ACQUISITION DE DONNÉES SUR LES RÉSEAUX SOCIAUX

Laurane Castiaux

LCLIG2140 Mars 2020

Acquisition de données sur les réseaux sociaux

Le rapport suivant détaille d'abord les différentes étapes que j'ai suivies pour collecter un corpus de tweets sur le thème du carnaval d'Alost, et décrit ensuite ce corpus. Je termine par une très courte synthèse sur la méthodologie et le contenu des tweets.

Tous les programmes utilisés pour ce travail (à l'exception du programme m'authentifiant sur Tweepy) sont disponibles sur mon github : https://github.com/Reyla98/TweetsAnalyses.

1. Récolte de tweets

1.1 Récolte par mots-clés

Pour commencer ma récolte de tweets, j'ai fait une recherche via Tweepy avec les mots clés « carnaval Alost » (programme utilisé : create_ndjson.py). Cela m'a donné 7082 résultats que j'ai enregistrés au format ndjson dans le fichier Alost.ndjson. Pour chaque tweet, j'enregistre son id, le screen name de l'utilisateur, l'id de l'utilisateur, la date de création de tweet, ses hashtags éventuels, sa langue et son texte complet.

J'ai ensuite fait la même recherche avec « Aalst carnaval ». J'ai ajouté les tweets à la suite du fichier précédent, ce qui m'a donné 18599 tweets. Parmi ceux-ci se trouvaient bien sûr des doublons ainsi que des retweets (qu'on ne peut pas filtrer avec api.search() de Tweepy). Je les ai supprimés via le programme rm_duplicates.py.

Sur les 4595 tweets restants, j'ai cherché quels étaient les hashtags les plus fréquents grâce aux fonctions countTags() et dictMaxValues() (dans le programme manipulate_tweets.py). Il y en a au total 867 différents, dont les 30 premiers sont :

[('Aalst', 334),
('carnaval', 233),
('aalstcarnaval', 227),
('Alost', 184),
('Carnaval', 163),
('aalst', 90),
('Belgium', 74),
('Belgique', 63),
('AalstCarnaval', 61),
('7dag', 59),
('ikbenblank', 42),
('Oilsjt', 41),
('Israël', 41),
('Antwerpen', 40),
('Gent', 40),
('antisemitisme', 39),
('Trump', 39),
('CLUCHA', 39),
('joodse', 39),
('autisme', 38),
('gedicht', 38),
('gedichten', 38),
('Italie', 38),
('Bernie', 38),
('Sanders', 38),
('Eden', 38),
('Hazard', 38),
('Wilder', 38),
('juist', 38),
('Brabant', 38),
('Laat', 38)]

Étant très étonnée de trouver des tags comme « #Sanders » ou « #Hazard » dans cette liste, j'ai affiché le texte des tweets contenant ces hashtags. Le texte se compose à chaque fois uniquement d'un lien, suivi des 27 mêmes hashtags en réponse à un tweet différent. Le lien mène vers une page de sensibilisation à l'autisme. J'ai alors cherché qui étaient les auteurs de ces tweets et il s'agissait toujours du même utilisateur. Sans doute veut-il/elle amener un maximum de gens vers cette page en utilisant des hashtags très suivis. Ces tweets n'ont donc aucun rapport avec le carnaval d'Alost (si ce n'est par l'utilisation du hashtag à succès « #aalstcarnaval »), je décide donc de les filtrer ; il me reste 4568 tweets.

La nouvelle liste des 30 hashtags les plus fréquents est la suivante :

[('Aalst', 296),
('carnaval', 195),
('aalstcarnaval', 189),
('Alost', 184),
('Carnaval', 163),
('aalst', 90),
('Belgique', 63),
('AalstCarnaval', 61),
('Oilsjt', 41),
('antisemitisme', 39),
('Belgium', 36),
('Carnaval2020', 30),
('oilsjt', 29),
('vrtnws', 29),
('Carnival', 27),
('deafspraak', 24),
('alost', 22),
('aalstcarnival', 21),
('7dag', 21),
('antisémitisme', 17),
('deochtend', 14),
('CarnavalAlost', 13),
('Carnival2020', 13),
('begov', 12),
('Unesco', 12),
('Antisemitism', 11),
('Israel', 11),
('belgium', 10),
('Aalstcarnaval', 10),
('carnavalaalst', 10),
('dwdd', 10)]

Dans cette liste, on retrouve des mentions à des lieux géographiques (Aalst, Oilsjt, Belgium), au carnaval (carnaval, Carnival2020), à des média (vrtnws, deafspraak, 7dag), à des institutions (begov, Unesco) et à l'antisémisime.

