

Master Sciences et Technologies de l'Information et de la Communication
STIC-B545 - Traitement automatique de corpus

Enseignants : Max De Wilde

Assistant : Louis de Viron

1ère session - Décembre 2022.

Rapport - TP4

L'exploration spatiale et la photographie dans le corpus CAMille

MAHIANT Alice

Table des matières

1	Modalités	3
2	Introduction	4
2.1	Choix du sujet	4
2.2	Constitution du corpus	4
2.3	Embûches : un avant-goût	5
3	Analyse	6
3.1	Répartition des fichiers selon les années et selon les journaux	6
3.1.1	Résultats et difficultés d'analyse	6
3.1.2	Quelques évènements relatifs à la photographie des astres	8
3.2	Etude des termes les plus fréquents	9
3.2.1	Résultats	9
3.2.2	Constitution de la liste de stopwords	10
3.3	Détection des hapax et mots les plus longs : indice de qualité de l'ocr plus qu'outil de recherche pertinent	10
3.4	Clustering : catégorisation des fichiers d'une année spécifique en 5 groupes distincts .	10
3.5	Words embeddings	11
4	Conclusion	12

Modalités

Sur la base des éléments méthodologiques et des enseignements techniques présentés lors du cours théorique, il est demandé dans le cadre de ce TP :

- de mobiliser les connaissances et compétences acquises tout au long du cours ;
- d'apporter un regard critique sur le traitement automatique de corpus en général ;

Les étapes à mettre en œuvre sont les suivantes :

- Choisissez une thématique très ciblée qui vous intéresse ;
- Rendez-vous sur <https://www.camille-ulb-kbr.be/>
> effectuez une requête en l'affinant avec des filtres (journal, années, etc.) ;
- Une fois la requête suffisamment précise, exportez les résultats en ZIP et/ou XLSX.
> attention, seuls les 1000 premiers résultats seront exportés ! ;
- Étudiez votre thématique de manière transversale dans votre sous-corpus en vous aidant des différentes techniques vues au cours :
 - exploration ;
 - fréquences ;
 - mots-clés ;
 - entités nommées ;
 - sentiment analysis ;
 - clustering ;
 - word2vec ;
 - etc.
- Si vous le souhaitez (optionnel), vous pouvez également faire appel à d'autres techniques qui n'ont pas été vues au cours pour enrichir votre travail ;
- Tout au long de l'analyse, prenez note de ce qui fonctionne bien ou moins bien afin de déterminer les avantages et les limites du traitement automatique de corpus dans un contexte historique ;
- Rédigez un rapport de 10-15 pages comprenant :
 - une introduction ;
 - une analyse ;
 - une conclusion ;
 - une courte bibliographie.
> votre document doit être compréhensible pour quelqu'un qui ne connaît pas le contexte du cours, par exemple un journaliste ou un historien intéressé par l'exploitation du corpus CAMille.
- Publiez votre code (notebooks) sur GitHub dans un dossier «tp4» et soumettez votre rapport sur l'UV au format PDF.

Introduction

Choix du sujet

Pour ce travail de traitement automatique de corpus, j'ai voulu travailler sur la question de la photographie de la Lune et de la Terre depuis l'espace. L'idée de départ était d'étudier comment la photographie était traitée avant qu'elle ne fasse partie intégrante de la presse papier de par, entre autres, l'apparition du photo-reportage. Parlait-on de photographie et d'évolutions techniques sans pouvoir montrer ces photographies ? Comment en parlait-on ? S'agissait-il d'articles à caractère purement scientifique ?

La photographie en tant que telle me semblait néanmoins trop vaste et j'ai donc décidé de me concentrer sur un sujet qui serait à coup sûr traité par la presse, à savoir l'exploration spatiale. Les photographies de la Lune - et plus tard de la Terre ou encore de Mars et Saturne vues de l'espace - ont en effet fait couler beaucoup d'encre. Je pense notamment à la première vue complète de la Terre vue de l'espace, une image publiée par la NASA en 1967 et obtenue par la combinaison de deux images satellites. Cette image avait fait l'objet d'une véritable campagne de la part de Stewart Brand, qui en fera d'ailleurs la couverture du dernier numéro du *Whole Earth Catalog*. Je songe également à "*Earth Rise*" prise en 1968 par l'astronaute William Anders qui sera même utilisée comme timbre poste, ou encore à "*The Blue Marble*" bien que celle-ci date de 1972, année qui n'est pas comprise dans le corpus CAMille. Ces clichés sont devenus des symboles du mouvement écologiste et ont marqué les esprits de l'époque. Si ceux-ci me sont familiers et me fascinent particulièrement, les ayant étudiés lors de mon cursus à l'ERG, je ne connaissais que très peu l'histoire des clichés antérieurs, s'il en est une. Il existe en effet d'autres images plus anciennes qui se sont vues vite dépassées par les progrès techniques des années 60 mais me permettent de faire des recherches sur l'ensemble du corpus CAMille et pas seulement sur la dernière décennie disponible.

