*** Projet MapReduce***

***« Healthcare Analytics »***

*Réalisé par :*

Riahi Farah

ING2 Info

**Introduction :**

Dans le domaine crucial de la santé, comprendre les relations entre différentes maladies est essentiel pour améliorer les soins.

Ce projet explore l'application de l'algorithme MapReduce dans cette optique, en utilisant un jeu de données spécialement conçu pour tester son efficacité.

J’ai volontairement fourni le jeu de données pour évaluer la capacité de MapReduce à analyser les occurrences et les fréquences des différentes maladies ainsi que des symptômes associés.

Cette approche prometteuse offre une gestion optimale des données massives, permettant des analyses parallèles et distribuées. En mettant en lumière les associations médicales significatives, cette étude vise à démontrer l'applicabilité de MapReduce dans la prise de décision dans le domaine de santé.

**Objectifs du projet :**

1. **Calculer la Fréquence des Maladies :**
   * Analyser la répartition des occurrences de différentes maladies dans le jeu de données.
   * Identifier les maladies les plus fréquentes pour mettre en évidence les problèmes de santé dominants.
2. **Identifier les Maladies Dominantes :**
   * Déterminer les maladies prédominantes en termes de fréquence.
   * Fournir des informations cruciales sur les principales préoccupations de santé au sein de la population étudiée.
3. **Analyser les Co-Occurrences de Maladies :**
   * Étudier les relations entre différentes maladies pour identifier les co-occurrences fréquentes.
   * Mettre en lumière les associations médicales significatives pour comprendre les liens entre les affections.
4. **Évaluer les Maladies Liées :**
   * Identifier les maladies qui peuvent potentiellement causer ou influencer l'apparition d'autres maladies.
   * Offrir des perspectives sur la manière dont certaines conditions médicales peuvent être des facteurs de risque pour d'autres.
5. **Analyser les Symptômes Liés aux Différentes Maladies :**
   * Explorer les données pour établir des connexions entre les maladies et les symptômes associés.
   * Fournir des informations précieuses sur les manifestations cliniques communes et les symptômes partagés entre différentes affections.**Haut du formulaire**

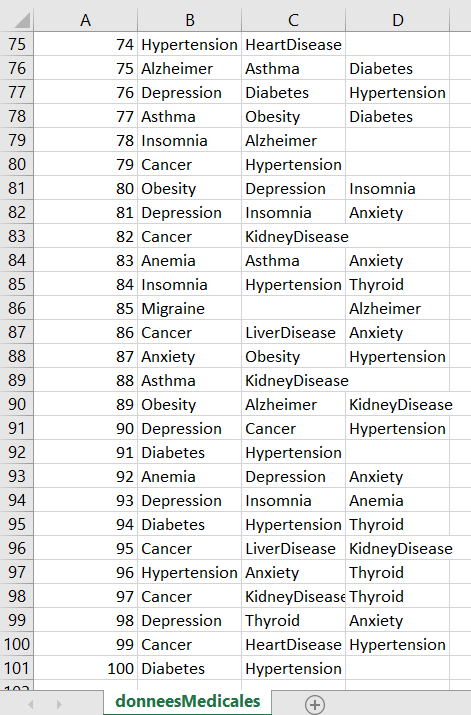
**Partie 1 :**

Préparation du jeu de données :

Imaginons un ensemble de données simplifié pour illustrer notre approche, comprenant 100 patients avec une variété de trois maladies différentes.

Cette modélisation simplifiée nous permettra de démontrer concrètement comment notre analyse, basée sur l'algorithme MapReduce, peut être appliquée pour extraire des informations

pertinentes sur la fréquence des maladies, les maladies dominantes, les co-occurrences, et les symptômes associés.

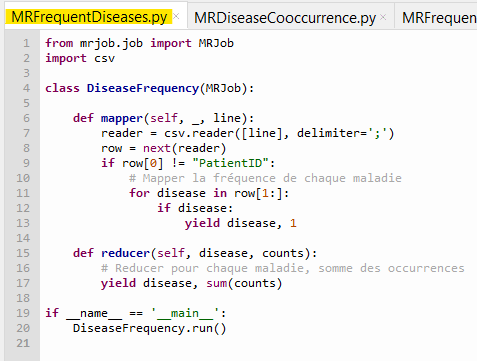


Objectif 1 :

