Partitionnement des données

Réalisé par

***Hiba Aneq, Ka-son Chau, Vincent Bédard***

***Groupe : 00001***

Présenté à

***Pierre-Paul Monty***

Dans le cadre du cours

***420-B53-VM DÉVELOPPEMENT EN ENVIRONNEMENT DE BASE DE DONNÉES***

16 décembre 2019

Le partitionnement des données par l’utilisation de centroïdes en tant que mots ainsi que points

**Hiba Aneq**

**Ka-son Chau**

**Vincent Bédard**

1. Résumé

La K-moyenne est une méthode de regroupement de donnée utilisée surtout en médecine, en science, en marketing, etc. Elle permet d’afficher des groupes, aussi appelée des centroïdes, ayant les même “patterns”. Malgré sa popularité, elle est peu, voire rarement, utilisé à des fins linguistiques, Ce qui est intéressant. L’intérêt créer et de pouvoir examiner jusqu’où la k-moyenne peut être efficace. Malheureusement, selon les solutions obtenues sur quatre expérimentations, elle ne permet point de trouver des synonymes, mais étonnamment elle arrive à afficher des groupes de mots dans certains centroïdes.

1. Introduction

Pour notre expérience, les distances ont été calculées par la méthode du moindre carré sur les positionnements de chaque mot. Doutant de l’efficacité de cette méthode et sur le fait de déceler un regroupement de synonymes dans un même centroïde, nos questionnements étant est-ce qu’on aura des résultats satisfaisants du point de vue que les mots sortis auront un lien entre eux sans pour autant être nommés synonymes, malgré le fait qu’il n’existe aucun synonyme parfait. Dans cet ordre d’idée, l’objectif de notre innovation est de permettre d’insinuer à quelle classe de mots le mot, en plus de sortir des synonymes, appartient en utilisant la méthode du KNN où en français la méthode des k plus proche voisins.

Tout compte fait, Ce problème est intéressant d’un point de vue informatique et sémantique. Trouver des synonymes informatiquement semble être une tâche chimérique où l’on ne peut confondre la langue et ses règles avec un programme informatique. On sait qu’il y aura une marge d’erreurs. En faisant ce travail, on se retrouve dans plusieurs impasses où ses questionnements subsistent, même à la fin.

1. Matériel

L’ensemble des données indispensable au fonctionnement de nos expériences sont bien entendus : la base de donnée SQL ite où sont stockés nos mots, nos cooccurrences entre chaque mot, ainsi que la taille de la fenêtre de cooccurrences. L’utilité de la base de donnée est de pouvoir faire les calculs nécessaires du K-moyenne en utilisant la cooccurrence entre chaque mot tout en sélectionnant les mots dans la taille de la fenêtre correspond. Ensuite, selon le choix de l’usager, les mots choisis en tant que centroïdes permettent de situer où se trouve les premiers points d’impact dans un univers avec plusieurs axes qui représentent chaque mot et leurs positions les uns envers les autres. Bref, dans notre cas, 25 000 axes par 25 000 axes. En d’autres parties, le choix aléatoire avec un nombre précis de centroïdes, quant à lui, implique que chaque centroïde peut se trouver à n’importe quelle position sans qu’on n’y puisse rien faire dû, surtout, au fait que la fonction aléatoire est multiplié par le nombre de mots présents dans la base de donnée.