J'ai relancé une recherche sur #aalstcarnaval et toutes ses variantes. Cela m'a donné seulement 166 résultats. Je suppose que j'en ai si peu du fait que cette recherche a été effectuée un jour plus tard que la première recherche (c'est-à-dire le 29 février 2020) et que l'API tweeter empêche de remonter au-delà de 7 jours. Après avoir à nouveau enlevé les doublons et les retweets, il reste 4681 tweets.

1.2 Récolte par *user*

J'ai voulu étendre ma recherche en passant par des comptes utilisateurs. Je me suis donc servi des fonctions countTweetsPerUser() et dictMaxValues() (dans le programme manipulate_tweet.py) pour chercher les 10 utilisateurs qui ont tweeté le plus dans mon corpus. Les premiers sont :

[('VaesNico', 38),
('TrendolizerVL', 34),
('HLN_BE', 32),
('INDEPEStilence', 29),
('batmatdc', 24),
('AntoineDerijck', 21),
('Mirthe10722243', 19),
('freddydumont3', 17),
('Vlaamse_Zaak', 16),
('sancdelrio', 16),
('msnbelgie', 16)]

Je me suis ensuite servi de la recherche avancée de Twitter pour faire une recherche sur les tweets postés par ces utilisateurs contenant soit le mot Aalst, soit le mot Alost. J'ai sélectionné parmi les résultats proposés ceux qui me paraissaient pertinent et j'y ai repéré des mots clés supplémentaires utilisés spécifiquement par ces utilisateurs lorsqu'ils parlent du carnaval d'Alost. J'ai donc lancé les recherches suivantes avec Twitterscraper :

- (Aalst OR Carnival) (from:VaesNico)
- Aalst (from:INDEPEStilence)
- ((#Aalst OR #Oilsjt) AND (#Carnival OR #Carnaval)) (from:batmatdc)
- (Alost OR carnaval) (from:AntoineDerijck)
- Aalst (from:HLN_BE) until:2020-03-01 since:2020-02-14

Dans le cas de l'utilisateur @HLN_BE, j'ai ajouté un filtre sur une période de temps allant du 14 février 2020 au 1er mars 2020, ce qui correspond aux dates de début et de fin des festivités liées au carnaval à Alost. J'ai décidé de préciser ces dates car ce compte étant lié à un média d'information, ils ont énormément de tweets, dont certains, plus ou moins anciens, qui

évoquent Alost sans lien avec le carnaval. Restreindre la période de recherche m'a permis de filtrer ces résultats.

J'ai enregistré ces tweets dans le fichier Alost_user.ndjson. J'ai voulu les séparer des autres étant donné que la méthode utilisée pour les récupérer est totalement différente de la méthode utilisée précédemment. Cela m'a ainsi permis de garder une trace d'où viennent les tweets récupérés.
Après coup, je me suis dit qu'il aurait été bien d'ajouter cela dans les métadonnées des tweets au moment de leur téléchargement.

Ce nouveau corpus contient 107 tweets, dont 55 en commun avec le premier corpus.

2. Analyse de mon corpus

Pour cette analyse, j'ai regroupé mes deux sous-corpus (en retirant les doublons) dans le fichier Alost_corpus.ndjson.

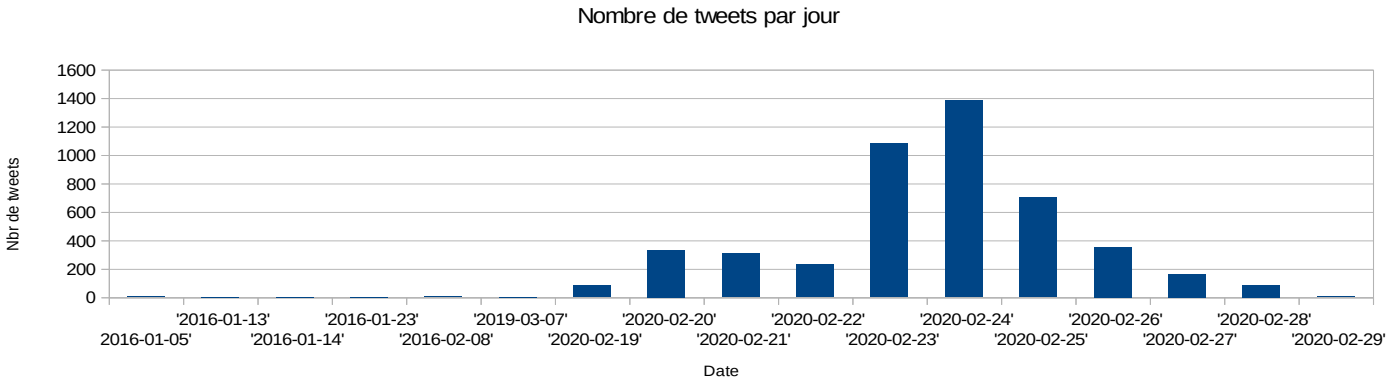
Pour m'aider à extraire des informations pertinentes, j'ai écrit la fonction « advancedSearch() » (manipulate_tweet_v2.py) et j'ai modifié mes fonctions précédentes pour qu'elles fonctionnent sur une liste de tweets (retournée par advancedSearch()) et non plus avec un fichier de tweets.[1]

