## 

| **Equipe cancer - LIVRABLE 3** |
| --- |

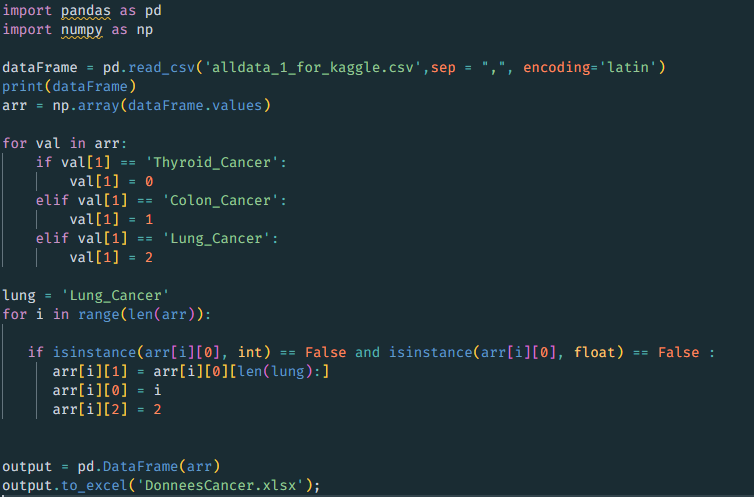
## **Analyse du jeu de données:**

**Extraction des données**

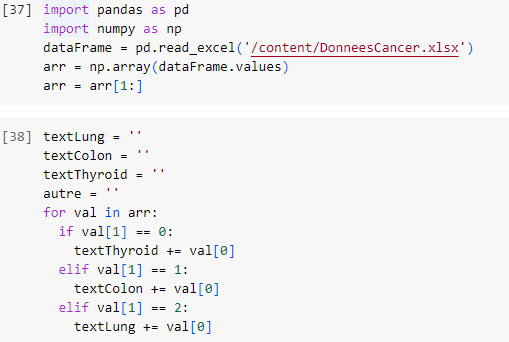
Étant donné que le fichier csv présentait de nombreux trous ou problèmes de formatage, nous avons dû réorganiser nos données avant de pouvoir les utiliser.

Pour cela nous avons utilisé un programme python et power query afin de nettoyer la donnée et d’avoir directement une forme X (les textes) et y (les types de cancer).

Nous avons aussi décidé de remplacer les noms de cancer par des indices.



Ensuite nous avons importé le fichier excel dans colab afin d’effectuer les wordclouds et les analyses.



**Traitement des données**

Lien du Google Colab : (Je vais bientot le changer car colab de test )****

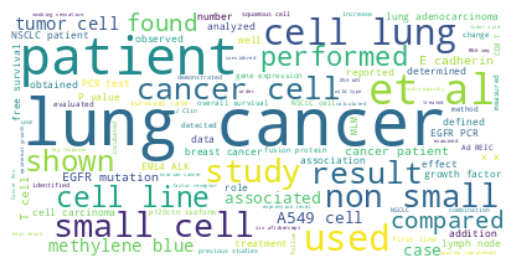
[Livrable3.ipynb](https://colab.research.google.com/drive/1-y-vNx-p9qAI44VvxwTEf5LhXxK5G6fd?hl=fr#scrollTo=_IT3NlfN6Z7Y)

Au cours de notre exploration du jeu de données nous avons utilisé la librairie des WordCloud pour identifier les termes prédominants associés aux textes. Ces visualisations nous ont permis de mettre en évidence les aspects les plus importants du jeu de données.