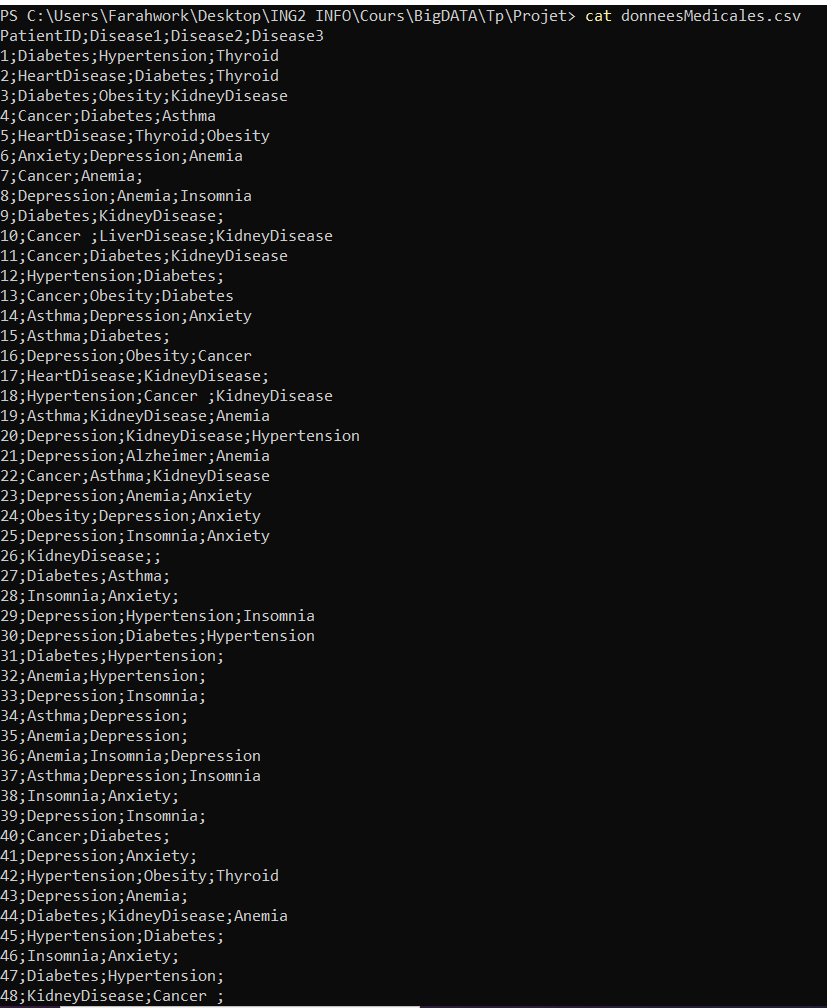
L’objectif de cette partie :

* Compter la fréquence de chaque maladie déterminer les maladies dominantes.
* Déterminer les cooccurrences des maladies.

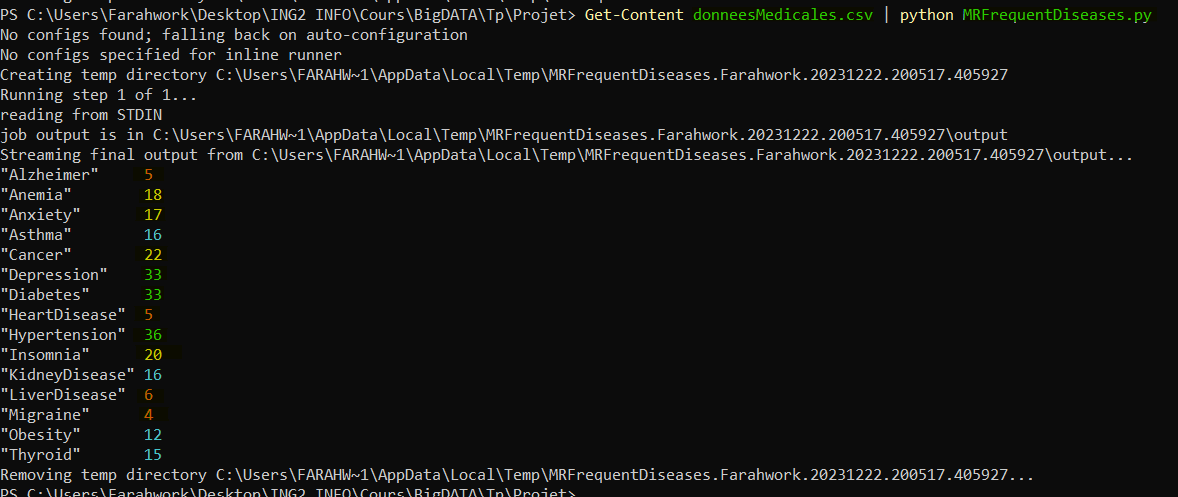
1. **Code MapReduce :**



* Explication de code :
  + La méthode mapper est appelée pour chaque ligne du fichier d'entrée.
  + Le CSV « donneesMedicales.csv » est lu ligne par ligne, et les valeurs sont séparées par des points-virgules.
  + On ignore la première ligne (l'entête) en vérifiant que la première colonne n'est pas "PatientID".
  + On itère ensuite sur les maladies (à partir de la deuxième colonne) et on émet un couple clé-valeur pour chaque maladie rencontrée avec une valeur de 1.
  + La méthode reducer est appelée pour chaque clé unique émise par les mappers.
  + Elle reçoit la clé (maladie) et la liste des valeurs associées (1 pour chaque occurrence).
  + Elle émet ensuite la maladie avec la somme des occurrences, fournissant ainsi la fréquence totale de chaque maladie.
  + Vérification délimiteur ;