Constitution du corpus

Pour constituer mon corpus, j'ai décidé de ne pas utiliser de filtres. Pour étudier les changements de discours sur la photographie avant et après son implémentation dans la presse il me fallait en effet, comme explicité plus haut, étudier l'ensemble du corpus CAMille. Pour ces mêmes raisons, je n'ai pas appliqué de filtres sur les journaux sélectionnés, ce qui sera détaillé plus loin dans l'analyse lorsque nous nous intéresseront à la répartition des journaux dans le corpus. Pour constituer mon corpus, j'ai donc sélectionné les fichiers correspondant à la recherche suivante :

terre AND (photo OR photographie OR photos OR photographies OR cliché*)
AND (planète OR planètes) AND lune

J'avais initialement repris "daguerriotype" comme alternative à "photo" mais cela n'affectait pas les résultats, je l'ai donc retiré. L'ajout du mot-clef "planète" m'a permis

de mieux cibler les articles qui m'intéressaient en écartant les simples prévisions météorologiques, passant de 20 268 résultats à 1632 fichiers. J'ai exporté ces fichiers par tranches de 50 ans afin de ne pas dépasser la limite de 1000 exportations simultanées.

Cependant, après exploration du corpus, entre autres en étudiant les mots les plus fréquents, il m'est apparu que peut-être ce mot-clé "planète" avait été mal choisi et devrait s'accompagner d'alternatives correspondant au champ lexical de l'exploration spatiale. J'avais également oublié de mentionner la possibilité de reprendre le mot "lunaire" comme alternative à "lune".

J'ai donc fait une étude parallèle sur les 1000 premiers fichiers des 3 229 qui apparaissaient en lançant la recherche suivante :

terre AND (photo OR photographie OR photos OR photographies OR cliché*) AND (lune OR lunaire) AND (planète OR planètes OR satellite* OR fusée* OR spatial OR astronaute OR astronomie)

Embûches : un avant-goût

Dès la constitution du corpus, j'ai été confrontée à différents problèmes dont trois principaux sur lesquels nous reviendrons plus en détail dans la suite de ce travail. Le premier consiste en la distribution hétérogène des différents journaux sur la totalité des années représentées dans CAMille. Le second grand problème concerne la qualité de l'OCR, qui demande soit d'anticiper les différentes fautes orthographiques possibles pour tous les mots-clefs que l'on reprendrait dans la requête constituant le corpus, ce qui est pratiquement infaisable, soit de simplement ignorer ces mots mal orthographiés et donc perdre de l'information. Le troisième principal obstacle relève de la séparation des fichiers en pages de journal et non en articles, ce qui nécessairement inclut un grand nombre d'informations non pertinentes au corpus, noyant les autres.

Analyse

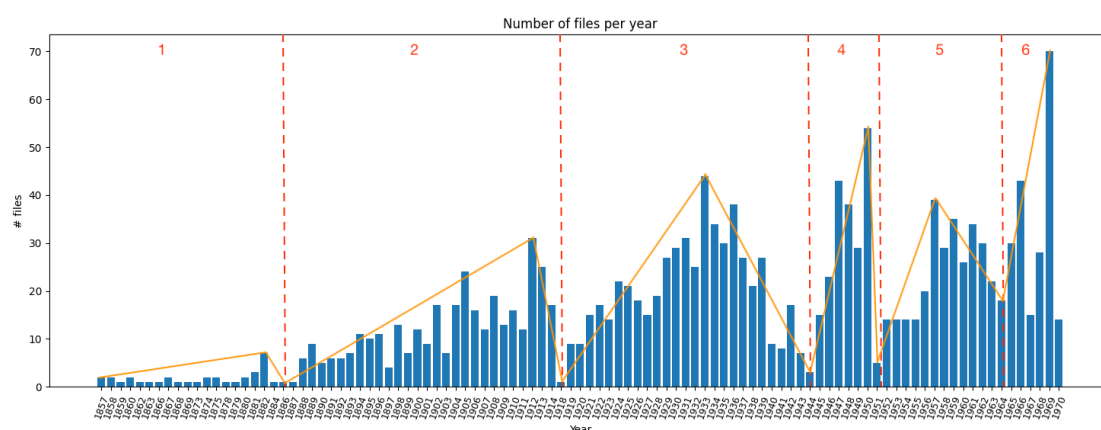
Répartition des fichiers selon les années et selon les journaux

