langue pour notamment supprimer la redondance de mots mais aussi supprimer les "stopwords" qui apporteraient un biais dans nos futures analyses. Nous avons tout d'abord essayé de traduire la totalité des mots en se connectant à une API permettant de faire ce travail. Malheureusement ce type d'application n'était pas forcément adapté à nos objectifs notamment à cause de la volumétrie de nos données. C'est donc pourquoi nous avons fait le choix de faire ce travail d'une manière plus manuelle.

Premièrement, nous avons essayé de répondre à cet objectif via la détection automatique de langage. Après quelques tests, les résultats obtenus n'étaient pas très probants, en d'autres termes nous avons constaté une mauvaise adéquation entre la langue prédite et la langue réelle. Pour pallier à ce problème, nous avons fait le choix de gérer le langage à partir d'une liste de "stopwords". Cela a été possible grâce au package nltk disponible sur Python qui inclut des corpus de "stopwords" en fonction de la langue. Nous avons créé un algorithme qui d'abord supprime les "stopwords" anglais et par la suite nous nous sommes attardé sur le cas des mots allemands. Notre objectif était de supprimer les publications allemandes du corpus, pour cela nous avons tout d'abord retenu les publications ayant un signe diacritique (accent, tréma, tilde ...) mais aussi les publications ayant au moins un "stopword" allemand ; pour finalement les supprimer de notre jeu de données.

Après avoir testé notre méthode expliquée précédemment sur des échantillons, nous avons souhaité l'appliquer sur l'intégralité de notre corpus. Nous nous sommes rendu compte que la complexité de notre algorithme était bien trop élevée pour obtenir des résultats dans un temps raisonnable. Lors de nos tests nous avons fait le constat que pour un même nombre d'individus le temps d'exécution est plus rapide pour plusieurs petites bases de données plutôt qu'une seule. Nous avons donc fait le choix de diviser notre jeu de données afin de réduire le temps d'exécution.

Afin de pouvoir observer l'évolution temporelle de chaque mot nous avons fait le choix de décomposer la date en années, mois et jours. De plus, pour appliquer nos méthodes, nous avons effectué une transformation concernant la mise en forme de nos données pour finalement obtenir un tableau avec en colonnes les années et en lignes les mots.

3.2 Analyse factorielle sur les données via la méthode de l'ACP

L'analyse en composante principale est une méthode d'analyse de données, plus généralement de la statistique multivariée. Elle permet d'explorer les jeux de données constituées de variables quantitatives. Elle consiste à transformer les variables liées entre elles en variables décorrélées les unes des autres, c'est ce que l'on appelle les composantes principales. Elle permet de réduire le nombre de variables et donc de réduire la redondance d'information. Si l'information associée aux deux premiers axes représente un pourcentage suffisant de la variabilité (ou inertie) totale du nuage de points, on pourra représenter les observations sur un graphique à deux dimensions, facilitant ainsi son interprétation.

Concrètement, il existe plusieurs applications pour l'ACP, tel que l'étude et la visualisation des corrélations entre les variables, l'obtention de facteurs non corrélés (utilisés dans des méthodes de modélisation) ainsi que la réduction de dimension.
Ici, nous allons utiliser l'ACP pour analyser les relations entre les variables, les individus mais aussi les relations croisant à la fois les individus et les variables. Dans notre cas, cela revient à analyser les ressemblances entre les années du point de vue des mots utilisés dans les titres des publications.

Afin de réaliser cette ACP, nous devions choisir des données pertinentes au vu de nos analyses. Il fallait donc perdre le moins d'information possible tout en ayant un nombre de données limité. L'importance des dimensions dans le cadre d'une ACP est essentielle notamment pour la représentation et les interprétations. Après de nombreux essais en faisant varier le nombre d'individus, nous avons choisi de retenir les 100 mots les plus présents dans la base de données. Pour revenir à notre tableau de données, les individus sont les mots, les variables les années et le croisement entre les deux correspond au nombre de fois où le mot apparaît dans l'année.