1. Méthodes

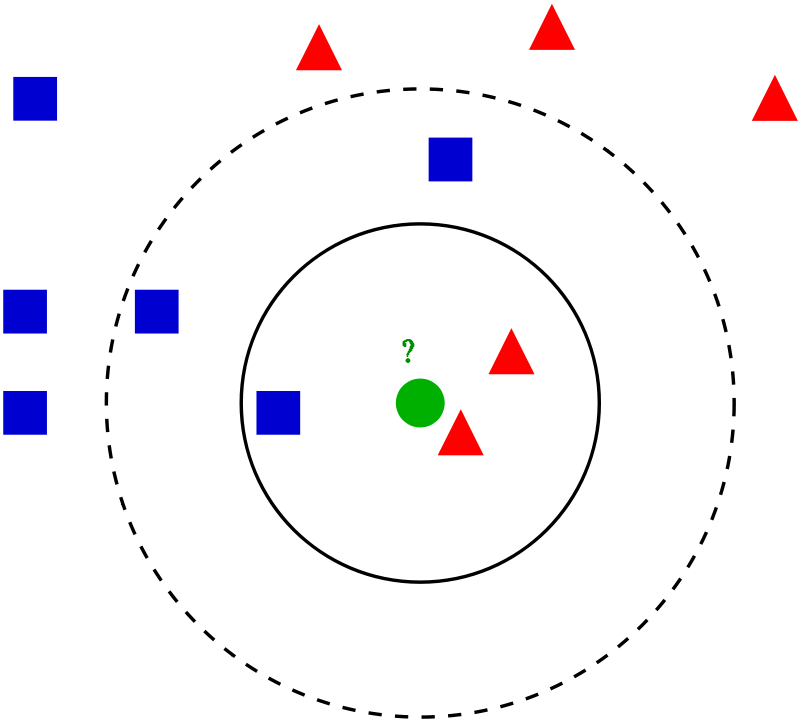
Cette expérience est définie par quatre étape. L’étape une est l’initialisation de nos centroïdes où on définit quels sont nos centroïdes initiaux, ainsi que leurs points bien qu’ils soient des mots ou des points aléatoires. Par la suite, l’étape deux est d’assigner les points au centroïde le plus proche en vérifiant la distance de chaque point par rapport à chaque centroïde en utilisant la méthode du moindre carré, connue sous la formule suivante :

(𝑎−𝑐)^2+〖(𝑏−𝑑)〗^2 𝑣𝑠 (𝑎−𝑒)^2+〖(𝑏 −𝑓)〗^2[[1]](#endnote-1)

Ensuite, la troisième étape est de faire la moyenne pour trouver chez chaque centroïde le barycentre en utilisant le calcul suivant :

𝐺=1/𝑛 (𝐴\_1+𝐴\_2+…+𝐴\_𝑛)[[2]](#endnote-2)

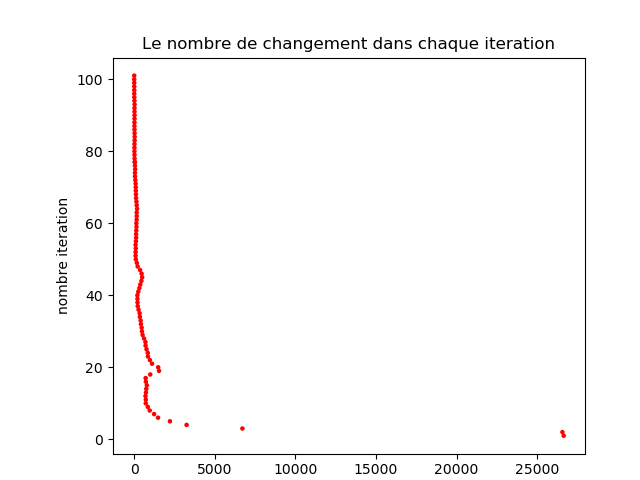
Où G est le barycentre, n est le nombre de points dans ce centroïde et A représente la valeur de chaque point. Enfin, on répète l'étape deux pour redistribuer les points dans chaque centroïde, car la distance peut fluctuer par rapport au nouveau centre et l’étape trois jusqu’à ce qu’il n’y est plus d'itérations dans chaque centroïdes. À ce moment-ci on affiche les mots qui se retrouvent dans chaque centroïde selon le nombre d’affichage maximale décidée.

[[3]](#endnote-3)Pour les k plus proches voisins, il se compose de trois étapes. La première est d’avoir la donnée où l’on cherche l’appartenance. Ensuite, on regarde autour de la donnée.

Puis, on détermine un certain nombre de mots avec leurs classes qui entourent la donnée, on peut leur donner un poids. Sinon, en dernier, la donnée prend la classe qui a le plus de vote.

1. Résultat

Les expériences ont été effectuées sur un nombre de centroïdes où notre but est de voir la corrélation entre le nombre de centroïdes choisis et le nombre d’itérations, ainsi que les résultats affichés.

Le premier est le nombre de changement dans chaque itération avec vingt centroïdes.