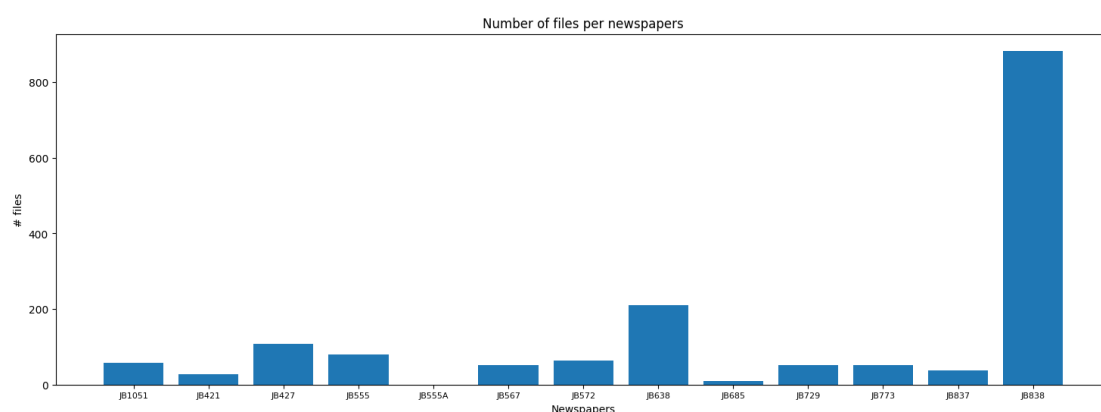
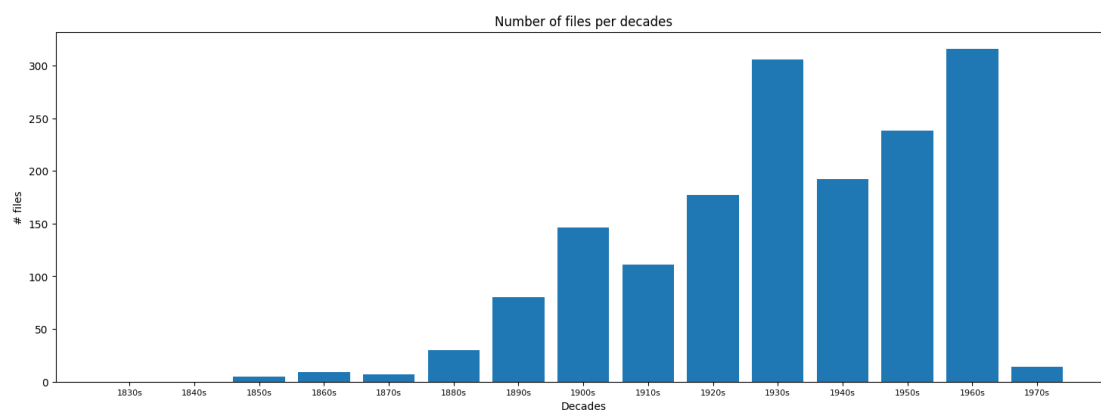
Résultats et difficultés d'analyse

