

Faculté de Lettres, Traduction et Communication

## Traitement automatique de corpus

STIC-B545

TP 4 : Travail final

Vankerckhoven

Vicky

000551082

MSTICS

<https://github.com/VickyVKV>

Année académique 2022-2023

## Table des matières

<b>Introduction</b>	<b>3</b>
<b>Choix du thème</b>	<b>3</b>
<b>Mise en place et résultats CAMille</b>	<b>3</b>
<b>Keyword</b>	<b>4</b>
<b>Nuage de mots</b>	<b>5</b>
<b>Entités nommées</b>	<b>5</b>
<b>Analyse de sentiments</b>	<b>7</b>
<b>Clustering</b>	<b>8</b>
<b>Conclusion</b>	<b>9</b>
<b>Bibliographie</b>	<b>10</b>

## Introduction

Un programme est en cours de financement par le gouvernement fédéral belge en vue de développer l'histoire du journalisme belge et mettre en place un système d'archive. Pour cela, la KBR (Bibliothèque royale de Belgique) et l'ULB se sont associées.

Notre travail est de ressortir des informations intéressantes du corpus CAMille. CAMille est le centre d'archives sur les médias et l'information, cette base de données reprend les journaux belges de 1830 jusqu'à 1970 (pour le moment) dans un but de recherches journalistiques. Pour ce dernier travail, je vais m'attarder sur les articles concernant la protection des animaux de 1900 à 1925.

Avant tout, qu'est ce qu'un corpus ? Un corpus regroupe un ensemble de textes d'une thématique choisie. Ici, le corpus défini est la protection des animaux. Cette étude va analyser les différentes informations que nous pouvons retrouver dans ce corpus à l'aide de nuage de mots, de mots clés et d'autres méthodes d'analyse textuelle.

## Choix du thème

J'ai décidé de travailler sur les articles concernant la protection des animaux de 1900 à 1925. J'ai établi une durée de 25 ans, le corpus contenant trop d'articles à analyser il a fallu délimiter une période donnée. Attention, pour certaines années j'avais une erreur 500 sur le site de CAMille.

Le sujet m'intéresse car dans les années 1900, il n'y avait aucun droit pour les animaux et aucune loi sur leur protection. Cependant, il existait des ASBL pour la défense de ceux-ci.

Il faut savoir qu'à cette époque, les animaux étaient considérés comme des objets. Ce n'est qu'après la guerre de 1914-1918 que les personnes ont commencé à se poser des questions sur le sort des animaux. En effet, les chiens ont été utilisés dans les tranchées lors de la première guerre mondiale et certains soldats s'y sont attachés. Aujourd'hui les droits des animaux sont en évolution, les familles accordent de plus en plus d'attention à leurs animaux de compagnie et les associations de protection animalière sont en perpétuelle évolution. Il y a encore du progrès à faire mais par rapport aux années 1900, nous pouvons observer une nette amélioration.

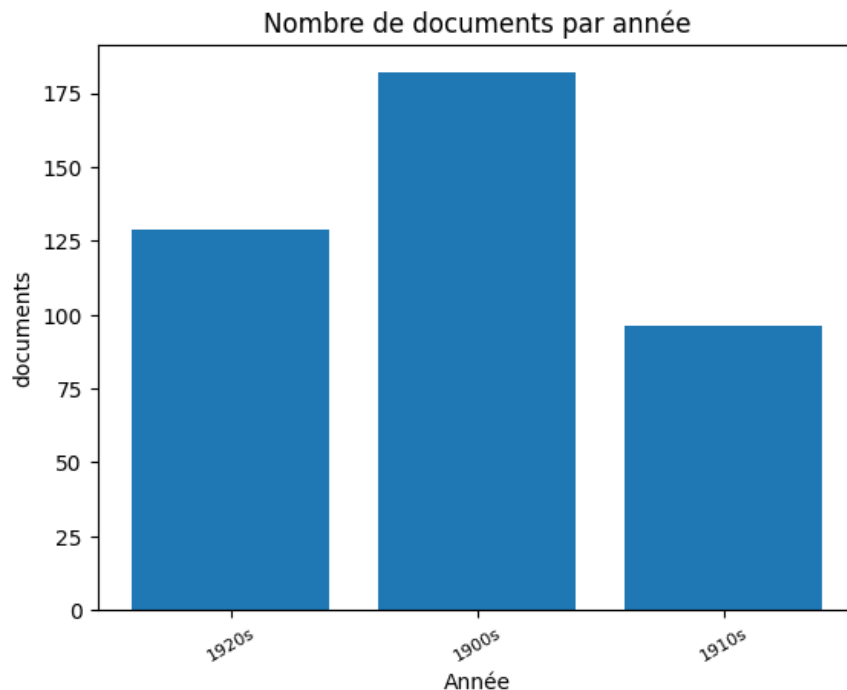
## Mise en place et résultats CAMille

Une fois mon thème décidé, j'ai fait mes recherches sur le corpus CAMille afin de voir combien de résultats en ressortaient. Pour ma durée de 25 ans, s'étalant de 1900 à 1925, j'ai trouvé 407 résultats. Il a donc fallu extraire ces résultats en zip, me donnant 407 fichiers texte. Une fois mes fichiers décompressés, je suis passée par python pour en faire un fichier all.txt reprenant tous les textes.