-----------------------------------------------------------

* Résultats du 1 er centroïde
* ('sentait', 954.5759368836294) est un VER
* ('disait', 963.5502958579882) est un NOM
* ('pourtant', 1030.293885601578) est un VER
* ('voyait', 1067.2426035502958) est un NOM
* ('restait', 1078.8836291913217) est un NOM
* ('semblait', 1203.0118343195268) est un NOM
* ('faisait', 1271.319526627219) est un NOM
* ('devait', 1336.3195266272191) est un VER
* ('faut', 1417.524654832347) est un NOM
* ('regardait', 1483.5246548323476) est un VER

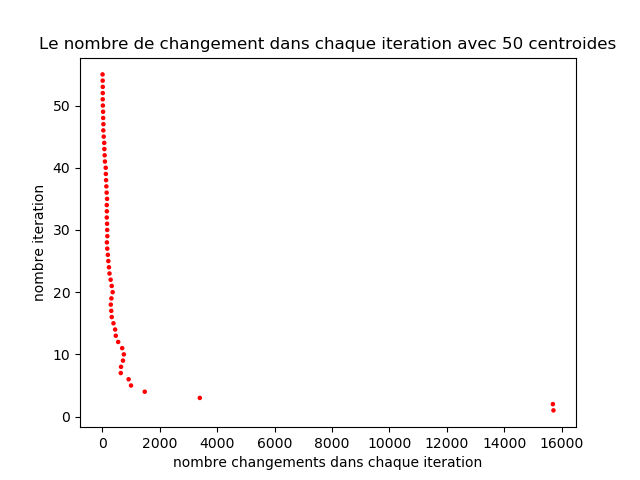
-----------------------------------------------------------

* Résultats du 4ème centroïde
* ('douleur', 135.61610509460928) est un NOM
* ('désespoir', 150.5621631858956) est un VER
* ('souffrance', 158.73228766722337) est un VER
* ('révolte', 159.22191422323996) est un VER
* ('crime', 159.5912088290491) est un VER
* ('chute', 163.55109817897997) est un NOM
* ('intérieur', 164.90241214855118) est un NOM
* ('sentir', 165.41555184426346) est un VER
* ('garder', 166.8650677502109) est un NOM
* ('chien', 169.2274467266978) est un NOM

-----------------------------------------------------------

* Résultats du 5ème centroïde
* ('pavillons', 108.6729044559) est un NOM
* ('bâtiments', 120.20231622060588) est un VER
* ('marchandes', 124.05945907774876) est un NOM
* ('galibots', 126.10987924581597) est un VER
* ('tailles', 127.82416496010167) est un VER
* ('herscheuses', 130.6981145399336) est un VER
* ('voies', 132.1350893298496) est un VER
* ('terres', 138.00903890968152) est un NOM
* ('haveurs', 138.26113975001766) est un NOM
* ('chefs', 138.60567756514368) est un VER

-----------------------------------------------------------

Le deuxième graphique est sur le nombre d’itérations avec 50 centroïdes.

-----------------------------------------------------------

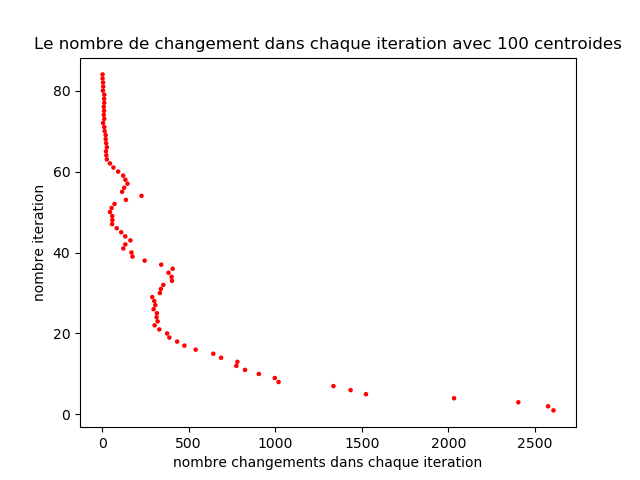
Résultats du 1er centroïde

* ('attendant', 285.75206611570246) est un NOM
* ('regardant', 325.3884297520661) est un NOM
* ('face', 439.5702479338843) est un NOM
* ('faisant', 523.5702479338843) est un NOM
* ('avant', 628.4793388429753) est un VER
* ('voyant', 680.8429752066116) est un VER
* ('Angleterre', 783.7520661157022) est un VER
* ('cas', 791.9338842975206) est un VER
* ('sûr', 860.6611570247934) est un VER
* ('autant', 1376.297520661157) est un ADJ

-----------------------------------------------------------

Résultats du 4ème centroïde

* ('conversation', 288.1404320987654) est un VER
* ('cour', 344.91820987654324) est un VER
* ('voiture', 447.6959876543209) est un ADJ
* ('mer', 507.9737654320987) est un NOM
* ('procureuse', 518.1404320987655) est un NOM
* ('veille', 579.1404320987655) est un VER
* ('dame', 592.0848765432098) est un VER
* ('fenêtre', 619.9737654320988) est un VER
* ('rochelle', 630.9182098765433) est un VER
* ('force', 655.3070987654322) est un NOM