De plus, avec la présence de l'effet d'échelle dûe au fait que le nombre de publications augmente dans le temps, nous avons décidé d'appliquer une ACP normée qui intègre directement un processus de standardisation des variables permettant d'homogénéiser celles-ci et donc de supprimer les différences d'ordre de grandeur.

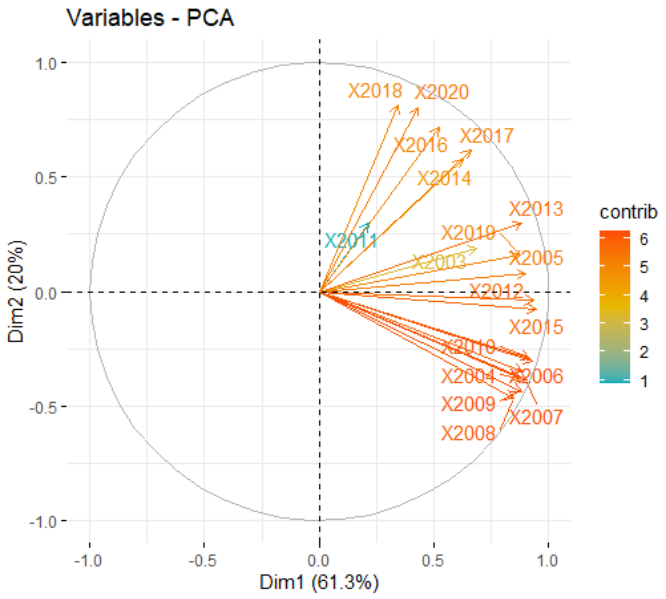
Tout d'abord, nous devons choisir le nombre d'axes principaux. Ce choix est appuyé par le critère de Kaiser-Guttman qui repose sur une idée simple. La moyenne des valeurs propres dans le cas d'une ACP normée vaut 1, nous considérons un axe intéressant si sa valeur est supérieure à 1. Ici les deux premières valeurs propres (11.6 et 3.8) sont supérieures à la moyenne tandis que la troisième (0.94) et les suivantes sont inférieures à 1. De plus, si nous choisissons de conserver les deux premiers axes, nous conserverons 81.3% des données et cela nous permettra de représenter les résultats sur un graphique à deux dimensions. Grâce à ce choix, nous pourrons mieux visualiser et interpréter les relations présentes dans notre jeu de données.

Dans un deuxième temps, nous allons construire les nuages de points projetés. Chacun de ces nuages (variables et individus) est construit en projection sur les plans factoriels. En examinant ces derniers, nous pourrons visualiser les corrélations entre les variables ainsi qu'identifier les individus qui se ressemblent. En d'autres termes, nous pourrons apprécier les mots qui sont proches et qui auront tendance à être utilisés de

manière similaire au cours du temps.

Ensuite nous allons interpréter les deux axes retenus, pour se faire nous allons étudier grâce au nuage de points quels sont les variables et individus qui participent le plus à la formation de cet axe.

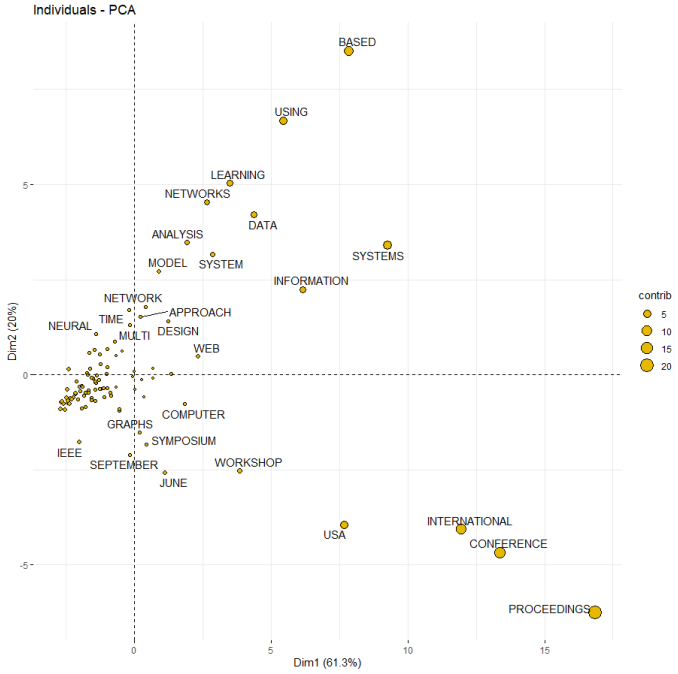
Figure 1: Projection des variables sur le plan factoriel