Une partie des manipulations faites sur le corpus sont disponibles dans le programme analyses.py. Bien entendu, j'ai modifié certaines lignes de commandes au gré de mes besoins.

2.1 Date

La grande majorité des tweets ont été publiés entre le 19 février 2020 et le 29 février 2020, avec un maximum de 1383 tweets le 24 février. 1 tweet date de 2019 et 13 de 2016 ; tous ces tweets ont été publiés par @batmatdc. Cette disparité vient de l'utilisation de Twitterscraper pour récupérer les tweets de cet utilisateur et pas des autres.

Étant donné que cette récolte a pour but de s'entraîner à utiliser différentes méthode d'acquisition de tweets, j'ai choisi de les laisser dans mon corpus pour continuer mes analyses. Dans une autre situation, j'aurais dû redéfinir mon sujet de recherche, soit en me limitant aux tweets de 2020, soit en utilisant Twitterscraper pour faire une recherche avec les mêmes mots-



[1] Il aurait été idéal de prévoir une fonction équivalente à advancedSearch() qui fonctionnerait sur une liste, afin de pouvoir faire des recherches en cascade sans à chaque fois reparcourir l'ensemble du corpus. Je ne l'ai pas fait car la taille relativement réduite de mon corpus me permet de le reparcourir entièrement à chaque recherche.

clés que ceux utilisés sur Tweepy pour ainsi avoir des tweets plus anciens d'autres utilisateurs et donc avoir un corpus représentatif du carnaval d'Alost en général et non pas seulement de cette année.

2.2 Langue

Il y a 18 langues représentées dans ce corpus. Comme on peut s'y attendre, les langues majoritaires sont le néerlandais (3055 tweets), le français (965 tweets) et l'anglais (409 tweets). Suivent ensuite l'espagnol (109 tweets) et le catalan (28 tweets). Les autres langues représentées sont le portugais (23), l'allemand (11), l'estonien (10), le roumain (4), le danois (3), le polonais (3), l'italien (2), l'hindi (1), le créole haïtien (1), l'indonésien (1), l'hébreux (1), l'islandais (1) et le turc (1).

105 tweets ont une langue « indéterminée » car ils ne contiennent que des hashtags, mentions, images ou URLs.

La raison pour laquelle j'ai autant de tweets en espagnol et catalan dans mon corpus (et non pas, par exemple, en italien) tient sûrement au fait que dans ces deux langues, *carnaval* se dit « carnaval » et *Alost* « Alost ». J'ai donc dû récolter ces tweets lors de ma recherche sur ces deux mots-clés visant principalement à récolter des tweets francophones.

2.3 Utilisateurs

Le corpus se compose de tweets provenant de 2796 utilisateurs différents. La répartition parmi les principales langues est la suivante :

- Néerlandais : 1744 utilisateurs
- Français : 715 utilisateurs
- Anglais : 254 utilisateurs
- Espagnol : 66 utilisateurs
- Catalan : 22 utilisateurs

2.4 Contenu des tweets

Je vais maintenant m'attarder sur le contenu des tweets dans quelques-unes des langues de mon corpus.

Nérlandais

Plusieurs médias flamands ont abordé le sujet du carnaval d'Alost. C'est notamment le cas de la VRT, en particulier dans leur journal télévisé et dans leur émission *De zevende dag*, et du journal Het Laatste Nieuws (HLN).

J'ai dans mon corpus 148 tweets mentionnant/faits par la VRT ou une de leurs émission et 180 pour HLN. Il y a 7 tweets en commun dans ces deux sous-corpus.