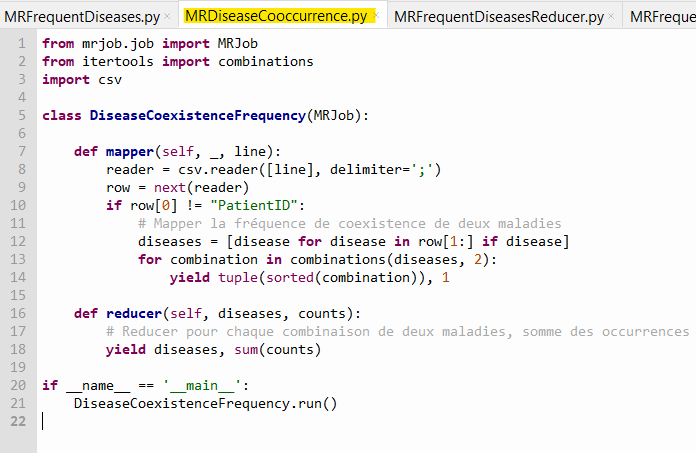


Exécution :

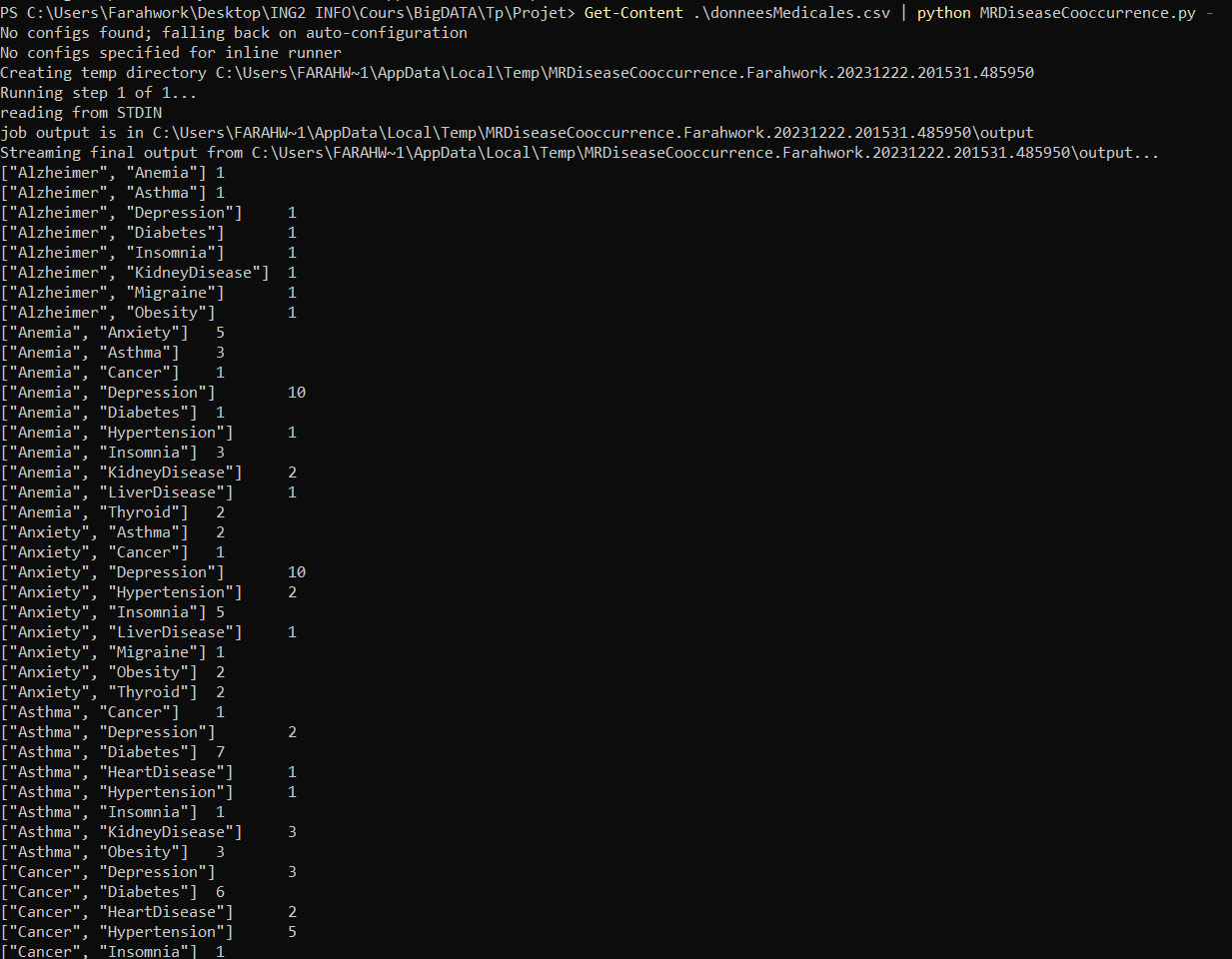


* En interprétant ces résultats, on peut observer que, par exemple, "Hypertension" a une fréquence de 36, indiquant qu'elle est présente 36 fois dans le jeu de données. De même, "Depression" et "Diabetes" ont tous deux une fréquence de 33, suggérant qu'elles sont également des conditions médicales fréquentes dans cet ensemble de données. Ce type d'information peut être extrêmement utile pour comprendre les tendances médicales dominantes dans la population étudiée.