J'ai pu voir qu'il me manquait 4 années dans mon fichier : 1916, 1917, 1918 et 1919. Je peux également observer qu'il y a 2 journaux qui sortent du lot avec plus de fichiers que les autres : le journal de Charleroi et Le Soir. Pour ceux présentant le moins de résultats, il s'agit de L'Indépendance belge (édité en Angleterre) et Le Drapeau Rouge. Quant à elle,

l'année représentant le plus de résultats est 1925 avec 51 fichiers. Cela pourrait s'expliquer par la création de l'ASBL CROIX BLEUE qui fut fondée en 1925. Mais nous le découvrirons peut-être via nos analyses.

Nous pouvons voir via le graphique suivant, le nombre de fichiers par décennie. Bien que 1925 soit l'année comprenant le plus de fichiers textes, la décennie 1900 comprend le plus de documents.



## Keyword

Pour sortir les mots et expressions du texte, j'utilise Yake. Les bigrammes ressortis principalement sont les suivants :

'Société belge', 'Société royale', 'Conseil communal', 'sociétés belges', 'grand nombre', 'Paris Paris', 'gouvernement belge', 'JEUNE HOMME', 'cours d'une', 'Congo belge'.

Nous pouvons trouver une logique pour : Société Royale, en effet cela correspond à la Société royale protectrice des animaux qui aide les animaux en détresse depuis 1863. C'est une société belge donc cela pourrait également expliquer le bigramme 'Société belge' et 'sociétés belges'. Pour ce qui est de 'Conseil communal' et 'gouvernement belge', les droits des animaux sont énormément passés par des conseils communaux ainsi que par le gouvernement etc avant d'arriver à quelque chose.

Pour le bigramme 'Congo belge', je pense à la convention pour la protection des animaux au Congo en mai 1900. Cela a donc toujours un intérêt dans notre analyse et nous prouverait que les journaux en ont informé la population.

Par contre, pour les bigrammes 'JEUNE HOMME', 'grand nombre', 'cours d'une', 'Paris Paris', ils n'apportent rien à mon analyse. Effectivement, les trois bigrammes ne sont pas assez explicites et ne me permettent pas d'identifier les informations précises qu'ils apportent.

Afin d’extraire mon nuage de mots, j’ai créé un fichier “protection\_clean.txt”. Ainsi je n’ai plus de caractères spéciaux qui peuvent poser problèmes lors de la génération du nuage de mots.

[illegible]

Nous pouvons observer aussi “Prix” qui pourrait être lié au futur prix Nobel de la paix de Ludwig Quidde qui était un des principaux opposants à la vivisection.

Lors de l'extraction des entités nommées, le logiciel SpaCy a pu ressortir les différentes personnalités :

- 5

Pour les autres personnalités, je ne sais pas à qui elles réfèrent, retrouvant donc : Ides L, B CAMARADE, LE SUCCESSEUR DE, M. Max parmi les propositions. Après quelques recherches, je n'ai rien trouvé de concluant.

Au niveau des lieux ressortis, nous en retrouvons une assez belle quantité passant de pays à des villes voir même à des rues :

- Londres
- Bruxelles
- France
- Everaert
- rue Blaes
- Paris
- Grand-Duché de Luxembourg
- Etat luxembourgeois
- province d'Anvers
- Bruxelles-Anvers
- UNE EXPOSITION EN 1930

Nous pouvons voir dans notre logiciel que Londres a été nommée 3 fois, Bruxelles et France 2 fois sur notre corpus. Pour les autres, ils ont été nommés qu'une fois.

Il y a cependant des noms qui ne me disent rien et dont je n'ai pas d'informations tels que Clyde, Everaert, Huilera. Ainsi que la Société Prince-Henri qui correspondrait plus à une organisation qu'à un lieu. Ainsi que Suissesses qui correspond plus à une population qu'à un lieu.

Enfin, voici les organisations qui sont apparues :

- Syndicat des Employés
- Syndicat des Cuirs et Peaux
- Cabinet de Conseil
- Conseil
- Comité
- Havas

Nous retrouvons également Angleterre qui devrait se trouver dans les lieux ainsi que ministre des Travaux qui devrait plutôt être dans les personnalités.

