JESSICA CHAN ET SEBASTIAN PEREZ

420-C62-IN Données, mégadonnées et intelligence artificielle II

Groupe 00001

Travail pratique 3

Intelligence artificielle II en techniques de l’informatique

**RAPPORT DE LABORATOIRE**

Travail présenté à

M. Pierre-Paul MONTY

Cégep du Vieux Montréal

Mardi, 13 décembre 2022

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nombre centroïdes | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster4 |
| 25 | mores, bourgeois, ténèbres, jambes, fosses ,petits,  épaules, enfants, mineurs, bois  **nombre de mots : 10** | camarades, hommes, mousquetaires , halles, autres ,yeux , tous , sous , étaient ,dont  **nombre de mots : 10** | pour, avec, mais, lui, plus, sur, par, ils , qui, chevalier  **nombre de mots : 10** | imp, jouvet, furne, édit, raçon, cie  **nombre de mots : 6** | nouveau, montsou, venait, près, buckingham, beauté, île, rossinante, leurs, coups  **nombre de mots : 10** |
| 50 | préjudiciable, engagerait, transcrivit, apitoiement, acquiesçait, versent, disposât, stagnation, gagnerais, accommodera  **nombre de mots : 10** | décupler, cadencement, territoire, nique, sustente, escarpé, prieure, chevalerie\_,  pendirent, dindon  **nombre de mots : 10** | \_philis\_,  égrossir, scrofule,  avènement,  encombrement,  révocation,  islamisme, longanimité,  mésintelligence,  nonchalance  **nombre de mots : 10** | compagnie, fosse, porte, manche, route, tête,  maheude, normande,  terre, lisa  **nombre de mots : 10** | cherchions  morfondons,  priâmes, déjeunerons,  rappelions, marchions,  retrouvions, professons,  espérions, mordicus  **nombre de mots : 10** |
| 75 | porthos, aramis, milady, bonacieux, athos  **nombre de mots : 5** | abrutissement, hochets, fruitière, navigation, choisies, sommets, encaissement, boitait, confection, contraction  **nombre de mots : 10** | ça, nous, cardinal,  tu, maître, héros, grâce  **nombre de mots : 7** | taille, bande, vertu, table, princesse, bouche, gouvernante, nièce, grève, mine  **nombre de mots : 10** | amidon, manèges, cafards, dizains, coquetteries, salées, araignées, compotiers, couperets, pastilles  **nombre de mots : 10** |
| 100 | résolut, passait, dut, arrivait, leva, fallut, quoiqu, aperçut, regarda, toutefois  **nombre de mots : 10** | telle, grande, bonne, pierre, fois, heure  **nombre de mots : 6** | planchet, catherine, rossinante, duc, esprit, curé, felton  **nombre de mots : 7** | remerciée, remerciera, dota, mourût, rentrât, abîmera, préférât, assidûment, baveuse, surmontait  **nombre de mots : 10** | dû, voulu, vu, eu, donné, pu, pris, ait  **nombre de mots : 8** |

**Ce tableau et les suivants partagent les propriétés suivantes : Taille de fenêtre : 5** **nombre maximum de mots par clusters : 10**

Dans le cadre du laboratoire nous avons décidé d’étudier, tout d’abord, l’impact du nombre de centroïdes sur nos résultats. Nous avons donc choisi d’inspecter le contenu des 5 premiers clusters de différents ensembles de données avec un nombre de centroïdes variable. Notre lexique a été recueilli à partir des 4 textes : Don Quichotte de Miguel de Cervantes, Germinal de Émile Zola, Les trois mousquetaires d’Alexandre Dumas et Le ventre de paris d’Émile Zola

En premier lieu, nous allons faire une analyse profonde de l’ensemble de données avec 25 centroïdes car nous trouvons la variété de ses clusters particulièrement intéressante. Afin de d’alléger le rapport qui dépasse déjà le nombre de pages demandées, nous allons seulement faire l’analyse profonde d’un ensemble des données.

## Analyse profonde de l’ensemble de données avec 25 centroïdes

Tout d’abord, cet ensemble des données possède des champs lexicaux relativement clairs dans ses clusters et c’est pour cela que nous l’avons choisi pour faire une analyse en profondeur. Lorsque nous regardons le premier ensemble de données, le cluster 0 regroupe un champ lexical propre à Germinal dans lequel nous pouvons observer la réalité décrite par l’auteur des enfants exploités qui travaillent comme des mineurs. On remarque ici un cluster peuplé principalement par un seul de nos textes.

Si l’on observe maintenant le cluster 1, le champ lexical qui se dessine est plus proche des Trois mousquetaires avec des mots comme « camarades », « hommes » et « mousquetaires ». D’autres mots qui indiquent l’appartenance à un tout sont également présents comme « tous » et « autres » propres à l’esprit de camaraderie du livre.

Le cluster 2 contient principalement des prépositions à l’exception du mot « chevalier » qui vient probablement de Don Quichotte. Ce mot étant le seul de sa catégorie est également celui qui est le plus loin de son centroïde. Cela illustre le peu de connexion entre ce mot et le restant de la population du cluster. Ce cluster est également celui qui a la plus grande de distance entre ses membres et son centroïde avec presque quatre fois plus de distance que les autres quatre clusters de cet ensemble des données.