Pour le premier corpus constitué, on constate sur le graphique de distribution des fichiers selon les années 6 blocs plus ou moins marqués qui forment des vagues progressives. Ces résultats sont difficiles à analyser, tout d'abord parce que je ne suis pas arrivée à représenter les années "vides" sur le barchart. De plus, ces résultats doivent être confrontés à la distribution, hétérogène, des différents journaux selon les années. Toute les années ne sont pas fournies uniformément dans le corpus CAMille, ce qui rend ce genre de statistiques peu pertinentes. En effet, sur les 12 journaux repris dans le corpus, 10 d'entre eux ne sont plus représentés à partir de 1950. Ainsi, seul *Le Soir* couvre la période allant jusqu'à 1970, le *Drapeau Rouge* couvrant pour sa part la période allant jusqu'en 1966. Le biais est donc énorme sur les statistiques que l'ont pourrait faire sur l'évolution de la présence des mots-clefs recherchés selon les années. Pour contrer ce biais, j'avais initialement pensé ne me concentrer que sur le journal *Le Soir* et proposer une analyse plus ciblée. Malheureusement celui-ci s'est créé en 1887, or je voulais prendre en compte dans mon étude la période précédant l'apparition de photographies dans la presse.

Sur le dernier graphique, présentant la distribution des fichiers selon les journaux, on constate sans surprise que *Le Soir* se démarque fortement avec plus de 50% des fichiers lui étant attribués face à 12% pour *la Meuse* et 6,5 % pour *la Libre Belgique*, si l'on ne reprend que les trois journaux les plus représentés.

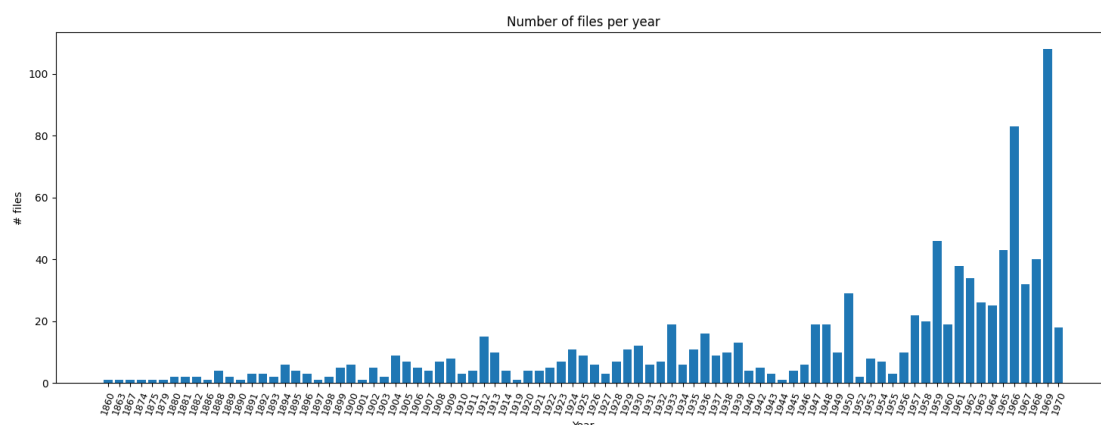
Distribution des fichiers du corpus 1

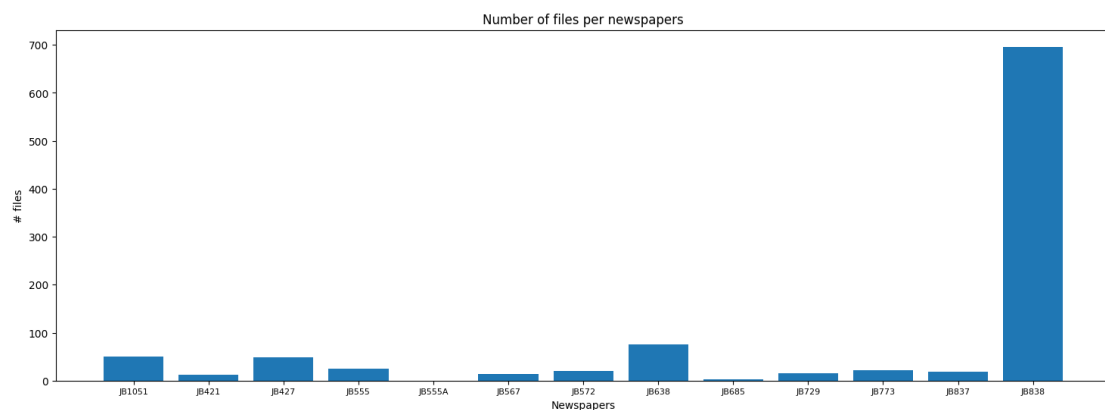
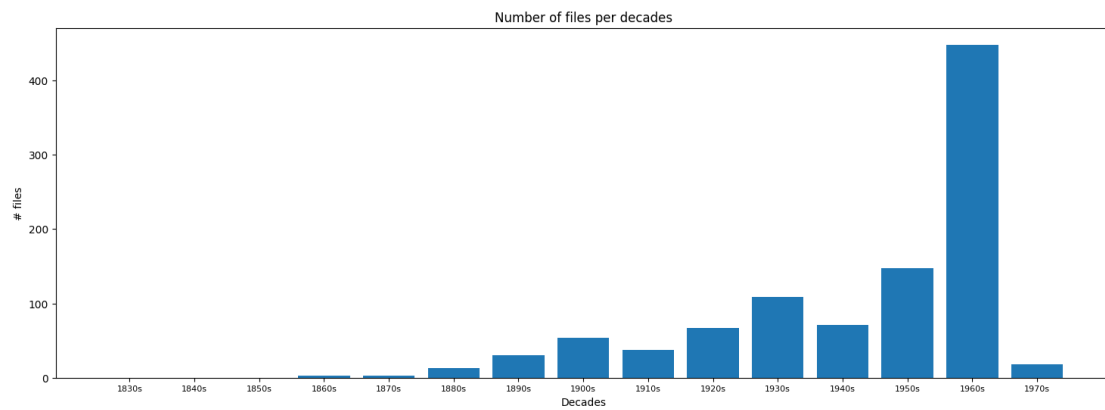