Mis à part ça, SpaCy a assez bien analysé le corpus pour les entités en retrouvant plusieurs d'entre elles, une meilleure analyse serait toujours possible. Peut-être qu'un autre logiciel permettrait une analyse plus poussée.

## Analyse de sentiments

Pour l'analyse de sentiments, j'ai choisi 10 phrases à analyser comportant des liens avec la protection des animaux. L'analyse de sentiments consiste en l'interprétation des sentiments dans les données textuelles.

Je vais donc retranscrire les résultats dans un tableau reprenant le pourcentage de subjectivité et de polarité (positive, négative, neutre).

Analyse de sentiments		
	Polarité	Subjectivité
1	18% positif	0.33999999999999997%
2	12% négatif	0.3333333333333333%
3	neutre	parfaitement objectif
4	neutre	0.15%
5	neutre	parfaitement objectif
6	40% négatif	0.2%
7	neutre	parfaitement objectif
8	20% positif	0.2%
9	20% positif	0.275%
10	21% positif	0.0875%

Sur l'échantillon extrait, je peux déduire que le sujet de presse "Protection des animaux" est entre neutre et positif. L'échantillon s'est basé sur plusieurs années différentes. Pour ce qui est des polarités négatives, il s'agit d'extraits de conférences ou d'avis des Sociétés de protection des animaux envers certaines pratiques.

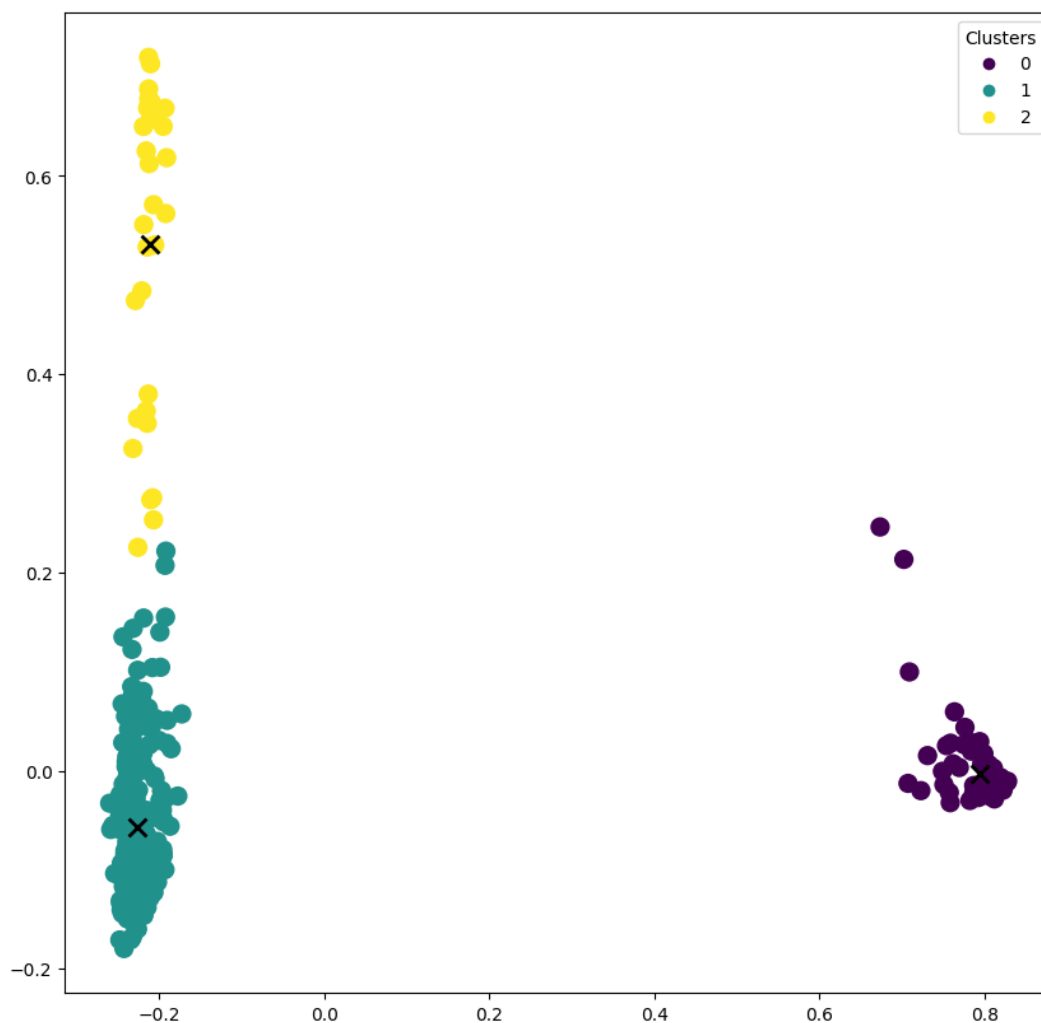
Vous pouvez retrouver les phrases sélectionnées dans le fichier `entites_sentiments.ipynb` se trouvant sur Github.