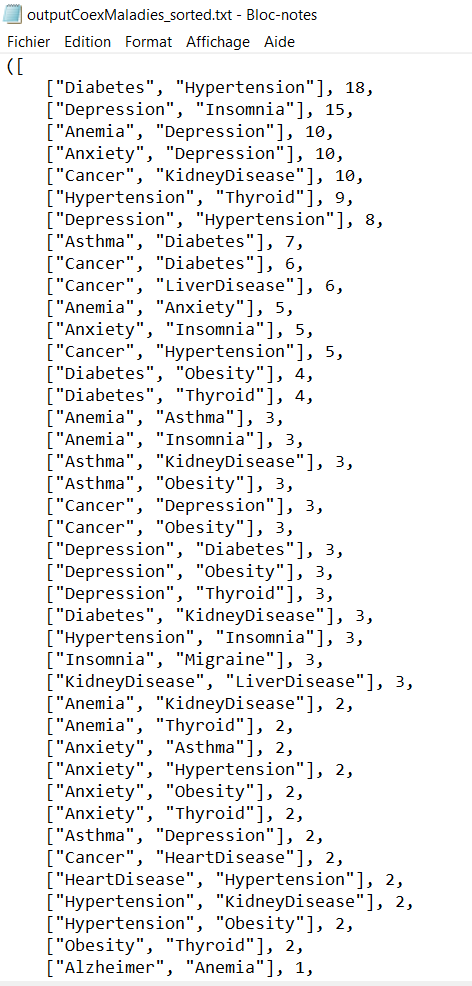
1. **Code MapReduce :**

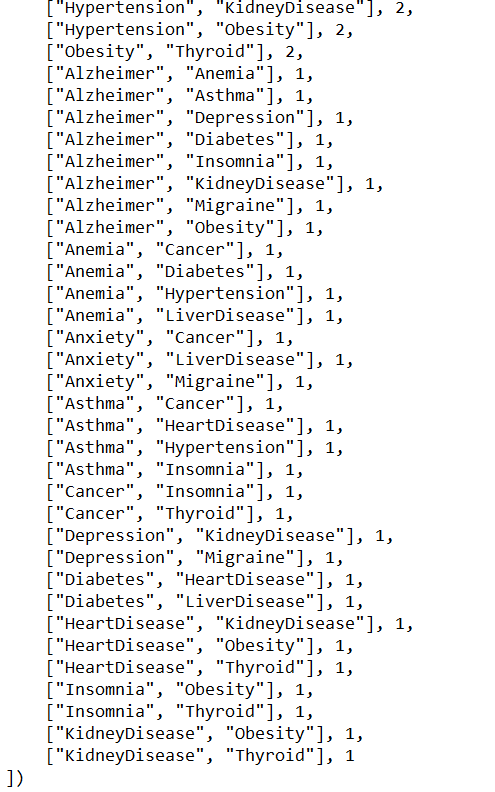


* Explication :
  + Le mapper crée une liste diseases des maladies pour le patient actuel
  + Génère toutes les combinaisons possibles de deux maladies et émet un couple clé-valeur pour chaque combinaison rencontrée avec une valeur de 1.
  + Le reducer reçoit la clé (combinaison de deux maladies) et la liste des valeurs associées (1 pour chaque occurrence).
  + Emet ensuite la combinaison de maladies avec la somme des occurrences, fournissant ainsi la fréquence totale de coexistence de ces deux maladies.

Exécution :

* + L’output sera mis dans un fichier outputCoexMaladies.txt pour réaliser un tri décroissant en fréquence de coexistence.
  + Le résultat est affiché dans ce fichier outputCoexMaladies\_sorted.txt.