Sur le graphique ci-dessus, nous pouvons donc remarquer qu'il y a trois grands groupes de variables (et l'année 2011 qui est un peu exclue de ces groupes). Ainsi, l'axe 1 possède une caractéristique remarquable puisqu'aucune des variables n'a de coordonnées négatives sur cet axe. De plus, seules les variables 2011, 2018 et 2016 apportent une contribution faible à cet axe avec une coordonnée inférieure à 0.5. Toutes les autres variables contribuent positivement et de manière significative. Évidemment il existe des années qui ont une coordonnée plus élevée que d'autres, par exemple 2015 et 2012 qui ont des coordonnées supérieures à 0.93 parallèlement à 2014 et 2017 qui sont situés à environ 0.65. Pour conclure sur cet axe, il semble instaurer une échelle entre les années.

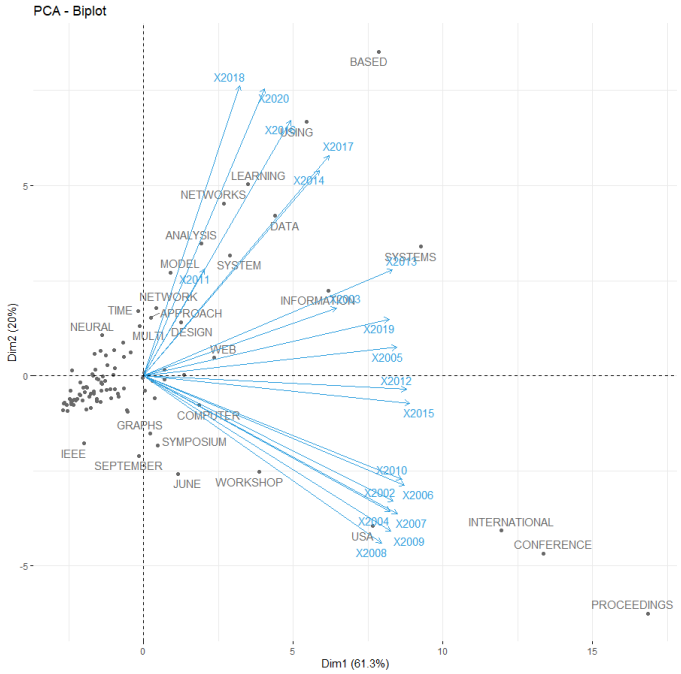
Concernant l'axe 2, il semble différencier les années les plus anciennes de celles plus récentes au vu du graphique, des contributions mais aussi des coordonnées des variables. En effet, les variables ayant une coordonnée négative sur cet axe semblent globalement correspondre aux années les plus anciennes (avant 2010). Tandis que les variables avec une coordonnée positive représentent les années plus récentes (entre 2014 et 2020). Il y a cependant quelques exceptions à noter puisque 2015 est projetée du côté négatif alors que les années 2003 et 2005 du côté positif.

Figure 2: Projection des individus sur le plan factoriel



Ensuite, sur la figure 2 nous pouvons d'abord remarquer qu'une grande majorité d'individus sont concentrés à un même endroit, près du croisement des deux axes. Ils ont donc des coordonnées très faibles et par conséquent ne contribuent que très peu aux deux axes. Cependant nous pouvons remarquer quelques individus qui contribuent fortement à la construction des axes, comme les mots "BASED", "USING"," SYSTEMS" ou encore "PROCEEDINGS". Leur place dans le nuage de points permet d'expliquer leur contribution aux axes.