Nous pourrions aller plus loin dans l'analyse en utilisant le topic modeling (méthode non supervisée déterminant des thèmes). Cela permettrait de nous aider à découvrir la structure sémantique de la presse pour la protection des animaux. Le résultat donnerait un schéma représentant les différents sujets abordés sur le thème permettant donc une meilleure vue d'ensemble et potentiellement une analyse plus poussée.

## Clustering

Le clustering est une méthode d'analyse permettant de séparer en groupe les thèmes dans un corpus. Afin de regrouper les thèmes plus facilement, j'utilise TF-IDF qui va me permettre de voir l'importance d'un terme dans une collection.

Pour la création du clustering, j'utilise K-means qui est un algorithme non supervisé, c'est-à-dire qui n'a pas besoin d'intervention humaine. La seule intervention est de donner le nombre de clusters souhaités. J'ai choisi ici 3 clusters après plusieurs tests :



Le nombre de 3 clusters me semble pertinent car aucun point ne se mélange vraiment et nous voyons une répartition censée. 2 clusters relient les clusters vert et jaune sur le graphique. Cela ne me semblait pas logique au vu de la distance entre les points. 4 était par contre de trop séparant des points qui s'entremêlent.

Nous pouvons remarquer que nous avons 3 clusters de taille différente. Le vert (n°1) contient la majorité des fichiers avec 286 en sa possession. Le plus petit cluster (0) contient 31 documents. Quant au dernier (2), il contient 90 fichiers.

J'ai pu analyser que le cluster 2 ne contient que 2 journaux : Le Petit Bleu allant de 1900 à 1913 et Le Soir passant par toutes les années.

Le petit cluster quant à lui ne contient pratiquement que des fichiers entre 1910 et 1915 avec une tendance axée sur 1913.



## Conclusion

En conclusion, pour le thème “protection des animaux” dans le corpus CAMille s'étendant de 1900 à 1925, nous pouvons remarquer qu'il manquait 4 années qui sont 1916, 1917, 1918 et 1919. Je n'ai pas su déduire s'il y avait une raison à ce manque dans le corpus. Ces 4 années manquantes expliquent que la décennie 1910 était celle comportant le moins de documents, et la décennie 1920 contenait quant à elle le plus de fichiers.

Nous pouvons analyser que les principaux mots clés qui sont apparus dans notre corpus sont liés à la politique et aux potentiels noms d'ASBL pour la protection des animaux. En effet, le sujet des animaux était très politique à l'époque - et encore aujourd'hui -, ceux-ci étant considérés comme des objets sans droits. Le nuage de mots nous a confirmé cette observation avec les mots clés.

De manière générale sur l'échantillon relevé, les articles qui ont été écrits sont soit neutre soit positif et principalement subjectif. Lorsque des articles comprenaient du négatif, après vérification, il s'agissait d'extraits de parole directe des associations envers certains droits non respectés etc. Dans les entités nommées, nous retrouvons encore des personnalités politiques tout comme des organisations du même type.

Enfin, le clustering nous a permis de définir 3 types différents de thème dans le corpus. Un petit, un moyen et un grand comprenant chacun des caractéristiques différentes. Le plus petit ne contenant que des fichiers entre 1910 et 1913 en majorité et le moyen que des articles des journaux “Le Petit Bleu” et “Le Soir”.

Suite à cette analyse, je peux donc affirmer que le sujet de la protection des animaux était un sujet très politique pour l'époque, les droits des animaux n'étant pratiquement pas reconnus. Le droit de ceux-ci est encore aujourd'hui remis en question et en évolution.

## Bibliographie

Dossche, Florence. Delvaux, Paul-Henry., *Le droit des animaux: perspectives d'avenir*, Bruxelles : Larcier légal, 2019.

Brucker R., Sellier J., “L’histoire du concept du droit des animaux, du mouvement animaliste et du véganisme”, dans *Allemagne d'aujourd'hui*, No. 230, 28.11.2019, pp. 140-156.

Pouillard V., “Conservation et captures animales au Congo belge (1908-1960). Vers une histoire de la matérialité des politiques de gestion de la faune”, dans *Revue historique*, No. 679, 3.2016, pp. 577-604.

La Croix Bleue de Belgique, *Historique de la Croix Bleue de Belgique*. Société royale La Croix Bleue de Belgique. <https://www.croixbleue.be/la-croix-bleue-de-belgique/>, consulté le 26 novembre 2022.