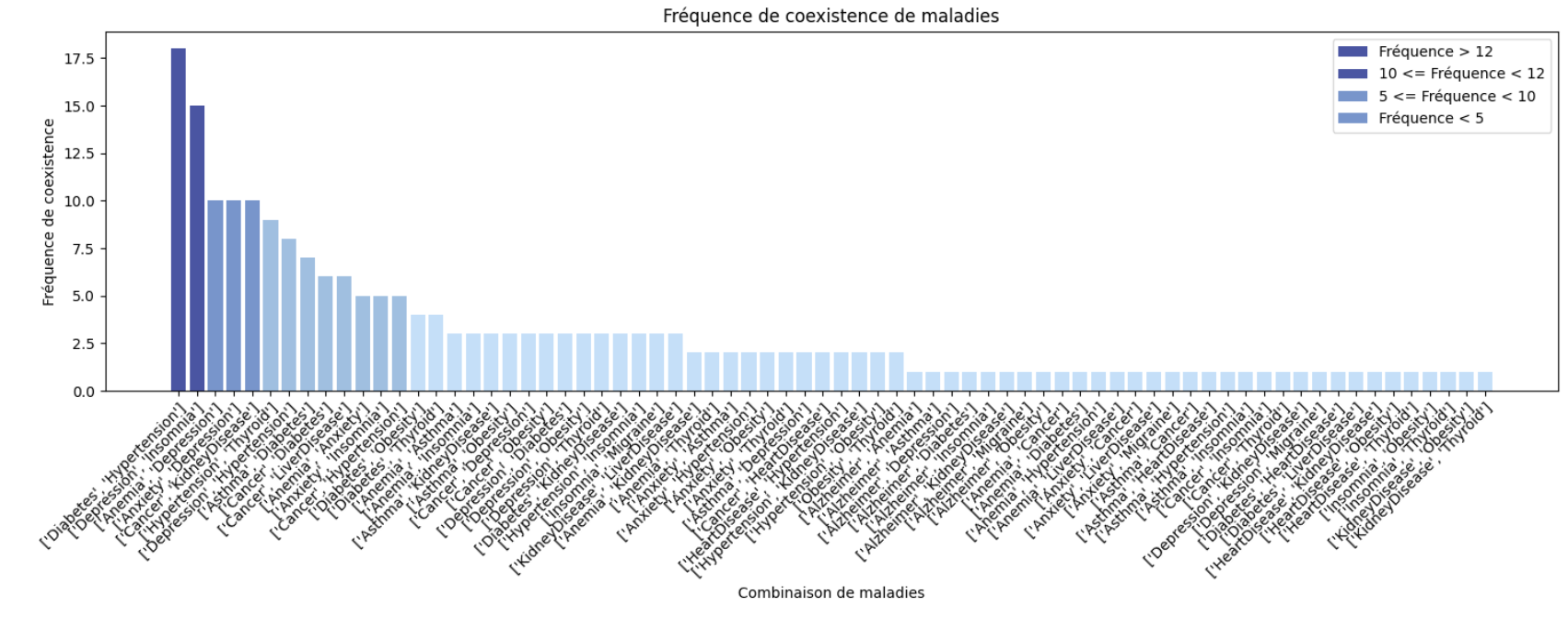


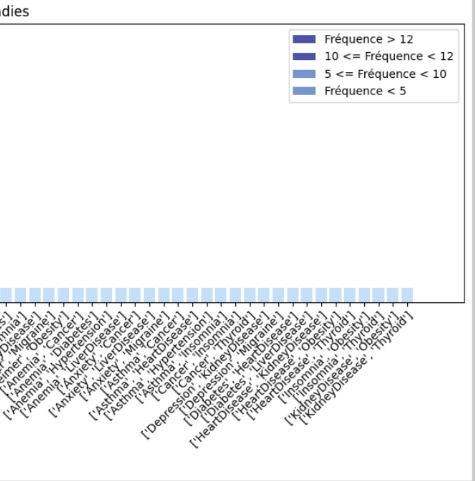
* Les exemples les plus fréquents :
  + Hypertension + Diabete :

L'hypertension et le diabète sont deux conditions médicales courantes associées à des facteurs de risque similaires, tels que le mode de vie, l'alimentation et la génétique.

Les patients atteints de diabète sont plus susceptibles de développer une hypertension, et vice versa, en raison des liens complexes entre la régulation du glucose et de la pression artérielle.

* + Dépression + Insomnia :
  + La dépression et l'insomnie sont souvent liées, formant un cercle vicieux où l'une peut aggraver l'autre.
  + Les troubles du sommeil, tels que l'insomnie, sont fréquemment observés chez les personnes souffrant de dépression en raison de changements dans les schémas de sommeil et de la détérioration de la santé mentale.