Le cluster 3 contient des mots avec peu de rapport entre eux. Il est difficile d’émettre une hypothèse sur le lien entre les mots car à l’exception d’« édit » et « rançon » les mots du cluster ne figurent pas dans un dictionnaire. Ce cluster est également celui avec le moins de mots dans l’ensemble des donnes avec 25 centroïdes. Malgré le peu de sens entre les mots, leur distance est relativement semblable à celle du cluster numéro 1, sauf que, à différence du cluster 3, le cluster 1 a un champ lexical très clair.

Le cluster 4 prend des mots de plusieurs livres et les regroupe selon la thématique du voyage. « Montsou » étant le nom de la mine dans laquelle le personnage principal de Germinal se fait embaucher, « Buckingham » étant une ville d’Angleterre et « Rossinante » étant une référence au cheval de de Don Quichotte. De plus, des mots comme « île », « nouveau », « venait », « près » et « beauté » enrichissent la thématique du voyage à l’intérieur de ce cluster. À la suite du cluster 0, ce cluster est celui qui est le plus proche de son centroïde et il est intéressant d’observer ce résultat bien que les membres du cluster n’appartiennent pas à un même texte.

## Analyse de l’ensemble avec 50 centroïdes

L’ensemble de données avec 50 centroïdes est le seul à avoir ses cinq clusters complètement peuplés. Cela est curieux car l’ensemble de données précèdent a moins de centroïdes et par conséquent nous aurions cru que puisqu’il avait moins de clusters, ces derniers seraient plus volumineux. Nous nous sommes trompés dans notre hypothèse. Cet ensemble de données est de loin celui qui possède la plus petite distance entre ses clusters et leur centroïde respectif. Nous avons de la difficulté à trouver une hypothèse pour ce phénomène. La distance la plus grande d’un mot à son centroïde dans cet ensemble de données étant inférieure à 10 000.

Les champs lexicaux sont moins clairs que dans l’ensemble précédant également. Le lien le plus significatif entre les mots du cluster 0 est la classe grammaticale de la plus grande partie du cluster. Presque la totalité de sa population sont des verbes. Ce phénomène se reproduit aussi avec le cluster 4. Nous pensons que cela se doit au fait que les mots se dissipent un peu plus dû au nombre de centroïdes qui augmente.

## Analyse de l’ensemble avec 75 centroïdes

L’ensemble de données avec 75 centroïdes est celui qui a le cluster le plus petit avec une population de 5 pour le cluster 0. Étant donné que le nombre de centroïdes augmente nous pensons qu’il est raisonnable que certains clusters diminuent en taille. Cet ensemble est également celui qui a le mot le plus éloigné des quatre ensembles de données que nous avons analysés à présent. Ce mot est : « grâce » avec une distance supérieure à 80 000 dans le cluster 2. Nous pensons que cela est simplement un hasard étant donné que nous avons choisi les 5 premiers clusters de chaque ensemble de données que nous avons étudié.

Le cluster 1, 3 et 4 ont tous des distances très basses bien que le mot le plus loin de son centroïde se trouve dans cet ensemble de données. Les clusters 0 et 2 se trouvent un peu plus loin que les autres clusters de cet ensemble mais, en général, à l’exception du mot « grâce », les distances sont meilleures que le cluster 2 de l’ensemble de données avec 25 centroïdes. Nous pensons que cela se doit au fait que les clusters se font plus petits et par conséquent, les données aberrantes se déplacent dans un cluster qui leur est plus approprié.

## Analyse de l’ensemble avec 100 centroïdes

L’ensemble de données de 100 centroïdes est celui qui possède la plus grande fluctuation au niveau du nombre de mots par clusters avec le cluster 1, 2 et 4 avec une population de 6,7 et 8 respectivement. Nous pensons que cela est normal car cet ensemble des données est celui avec le plus des centroïdes. Ainsi, plus il y a des centroïdes, plus la population générale de chaque cluster peut varier et par le fait même, diminuer de façon. Les deux autres clusters ont chacun une population de 10. Les distances entre les mots des clusters et leur centroïde respectif sont très basses. Les plus basses en général après celle de l’ensemble avec 50 centroïdes.

Nous retrouvons également dans cet ensemble des données le même phénomène que dans l’ensemble de 50 centroïdes. Dans le cluster 0, 3 et 4 au moins 80% de leur population sont des verbes et plus précisément des verbes au passé. Nous pensons que cela fait du sens étant donné que le temps de verbe le plus utilisé pour raconter une histoire est le passé.

## Expérience de notre cru

Pour l’expérience de notre cru nous avons implémenté un KNN. Dans le cadre du cours d’intelligence artificielle I nous avons codé un KNN en équipe de 4 mais notre sous-équipe a surtout développé la vue de ladite application. Alors, nous avions une idée de comment l’algorithme fonctionnait mais nous ne l’avions pas formellement codé. Nous avons donc profité de cette occasion pour avoir la chance de coder nous-mêmes notre propre KNN. Il y a eu des concepts qui ont été plus difficiles à comprendre notamment la position des mots dans un espace n-dimensionnel déterminé par la matrice de cooccurrence mais nous avons réussi à les maitriser et implémenter notre KNN.

## Réflexion sur les expériences

Ce cours est le premier contact que nous avons avec une quantité de données si grande. Ainsi, nous avons appris à travailler avec une telle quantité et éventuellement nous étions en mesure d’utiliser ces données. Nos expériences nous ont permis de voir l’influence du nombre de centroïdes dans notre algorithme de K-means. Elle nous a également permis d’analyser et de d’interpréter nos données afin de les présenter et les contextualiser. La réalisation d’un projet de cette envergure fut une première expérience dans laquelle nous avons énormément appris.