En analysant les résultats du second corpus, j'étais étonnée de voir que les résultats se rapportaient plus fortement aux évènements notables liés à l'exploration spatiale et la photographie listés plus bas, dont la majorité ont eu lieu dans les années 60, avec un moindre nombre d'articles pour les années précédentes que pour le premier corpus. Néanmoins, je pense que ces chiffres sont dus à la sélection partielle faite sur les documents. En effet, pour ce second corpus je n'ai sélectionné que les 1000 premiers fichiers et ce selon une classification par pertinence. Concernant la représentation des journaux, *Le Soir* domine toujours largement le corpus, avec près de 70% des fichiers, suivi là encore par *la Meuse* et la *Libre Belgique* avec respectivement 7,6 et 4,9%.

Distribution des fichiers du corpus 2





Quelques évènements relatifs à la photographie des astres

Bien que ces données soient à traiter avec précaution, on peut néanmoins voir que la représentation des termes se recoupe avec les évènements clefs de l'exploration spatiale, avec un pic notable début des années 30, puis un second en 1950, un troisième plus discret en 1958 et un dernier en 1969. Pour plus de clarté, j'ai dressé ci-après une courte liste des évènements qui me semblaient pertinents en termes de photographie et d'exploration spatiale :

- 1840 : un des premiers clichés de la Lune, en daguerréotype, par John W. Harper
- 1946 : première photographie de la Terre vue de l'espace par un missile V2 récupéré aux allemands et modifié par les USA ;
- 1957 : lancement du satellite Sputnik 2 ;
- 1958 : création de la NASA ;
- 1958 : exposition universelle de Bruxelles ;
- 1959 : première photographie de la Terre en orbite par le satellite Explorer-6, lancé par la NASA ;
- 1959 : première photographie de la face cachée de la Lune par Luna-3, envoyée par l'URSS ;
- 1964 : premières images de Mars depuis Mariner-4, envoyé par la NASA ;
- 1965 : première image et vidéo d'un Homme (Alexei Leonov) dans l'espace lors de la mission soviétique Voskhod-2 ;

- 1966 : première image de la Terre en disque plein par le satellite ATS-1 lancé par la NASA ;
- 1967 : premier cliché couleur de la Terre en disque plein obtenu par la combinaison des images de DODGE et de l'ATS-3, utilisé comme illustration de couverture du Whole Earth Catalog ;
- 1968 : multiples clichés capturés par le crew de la mission Apollo 8, dont "*Earthrise*" ;
- 1969 : première photographie d'un Homme marchant sur la Lune lors de la mission Apollo-11 ;
- 1969 : première image d'une éclipse solaire avec la Terre capturée par le crew de la mission Apollo-12 ;
- 1972 : "*The Blue Marble*", photographiée par le crew de la mission Apollo-17.