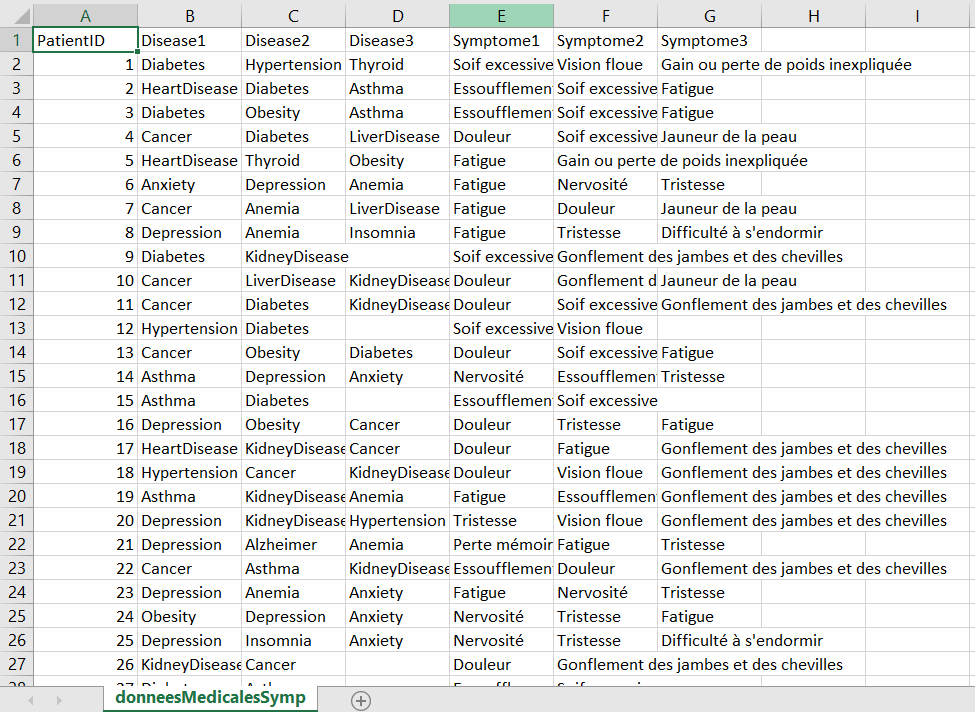


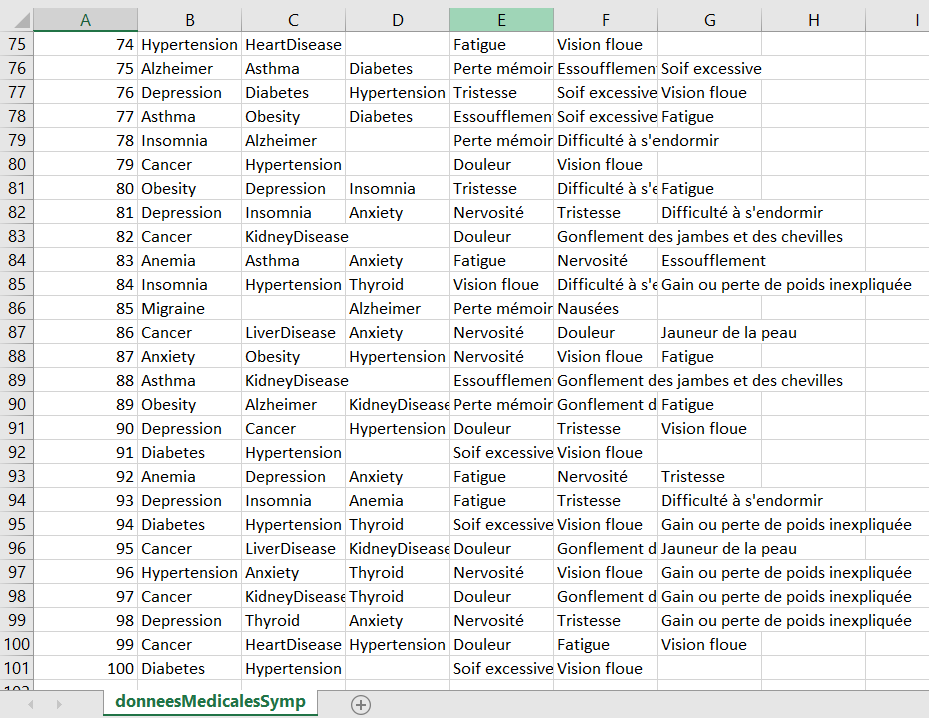
**Partie 2 :**

Utilisez MapReduce pour analyser la cooccurrence de différents symptômes dans les données médicales. Cela pourrait aider à identifier des associations entre certaines maladies et symptômes .

Préparation du jeu de données :

On ajoute quelques symptômes occurrents pour chaque patient.



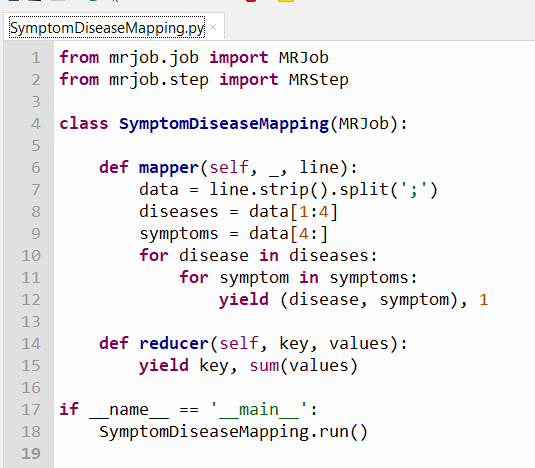


Objectif 2 :

L’objectif de cette partie :

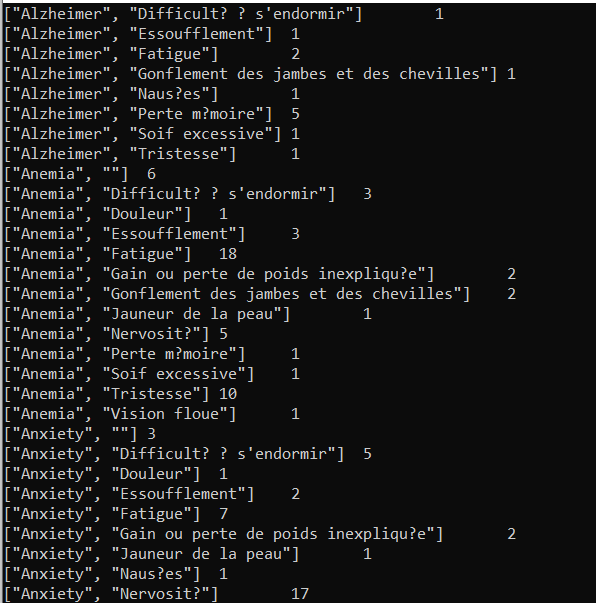
* Lier les symptômes les plus occurrents aux maladies associées.