J'ai voulu commencer par comparer les hashtags utilisés par ces deux médias, mais seulement 4 hashtags ont été utilisés plus d'une fois dans les tweets HLN, contre 17 pour la VRT. Une comparaison ne serait donc pas pertinente[2].

[2] Liste des hashtags pour chacun de ces sous-corpus disponible dans l'annexe 1

J'ai donc plutôt comparé les mots utilisés[3]. Si les sujets abordés se recoupent fortement (les caricatures, la politique, les juifs, etc.), on peut tout de même noter que les lecteurs de la VRT choisissent des termes forts tels que *jodenhaat* et *homohaat*, qui sont très peu utilisés par ceux de HLN. Ces derniers ont préféré commenter l'affirmation de la première ministre Wilmès disant que les caricatures du carnaval d'Alost « *brengt het samenleven in gevaar* ». On peut également noter que HLN aborde les conditions climatiques, ce qui n'est pas du tout le cas de la VRT.

Si on regarde à l'ensemble des tweets en néerlandais, on y trouve des hashtags en lien avec Charlie Hebdo (#jesuicharlie (7), #JeSuisCharlie (6), #charliehebdo (3) et #JeSuisAalstCarnaval (1)). Ils revendiquent tous la liberté d'expression en rappelant au gens leur soutient envers ce journal satirique au lendemain des attentats, il y a maintenant 5 ans.

Français

Du côté francophone, certains parlent de « fête » et de « rire », d'autres de « honte » et d' « humiliation ». Il n'y a pas de doute, ce carnaval est bien sujet à « polémique »[4].

Le thème de l'antisémitisme est très présent. Sur les 965 tweets en français, 307 contiennent au moins un mot commençant par *antisémit-*, 131 autres commençant par *juif-* ou *juiv-*.

Parmi les autres sujets plusieurs fois abordés, on retrouve les Nazis et les camps de concentration. La plupart des autres tweets partagent des photos ou donnent leur avis sur ce carnaval (une grande majorité le critique, mais certains le défendent, au nom de la liberté d'expression ou en disant qu'il ne s'agit que d'humour, que cela fait partie du folklore local).

Contrairement aux flamands, les francophones ne mêlent pas la politique au carnaval d'Alost. Très peu de politiciens ont donné leur avis sur le sujet et les citoyens ne les ont pas mentionnés à l'exception du maire de la ville, membre de la NVA, qui est mentionné quelques fois.

On retrouve en français, comme précédemment en néerlandais, des évocations de Charlie Hebdo. Ici les avis sont plus partagés : certains défendent le carnaval, tout comme ils avaient défendu le journal, d'autres critiquent ceux qui font le rapprochement entre les deux types de caricatures, d'autres encore pointent les deux du doigt.

Espagnol et Catalan

Avant de prolonger mes commentaires, je tiens à préciser que je ne parle ni espagnol, ni catalan, je me suis donc servie de Google Translate pour comprendre le contenu des tweets rédigés dans ces deux langues.

Comme évoqué plus haut, c'est un « hasard » que j'ai des tweets dans ces deux langues, mais il n'empêche que leur contenu m'intéresse, car il donne un point de vue étranger sur un événement belge.

[3] Liste des mots les plus fréquents utilisés disponibles dans l'annexe 2
[4] Liste des mots les plus fréquents en français disponible dans l'annexe 3

Comme précédemment, il n'y a pas un hashtag particulier qui ressort de ces sous-corpus, si ce n'est #ENMundo, en espagnol, qui est utilisé par un média du Venezuela pour indiquer que le tweet porte sur l'actualité internationale[5].

En ce qui concerne les thèmes abordés, on retrouve à nouveau beaucoup le sujet de l'antisémitisme (environ un cinquième des tweets). En espagnol, l'Unesco est évoqué dans environ un dixième des tweets : de manière générale, les hispanophones approuvent la décision d'avoir retiré ce carnaval de la liste des patrimoines immatériels de l'Unesco. Du côté catalan, l'ex-président Puigdemont est mentionné plusieurs fois, mais uniquement par trois utilisateurs différents. Ils font tous les trois le parallèle entre cet homme politique et les partis d'extrême droite flamands et ils déplorent l'antisémitisme du carnaval[6]. Je tiens à préciser que ce rapprochement avait déjà été fait explicitement en 2018 par les carnavaliers, puisqu'un char, nommé « NCA » (*Nieuw-Catalaanse Alliantie*) avait été créé aux couleurs de la catalogne et sur lequel était affichée la figure de Puigdemont.