-----------------------------------------------------------

-----------------------------------------------------------

Résultats du 1er centroïde

* ('clef', 98.4544) est un NOM
* ('puissance', 99.85440000000001) est un VER
* ('pointe', 102.2144) est un ADV
* ('beauté', 122.2144) est un VER
* ('douleur', 129.77439999999999) est un VER
* ('prison', 131.17439999999996) est un VER
* ('religion', 134.93439999999998) est un VER
* ('taille', 141.8144) est un ADJ
* ('comtesse', 141.81440000000003) est un VER
* ('mine', 144.4144) est un ADV

-----------------------------------------------------------

Résultats du 2ème centroïde

* ('défends', 11.286831935149403) est un VER
* ('vengerai', 11.67518144971251) est un VER
* ('retrouverai', 13.694598925440664) est un VER
* ('croirai', 14.490715430295033) est un VER
* ('regrette', 14.995569799227072) est un ADV
* ('verrai', 15.481006692430954) est un VER
* ('suivrai', 15.675181449712511) est un NOM
* ('reverrai', 16.45188047883872) est un VER
* ('demanderai', 16.490715430295033) est un VER
* ('perdis', 16.58780280893581) est un NOM

-----------------------------------------------------------

4. Discussion

Précédemment, on a insinué que notre objectif était d’afficher les groupes de mots de chaque centroïde qui sont des synonymes ou des groupes de mots ayant un lien tout en affichant à quelle classe de mot ils appartiennent. Ce but fut atteint du moment que l’on ait conscience que les résultats affichés peuvent être faussés par le poids du vote des k plus proche voisins, le nombre de points qu’on utilise pour le vote ainsi que l’emplacement des mots. On voulait aussi voir si le nombre de centroïdes a une conséquence sur le nombre d’itérations et sur le résultat. Les observations du au graphique de chaque expérimentation permet d’insinuer hypothétiquement qu’on ne peut prédire le nombre d’itération que cela peut prendre. Néanmoins, on voit tout de même que pour la majorité, les plus petits changements se font au fur et à mesure que le nombre d’itérations augmente. Aussi, plus qu’il y a de centroïdes, plus que l’on s’éloigne de cette image de la fonction exponentielle et que les changements dans chaque itération sont plus variés. Ensuite, par rapport aux résultats, l’itération à vingt centroïdes a sorti dans le premier centroïde, 9 verbes sur 10 mots, malgré le fait que le knn n’est pas fonctionné comme l’on voudrait et qu’il ait sorti certain en tant que nom. Le centroïde quatre semble former au moins quatre mots ayant un lien entre eux. En comparant à l’expérience deux avec 50 centroïdes à la première, on se rend compte que les résultats du premier centroïde n’ont aucun lien sauf les six mots sur dix qui finissent par -ant. Le reste des centroïdes n’affichent rien d’intrigant. Pour l’expérience à cent centroïdes, c’est le même résultat du point de vue que les résultats n’ont aucun lien entre eux sauf le deuxième où l’on a que des verbes et que le KNN ait fonctionnée. Après ces résultats, on est prêt à affirmer que notre hypothèse de base est approuvée sur le fait que l’on ne peut trouver des synonymes et que les k plus proches voisins n’est pas la méthode qui permet le mieux d’atteindre nos objectifs, surtout que l’on a eu majoritairement des résultats indésirables. Enfin, si l’on pouvait faire quelque chose autrement, ça serait surtout de soit utiliser le KNN d’une autre façon ou pour un autre objectif, soit ne pas l’utiliser complètement. Pour la K-moyenne, elle est utile pour certains domaines, mais non pour la langue quelle qu’elle soit. Car, il est dur de réduire des règles qui existent depuis des milliers d’années voire plus à un programme informatique. Il est plausible d’y arriver, comme l’a prouvé la méthode du k- moyenne pour réunir une classe de mot ensemble. Mais, on ne peut trouver des synonymes seulement avec des cooccurrences, la K-moyenne ou les k plus proches voisins.

1. PowerPoint VecteursCooccurrences, consulté le 18-12-2019 [↑](#endnote-ref-1)
2. PowerPoint VecteursCooccurrences, consulté le 18-12-2019 [↑](#endnote-ref-2)
3. PowerPoint KNN, consulté le 18-12-2019 [↑](#endnote-ref-3)