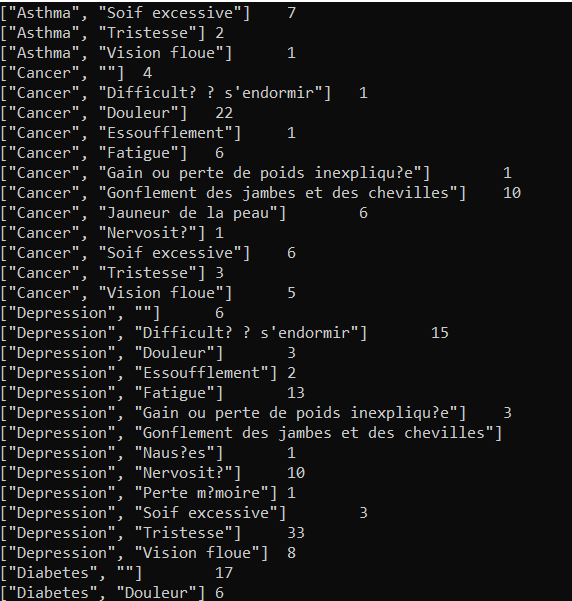
1. **Code MapReduce :**

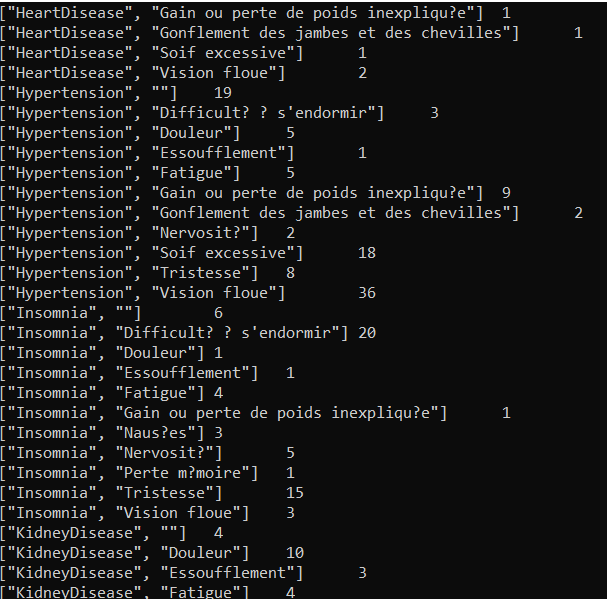


* Explication :
  + Le mapper extrait les informations sur les maladies (3 premières colonnes) et les symptômes (colonnes restantes).
  + Emet un couple clé-valeur pour chaque paire de maladie et symptôme rencontrée avec une valeur de 1.
  + Elle reçoit la clé (paire maladie-symptôme) et la liste des valeurs associées (1 pour chaque occurrence).
  + Elle émet ensuite la paire maladie-symptôme avec la somme des occurrences, indiquant le nombre de fois où cette paire a été observée dans le jeu de données.

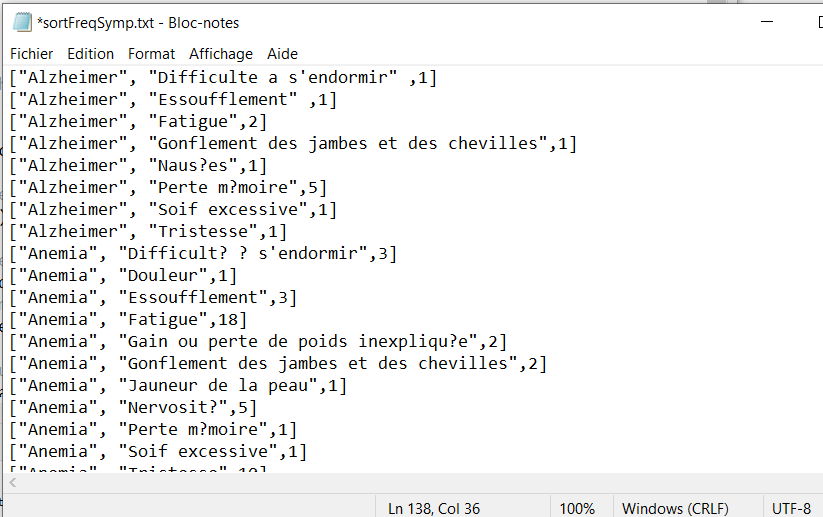
Exécution :