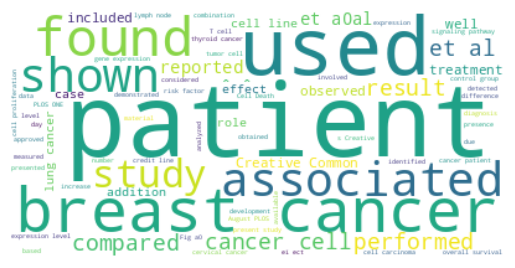
**Wordcloud : Colon**



**Wordcloud : Lung**



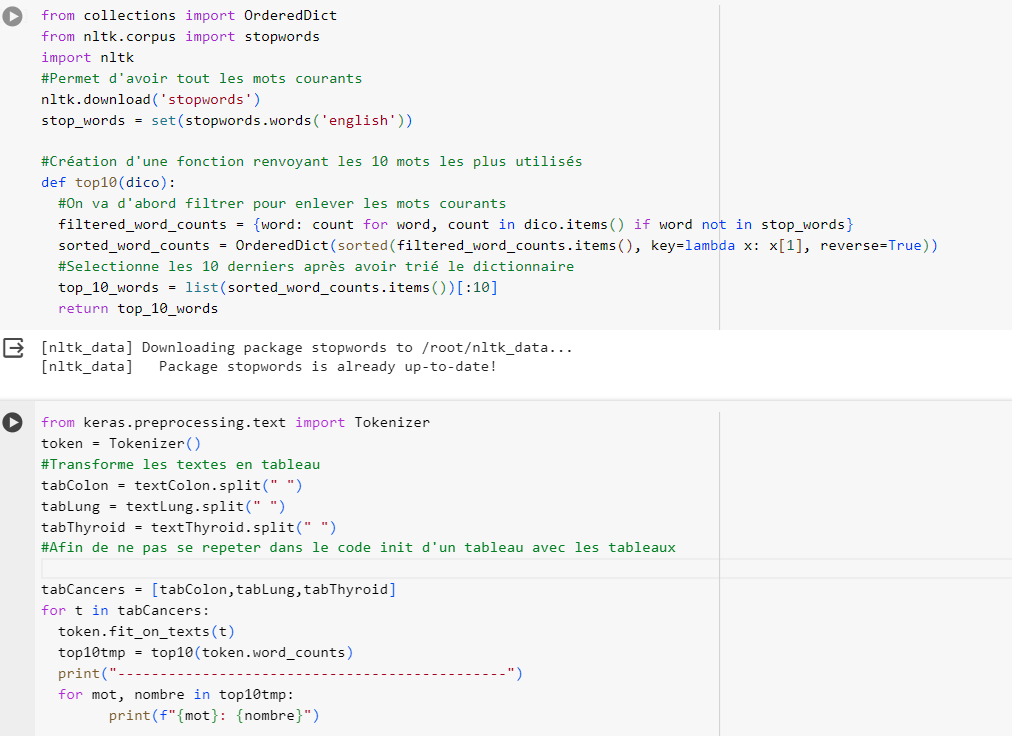
**Wordcloud : Patient**



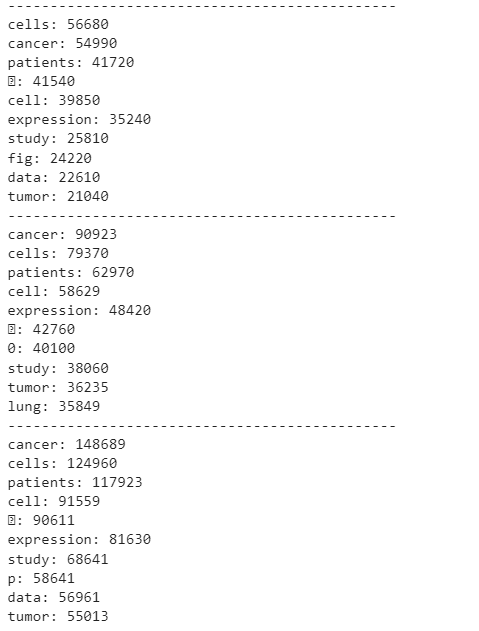
Le Word Cloud nous a permis de mettre en évidence les éléments clés du jeu de données de manière visuelle. Cette approche a facilité l'identification des tendances, les mots plus fréquents dans le Word Cloud vont nous aider à formuler des questions pertinentes.

En plus du word cloud nous avons essayé de représenter les 10 mots les plus fréquents à l’aide d’un diagramme en bâton l'objectif était de fournir l’objectif était de fournir un aspect plus formel et statistique.

Pour cela nous avons utilisé keras et nltk (Natural Language Toolkit) afin d’avoir nos 10 mots.



Après l’analyse des 10 mots les plus fréquents nous avons remarqué que les mêmes mots revenaient pour les 3 types de cancer.



Les résultats indiquent que les trois types de cancer (colon, poumon et thyroïde) ont des similitudes dans les termes les plus couramment utilisés pour décrire les cas. Les mots utilisés pour décrire les trois types de cancer peuvent indiquer des tendances ou des caractéristiques communes.

## **Pour la suite :**

Nous avons également essayé de faire un fit avec quelques textes du dataset cependant nous avons vite remarqué que l'exécution ne pouvait aboutir en raison d'une limitation de la RAM. Ce qui sera à surveiller lors du fine tuning du modèle