Anglais

En anglais, comme on peut s'y attendre, c'est à nouveau le thème de l'antisémitisme qui prédomine. L'Unesco est également mentionné. A part cela, aucun thème ne ressort de ces tweets.[7]

2.5 Lacunes du corpus

Comme déjà expliqué ci-dessus, la période temporelle couverte par le corpus n'est pas tout à fait homogène puisque quelques tweets sont antérieurs à 2020 alors que la grande majorité date de février 2020. De plus, deux méthodes très différentes ont été utilisées pour récolter les tweets (d'une part une recherche par mots clés avec Tweepy et d'autre part une recherche axée sur certains utilisateurs avec Twitterscraper). Comme déjà évoqué, j'ai fait ce choix dans un but pédagogique pour m'entraîner à utiliser différentes méthodes d'acquisition de tweets, mais pour constituer un corpus homogène, il aurait été mieux de ne se focaliser que sur une méthode, ou bien d'équilibrer l'usage de chacune.

Je reviens également sur le fait que j'ai récolté « par hasard » des tweets dans beaucoup de langues, alors que je me focalisais initialement sur le néerlandais, le français et l'anglais. En fonction de ce qui nous intéresse, il pourrait être pertinent de filtrer les autres langues ou bien d'étendre la recherche à d'autres mots clés dans les langues recherchées. Étant donné que toutes mes analyses sont basées sur des sous-corpus constitués des tweets d'une seule langue, cela ne m'a pas gêné ici.

3. Synthèse

3.1 Méthodologie

Pour récolter les tweets, j'ai adopté successivement une démarche *corpus-driven* ou *corpus-based* en fonction de mes besoins : j'ai sélectionné des mots clés comme point de départ, puis j'ai regardé quels étaient les mots les plus fréquent parmi les résultats, cela m'a permis de sélectionner d'autres mots-clés pour étendre ma recherche, et ainsi de suite.

---

[5] Liste des hashtags utilisés en espagnol et catalan disponible dans l'annexe 4
[6] Liste des mots les plus fréquents en espagnol et catalan disponibles dans l'annexe 5
[7] Liste des mots les plus fréquents en anglais disponible dans l'annexe 6

J’ai pu tester et comparer deux modules python différents pour cette tâche : Tweepy et Twitterscraper. Je trouve ce dernier plus simple d’utilisation, en particulier grâce aux nombreux filtres qui peuvent être utilisés lors d’une requête. Évidemment, le fait qu’il n’y ait pas de limite quant au nombre de tweets que ce module permet de télécharger est un atout non négligeable.

3.2 Contenu des tweets

Dans l’ensemble, les sujets abordés dans les différentes langues analysées sont assez semblables. L’antisémitisme est celui qui ressort le plus, aussi bien au niveau national qu’international.

En belgique, les politiciens flamands prennent part aux débats, alors que du côté francophone ce sont plutôt les citoyens qui s’expriment. À l’étranger, l’Unesco est souvent évoqué. Les Catalans font le rapprochement entre cet événement et leur histoire politique.

Il est plutôt amusant de constater que les défenseurs et détracteurs du carnaval d’Alost font appel au même argument qui pourrait se résumer à « souvenez-vous de Charlie Hebdo », mais le tournent chacun dans son sens.