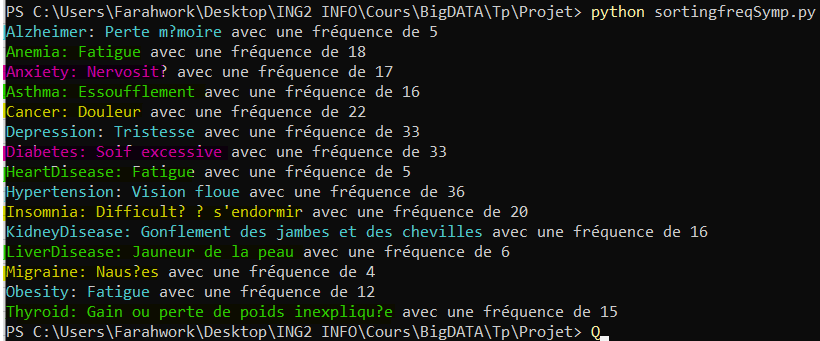




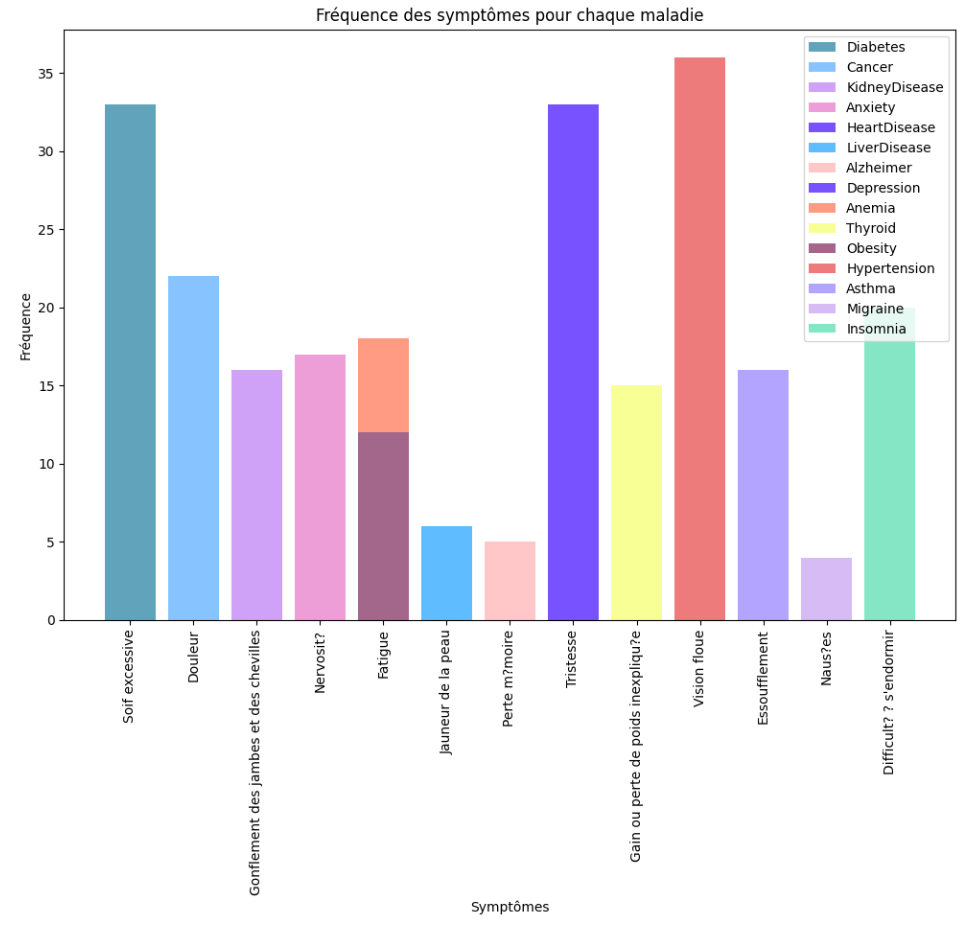


* + sortFreqSymp.txt pour réaliser un tri décroissant en fréquence de chaque couple : maladie : symptôme.





* Ces résultats fournissent des informations précieuses sur les associations entre les maladies et les symptômes dans le jeu de données médicales, aidant ainsi à mieux comprendre les manifestations cliniques courantes associées à chaque condition médicale.



Par exemple pour l’hypertension :

La Vision floue avec une fréquence de 36

* La vision floue est le symptôme le plus fréquemment associé à l'hypertension.
* Et ainsi de suite pour le reste des maladies.

Finalement et pour clarifier, les données analysées dans ce projet ne sont pas issues d'une source officielle et sont fournies par moi-même à des fins de test. Ces données sont utilisées dans le cadre de l'expérimentation et de la démonstration des concepts liés à l'application de l'algorithme MapReduce dans le domaine de la santé.

L'utilisation de données factices à des fins de test permet de développer et de valider les méthodes de traitement sans dépendre d'ensembles de données réels et sensibles.

Merci pour votre attention.

N'hésitez pas à me contacter pour toute question supplémentaire.