Figure 3: Projection des variables et des individus sur le plan factoriel



Le graphique ci-dessus est sans doute le plus intéressant puisqu'en plus de montrer la contribution des variables et des individus sur les axes, il permet également de montrer l'interaction entre eux. Ainsi, grâce aux interprétations précédentes, nous pouvons par exemple affirmer que les mots se situant proche de la projection des années sont plus utilisés ces années-là. Ainsi, les mots "INTERNATIONAL","CONFERENCE" et "PROCEEDINGS" sont employés dans les années les plus anciennes alors que "BASED" et "USING" eux, représentent les années les plus récentes. Nous avons vérifié ces observations dans nos données de base normalisées et cela correspond effectivement à nos résultats.

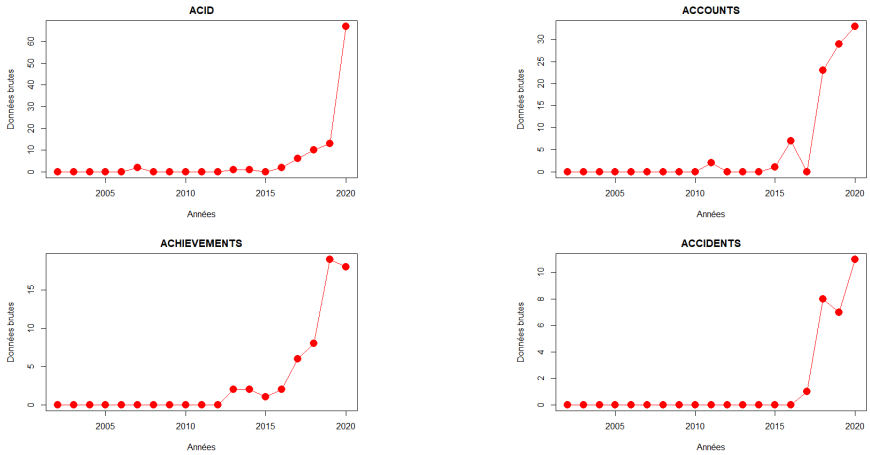
Pour conclure sur cette partie de notre article, cette ACP nous a permis de mieux comprendre nos données en visualisant les relations entre les variables et les individus. L'analyse de ces liaisons était compliquée à observer sur les données brutes, avec l'aide de méthodes factorielles comme l'ACP nous avons pu résumer l'information et tirer certaines conclusions.

3.3 Visualisation de l'évolution significative temporelle des mots

L'un des objectifs de notre article était d'étudier l'évolution temporelle des mots provenant d'une base de données importante en termes de volume. Lors de notre phase d'expérimentation, nous avons rencontré plusieurs types de problèmes, la première difficulté provenait de la variabilité au niveau de l'échelle de nos données. Pour prendre conscience de cette difficulté il est important de bien comprendre le monde de la recherche ainsi que la publication d'articles. Avec l'émergence d'internet ainsi que l'arrivée de sites comme DBLP la fréquence de publication est en constante évolution, ce qui traduit donc le fait de prendre en compte le nombre de publications pour pouvoir réaliser des analyses temporelles pertinentes.

Après quelques tests, nous nous sommes rendu compte rapidement que l'utilisation de données brutes n'avait pas de sens pour répondre à notre problématique. Sur les quatre graphiques présentés ci-dessous (fait avec les données brutes) on peut observer qu'il y a une forte évolution de l'utilisation de ces mots. Cette évolution n'est absolument pas pertinente car elle est uniquement basée sur l'effet d'échelle.

Figure 4: Graphiques représentant l'évolution temporelle des données brutes



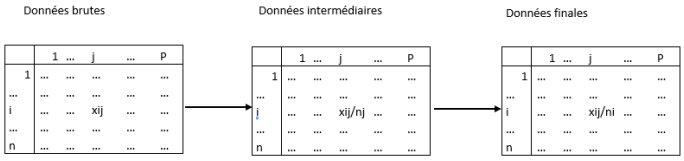
Pour pallier ce problème nous avons réfléchi à la transformation de nos données dans l'optique de supprimer l'effet d'échelle lié à l'augmentation du nombre de publications. Pour cela nous avons appliqué une formule permettant de transformer nos données en fréquences. Ci-dessous vous trouverez la formule permettant de calculer ces fréquences ainsi qu'un exemple détaillé sous forme de tableaux.