Etude des termes les plus fréquents

Résultats

Pour le premier corpus testé, les 5 premiers mots les plus fréquents du corpus sont, dans l'ordre :

terre > grande > paris > lune > premier.

Ces mots sont suivis d'un ensemble de mots se rapportant au monde politique, avec notamment *président*, *gouvernement* ou encore *ministre*.

Après voulu étudier spécifiquement les années qui se détachaient sur le premier graphique de distribution des fichiers sur les années, à savoir 1912, 1933, 1950, 1957, 1966 et 1969.

- 1912 : soleil > lune > moment > grande > point ;
- 1933 : paris > cap > mars > grande > france ;
- 1950 : lune > mars > grande > général > vie ;
- 1957 : satellite > terre > premier > grande > monde ;
- 1966 : mars > lune > premier > terre > première ;
- 1969 : lune > apollo > terre > lunaires > astronautes

En réalisant quelques recherches, j'ai appris qu'en 1912 avait eu lieu une rare éclipse solaire totale, observable en Europe et photographiée par de nombreux scientifiques et amateurs de l'époque. L'année 1933 est l'année dont les résultats sont les plus vagues, les 100 mots les plus fréquents ne faisant quasi pas mention d'astronomie et encore moins de photographie. L'an 1966 est, assez logiquement d'ailleurs, celui le plus clairement marqué par la thématique, les 50 premiers mots les plus fréquents s'y rapportant quasi tous.

Constitution de la liste de stopwords

Le principal problème de cette recherche de mots les plus fréquents consiste en l'application d'une liste de stopwords qui très vite devient lourde de conséquences et qu'il convient dès lors de manipuler avec précaution. J'ai ainsi progressivement créé diverses catégories, en plus des stopwords habituels tels que les déterminants, pronoms et abréviations relatives aux petites annonces de logement (qui sont légion dans le corpus), que j'ai combiné de manières différentes pour étudier les résultats produits. Un des problèmes que j'ai pu rencontrer concerne le double sens de certains mots tels que "mars", qui dans le cadre de l'astronomie est très à propos, mais n'est pas pertinent dans son emploi comme facteur temporel (le mois de *mars*). J'ai donc testé avec et sans prendre en compte les autres mois de l'année dans la liste de stopwords, ce qui n'a pas modifié le résultat, m'assurant ainsi que le terme "*mars*" était bien utilisé pour parler de la planète rouge.

Un autre grand problème concerne les noms des différents engins envoyés dans l'espace, qui sont malheureusement retirés lors de l'application de la fonction de nettoyage. En effet, ceux-ci non seulement se constituent souvent de tirets et de chiffres, telle que les différents engins du programme soviétique Luna, mais sont parfois des termes de deux caractères, tel que le missile V-2. Retirer cette condition de la fonction de nettoyage me permettrait de les récupérer effectivement au sein du corpus mais cela ne ferait que polluer encore plus le texte du corpus.