$$x_{i,j}(1) = \frac{x_{i,j}}{n_j} \text{ avec } n_j = \sum_{i=1}^{n} x_{i,j}$$

$$x_{i,j}(2) = \frac{x_{i,j}(1)}{n_i(1)} \text{ avec } n_i(1) = \sum_{i=1}^{n} x_{i,j}(1)$$

$x_{i,j}(1)$ correspond à la fréquence par année et $x_{i,j}(2)$ à la fréquence par mots

Figure 5: Tableaux résumant la transformation des données

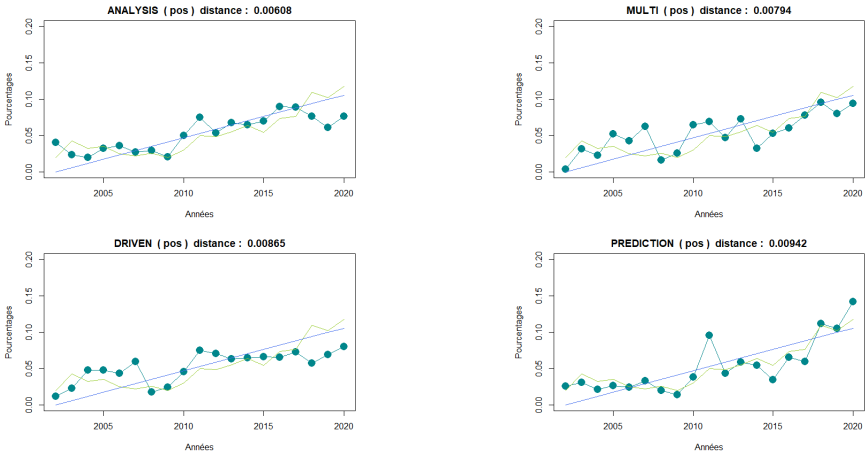


Suite à cette étape, nous avons pu commencer à analyser les évolutions des mots de notre base de données. Le second problème rencontré concerne la volumétrie de nos données, à ce stade de notre étude nous avions plus de 12000 mots avec chacune des fréquences par année. La question était donc de savoir comment détecter si un mot avait une augmentation d'utilisation significative, aucune évolution au cours du temps ou encore une diminution d'utilisation.

Afin de répondre à cette question nous avons décidé de regarder si la répartition des fréquences de chacun des mots était proche d'une droite en termes de distance. L'objectif était donc de calculer pour chaque mot une distance au carré entre différents types de droite et la fréquence d'utilisation d'un mot.

Premièrement nous avons calculé l'ensemble des distances par rapport à la droite $x = y$, en d'autres termes une évolution dite "linéaire". Etant donné que nous travaillions mot par mot la somme des points de la droite devait obligatoirement être égale à 1 pour conserver l'échelle utilisée. Suite à cette étape de calcul des distances nous avons sélectionné uniquement les mots ayant la distance la plus faible avec la droite $x = y$, ce qui veut dire dans notre cas les mots ayant connu une évolution linéaire la plus significative. Sur les quatre graphiques ci-dessous on observe les mots (courbe bleue avec des points) qui minimisent la distance au carré avec la droite $x = y$ (droite bleue) ainsi que l'évolution moyenne de l'utilisation des mots (courbe verte).