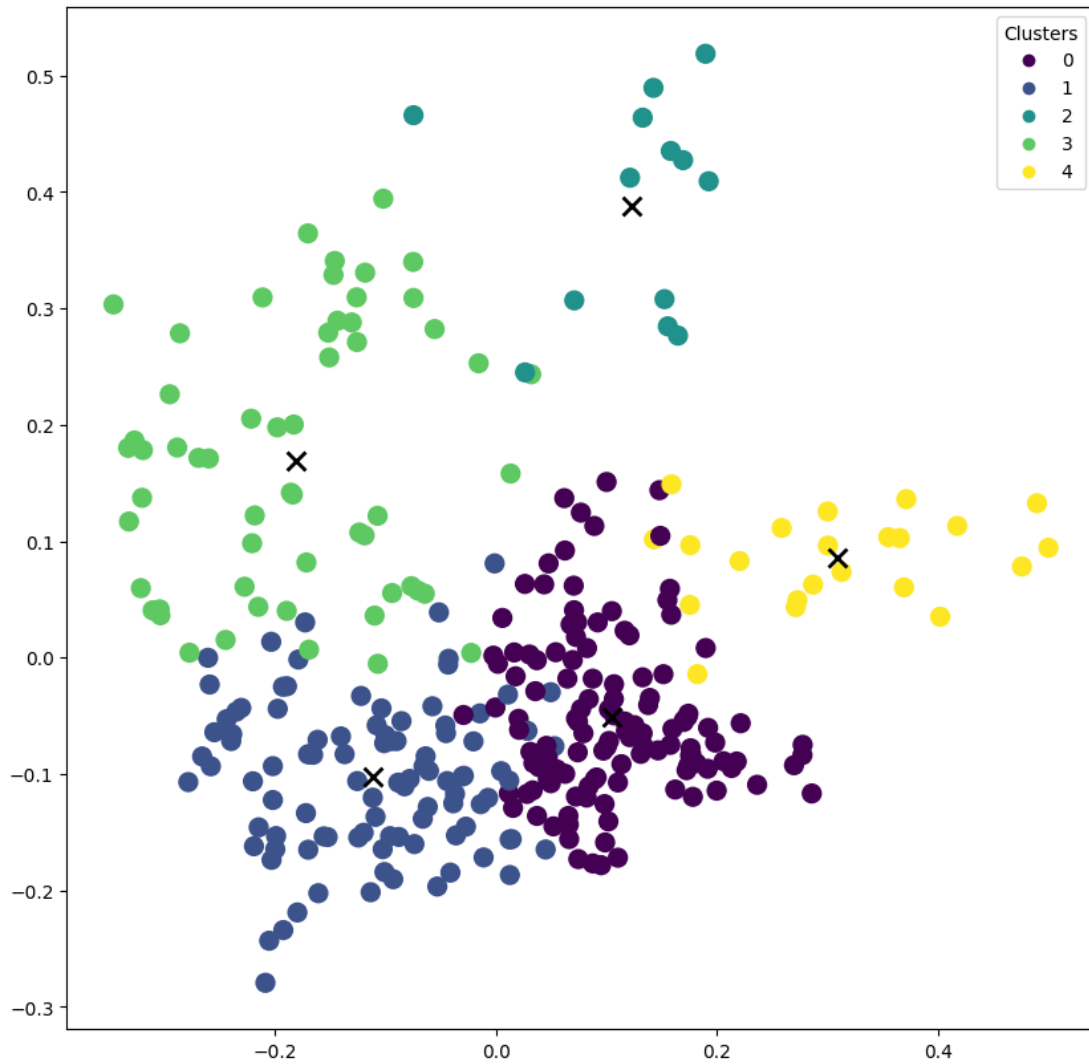
Détection des hapax et mots les plus longs : indice de qualité de l'ocr plus qu'outil de recherche pertinent

Les hapax ainsi que les mots les plus longs détectés dans le corpus permettent de rapidement rendre compte de la qualité de l'OCR utilisé lors du traitement des journaux. En effet, sur les 50 premiers hapax, seuls 6 sont des mots correctement orthographiés. En ce qui concerne les mots les plus longs, il faut parvenir au 997e mot pour obtenir un terme qui aie quelque peu du sens, à savoir "*chromatophotographie*."

Clustering : catégorisation des fichiers d'une année spécifique en 5 groupes distincts

Pour l'année 1966, on remarque que les fichiers peuvent être classés en 5 clusters différents. Il m'est toutefois difficile d'établir en quoi cette visualisation peut m'être utile dans le cadre de ma question de recherche.

Répartition des fichiers de l'année 1966



Words embeddings

Des observations que j'ai pu faire, celle qui m'a paru la plus intéressante concerne le fait que "cliché" est bien plus proche de "lune" que "photographie".

Conclusion

Je n'ai pu, par manque de temps et pour cause de maladie, écrire un rapport complet des essais réalisés sur le clustering et les modélisations word2vec reprises dans mon code. En conclusion, la qualité de l'OCR nuit véritablement à l'utilisation du corpus CAMille dans le cadre de la recherche. La riche présence de termes techniques mais également d'abréviations fait de la thématique des découvertes spatiales un cas particulièrement intéressant pour tester les limites dudit corpus. La découpe en pages de journal, et non en articles, noie les termes pertinents au sein d'un grand nombre d'articles qui ne le sont pas. Il aurait fallu appliquer une catégorisation selon les rubriques, ce qui est loin d'être une mince affaire.

Au sujet de ma thématique de départ, il apparaît que la photographie ne fait, sans réelle surprise, que peu parler d'elle dans le cadre de l'exploration spatiale au sein du corpus CAMille dans les années précédant l'inclusion quasi systématique de photographies dans la presse.