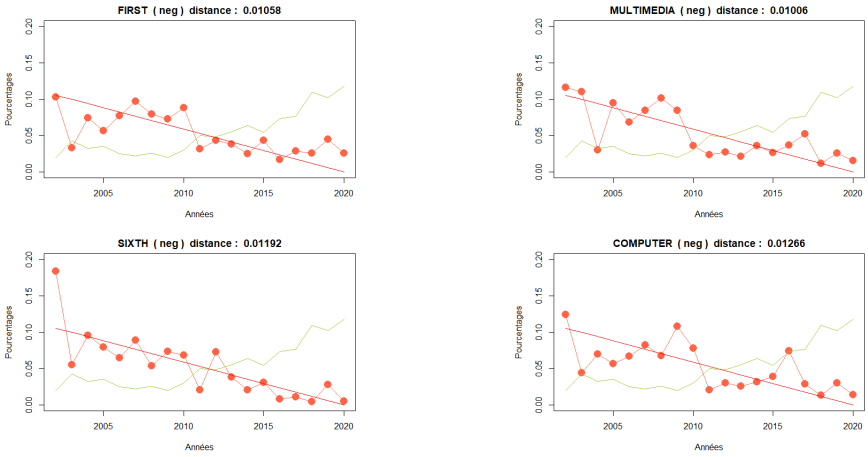
Figure 6: Graphiques des mots les plus proches de la droite linéaire positive



Dans un premier temps, on observe que chacune des distances est très proche de 0 ce qui traduit la proximité avec la droite $x = y$. Malgré l'utilisation de la distance au carré nous observons certaines fluctuations dans la fréquence concernant les mots ce qui traduit en quelque sorte les limites de notre technique.

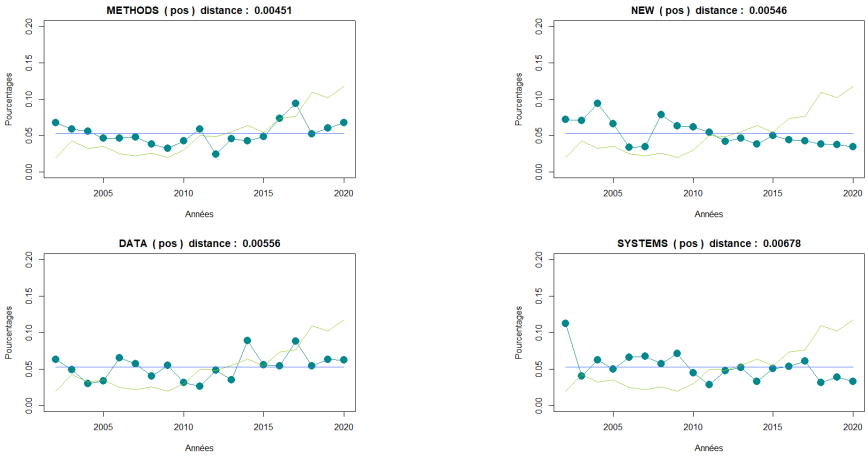
Suite à l'étude des distances avec la droite $x = y$ nous avons décidé de faire varier la droite pour essayer d'observer différentes évolutions. La deuxième droite étudiée correspond à la symétrie de la première, c'est-à-dire $x = -y$ qui représente donc une diminution linéaire. Grâce au même processus, nous avons pu calculer les distances au carré pour ensuite obtenir uniquement les mots les plus proches de cette droite. Avec les graphiques présentés ci-dessous on peut constater que l'étude de la diminution de la fréquence des mots est plus compliquée que celle portant sur l'évolution. Ce constat vient simplement du fait qu'un mot peut connaître une diminution au fil du temps mais il sera toujours légèrement utilisé dans la littérature.

Figure 7: Graphiques des mots les plus proches de la droite linéaire négative



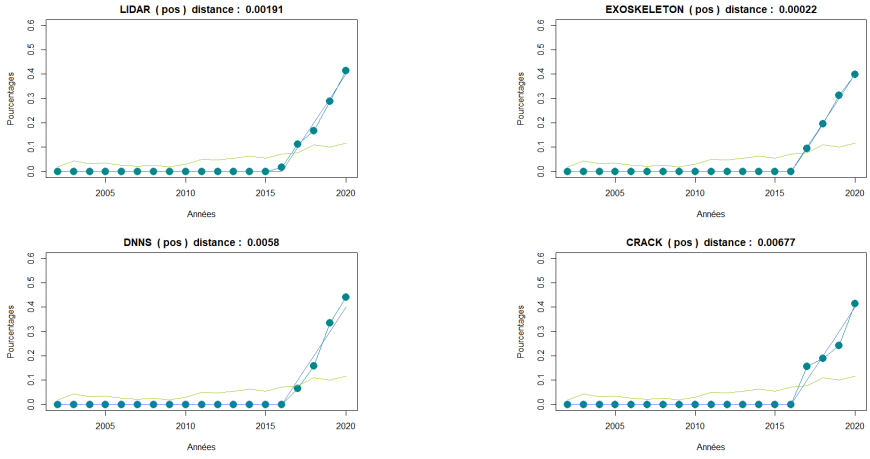
Le deuxième objectif de cette partie de notre article était d'analyser si certains mots connaissaient une utilisation constante dans le temps. Pour répondre à cet objectif nous avons utilisé la même méthodologie que précédemment en changeant notre base pour le calcul des distances. La nouvelle droite était donc égale à $x = c$ avec $c = 1/19$ ce qui nous a finalement permis d'obtenir les mots les plus constants dans le temps. Il est important de rappeler que l'utilisation constante d'un mot durant le temps est compliquée à étudier notamment à cause des tendances présentes dans la littérature. Chacun des graphiques ci-dessous connaît des fluctuations entre les différentes années, mais globalement nous pouvons voir que l'utilisation de ces mots est relativement stable dans le temps.

Figure 8: Graphiques des mots les plus proches de la droite constante



Enfin pour finir sur cette partie concernant la visualisation de l'évolution temporelle des mots nous avons fait le choix d'analyser les fortes évolutions positives ainsi que négatives. Pour cela, nous avons uniquement modifié le type de droite servant dans le calcul des distances. Cette droite prend une nouvelle forme qui cible uniquement les 4 dernières années, ce qui permet d'étudier plus particulièrement l'apparition de nouveaux mots dans le domaine de la recherche informatique. Comme vous pouvez le voir dans la figure 9, les distances entre la droite et les fréquences d'utilisation des mots sont très faibles ce qui traduit une bonne adéquation de la répartition des fréquences à la droite. Nous avons pu voir précédemment que les distances étaient globalement faibles mais en conservant quelques variations temporelles. Dans notre cas les variations sont minimes c'est donc pour cela que nous nous permettons de les analyser plus en profondeur. Pour prendre l'exemple du mot LIDAR, celui-ci représente quelque chose de connu depuis un certain temps (1930) pour être plus précis c'est une technique de mesure à distance fondée sur l'analyse des propriétés d'un faisceau de lumière renvoyé vers son émetteur. Cette technique était beaucoup utilisée notamment pour le contrôle de la vitesse en voiture ou encore la cartographie de la planète Mars. Nous pouvons donc nous demander pourquoi ce mot a connu une évolution dans son utilisation depuis 2016, en faisant des recherches sur cette technique nous avons pu la lier avec la démocratisation de la recherche sur les voitures autonomes.

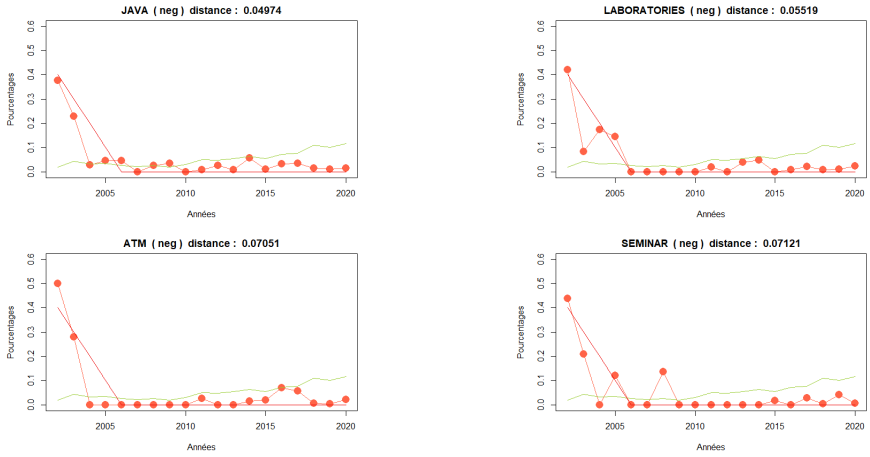
Figure 9: Graphiques des mots les plus proches de la droite d'évolution positive sur les 4 dernières années



Après avoir étudié l'évolution positive de l'utilisation des mots sur les 4 dernières années, nous nous sommes orientés sur la symétrie de la droite utilisée précédemment. Cette nouvelle droite traduit d'un point de vue théorique la forte diminution d'utilisation des mots, la droite commence donc à une fréquence élevée en 2002 pour finir à une fréquence égale à 0 en 2006, de plus cette fréquence restera stable jusqu'en 2020. Comme dans la première étude avec les droites $x = y$ et $x = -y$, on observe qu'il est plus difficile d'étudier la diminution des mots que l'évolution. Sur les graphiques ci-dessous on peut observer les mots les plus proches de la droite en termes de distance,

on constate une légère fluctuation mais une logique plutôt pertinente. Pour prendre l'exemple du mot JAVA, on distingue clairement que sa fréquence d'utilisation a baissé ces dernières années, ce qui est clairement en rapport avec l'actualité du domaine informatique.

Figure 10: Graphiques des mots les plus proches de la droite d'évolution négative sur les 4 dernières années



Pour résumer cette partie de notre article concernant la visualisation de l'évolution significative temporelle des mots, nous avons eu la possibilité de faire certains choix comme la métrique ou encore la forme des droites pour le calcul des distances. Ces différents paramètres sont certes discutables mais pour répondre à notre problématique ils sont justifiés. Les résultats obtenus dans cette partie sont, de notre point de vue, très pertinents car ils répondent entièrement à nos questions.

4 Conclusion

Notre problématique était d'analyser temporellement les mots issus d'une base de données bibliographique avec pour sujet l'informatique. Pour ce faire nous avons donc choisi deux axes différents. Le premier avec une analyse factorielle et le second avec une approche plus mathématique ; la comparaison d'un graphe de l'évolution d'un mot avec une courbe choisie en fonction de ce que nous voulions.

Dans la première approche, l'ACP nous a permis d'obtenir plusieurs résultats. En l'occurrence, les années qui se ressemblent en terme d'utilisation des mots, mais également les mots qui sortent du lot, et qui permettent de différencier les groupes d'années qui se ressemblent. Elle nous a, en outre, permis de mieux comprendre les données.

Pour parler de la deuxième partie de notre article, nous avons pu mettre en place une méthode d'analyse temporelle grâce à nos connaissances dans le domaine de l'étude des données textuelles. Cette méthode portant sur la détection de mots connaissant des évolutions significatives dans le temps, nous a permis d'obtenir un corpus de mots qui résume au mieux les phénomènes observés.

Cet axe d'analyse peut être amélioré en faisant varier davantage les paramètres de notre méthode. Premièrement il peut être intéressant d'avoir comme base de calcul différentes droites selon nos hypothèses ou encore des lois statistiques. La création d'un algorithme modifiant automatiquement les droites pourrait être une continuité dans le cadre de notre travail de recherche. De plus, la notion de distance étant un concept relativement vaste, le fait de changer de type de distance pourrait fournir d'autres résultats à ces analyses. Enfin, la dernière amélioration possible serait de nuancer notre période d'analyse notamment en faisant varier la granularité au niveau de l'utilisation des mots d'un point de vue temporel.

References

[1] Paul Rayson Alistair Baron and Dawn Archer. *Word frequency and key word statistics in historical corpus linguistics.*

[2] Christian Baudelot. *Aimez-vous les maths ? Une analyse statistique de données textuelles.* 1991.

[3] Martin Gerlach and Eduardo G Altmann. *Scaling laws and fluctuations in the statistics of word frequencies.* 2014.

[4] Thomas Risse Nina Tahmasebi and Stefan Dietze. *Towards automatic language evolution tracking.* 2011.

[5] Florencia Reali and Thomas L. Griffiths. *Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift.* 2009.

[6] Grenon V. *Méthodes factorielles en statistique textuelle. Application à l'analyse du discours en matière d'informatique scolaire au québec.* 2003.