4. Annexes

Annexe 1 : Liste exhaustive des hashtags par/au sujet de la VRT et HLN

VRT :	**HLN :**
[('vrtnws', 34), ('aalstcarnaval', 12), ('Aalst', 11), ('carnaval', 5), ('openvld', 4), ('sp_a', 4), ('cdenv', 4), ('groen', 4), ('7dag', 4), ('aalstcarnival', 3), ('nva', 3), ('deafspraak', 3), ('vtmnieuws', 3), ('aalst', 3), ('openbareomroepvanmijnvoeten', 2), ('Vlamingen', 2), ('zevendedag', 2), ('Carnaval2021', 1), ('magnjet', 1), ('achterlijkelangetenennitwitsgohome', 1), ('ikbenboos', 1), ('jesuischarlie', 1), ('racism', 1), ('Xenophobia', 1), ('perceptie', 1), ('Vlaanderen', 1), ('Wallonië', 1), ('Aalstcarnaval', 1), ('Antwerpen', 1), ('terzaketv', 1), ('hln', 1), ('België', 1), ('democratie', 1), ('dictatuur', 1), ('Bartsomers', 1), ('deochtend', 1), ('Labour', 1), ('Liebor', 1), ('klaagmieren', 1), ('oilsjtcarnaval', 1), ('oilsjt', 1), ('Hogent', 1), ('weesslim', 1), ('MSM', 1), ('politics', 1)]	[('aalstcarnaval', 4), ('Aalst', 3), ('carnaval', 3), ('AalstCarnaval', 3), ('aalst', 2), ('racisme', 1), ('youthforbelgium', 1), ('begov', 1), ('vvm', 1), ('freelol', 1), ('goforit', 1), ('belgiebarst', 1), ('notmygov', 1), ('notmypremier', 1), ('Carnaval', 1), ('BendevanNijvel', 1), ('Hanau', 1), ('JosHaex', 1), ('EvenwichtigeBerichtgeving', 1), ('Antisemitism', 1), ('sarcasm', 1), ('FierOpOilsjtCarnaval', 1), ('vrtnws', 1), ('vtmnieuws', 1), ('hln', 1)]

Annexe 2 : Mots les plus utilisés par/au sujet de la VRT et HLN

Source de la liste de mots-stop en néerlandais : https://github.com/stopwords-iso/stopwords-nl

VRT :	**HLN :**
[('aalst', 133), ('carnaval', 125), ('@vrtnws', 54), ('vrt', 40), ('joodse', 39), ('#vrtnws', 34), ('karikaturen', 32), ('nws', 30), ('nieuws', 27), ('**jodenhaat**', 20), ('optocht', 20), ('voil', 20), ('jeanetten"', 20), ('**homohaat**', 19), ('@de7dedag', 15), ('#aalstcarnaval', 12), ('#aalst', 12), ('carnaval.', 10), ('feest', 10), ('hand', 10), ('@vtmnieuws', 9), ('@stadaalst', 9), ('minister', 9), ('israël', 8), ('regering', 8), ('jan', 8), ('burgemeester', 8), ('@basselaerebart', 7), ('@nieuwsblad_be', 7), ('zegt', 7), ('[opinie]', 7), ('vlaamse', 7), ('linkse', 7), ('mee', 6), ('lachen', 6), ('@hln_be', 6), ('gehoopt', 6), ('uitgestoken', 6), ('calvo.', 6), ('helaas', 6), ('spreektijd', 6), ('slagen', 6), ('gordel', 6), ('uitgedeeld,', 6), ('jambon,', 6), ('burgemeester,', 6), ('steun', 6),	[('aalst', 153), ('@hln_be', 143), ('carnaval', 130), ('premier', 22), ('carnaval:', 21), ('joodse', 17), ('stoet', 17), ('wilmès', 15), ('live', 15), ('jaar', 14), ('karikaturen', 12), ('minister', 12), ('uur', 12), ('joden', 10), ('brengt', 9), ('israëlische', 9), ('belgië', 9), ('media', 9), ('scherp', 8), ('reageert', 8), ('persbericht', 8), ('**samenleven**', 8), ('kijk', 8), ('**windstoten**', 8), ('wagen', 8), ('vlammend', 7), ('gevaar”", 7), ('**stormweer**:', 7), ('begint', 7), ('morgen', 7), ('@sophie_wilmes', 7), ('lachen', 7), ('gelachen', 7), ('afbreuk', 6), ('gespot', 6), ('zit', 6), ('omstreden', 6), ('uitrijden', 6), ('verbieden:', 6), ('roept', 6), ('#aalstcarnaval', 6), ('reactie', 5), ('reputatie', 5), ('start,', 5), ('kostuums', 5), ('afgelast', 5), ('veiligheidscel', 5),

('herman', 6), ('goethem', 6), ('uiteraard.', 6), ('genuanceerd', 5), ('jaar', 5), ('wagen', 5), ('wereld', 5), ('lacht', 5), ('buitenlandse', 5), ('zaken', 5), ('#carnaval', 5), ('media', 5), ('@de_nva', 5), ('israëlische', 5), ('verbieden', 5)]	('samen,', 5), ('beslissing', 5), ('"jodenkoppen"', 5), ('politiek', 5), ('zwarte', 5), ('carnavalisten', 5), ('overlever', 5), ('aanslag', 5), ('bende', 5), ('bende,', 5), ('vond', 5), ('probleem', 5)]

Annexe 3 : Mots les plus utilises en français

Source de la liste de mots-stop en français : https://github.com/stopwords-iso/stopwords-fr

[('carnaval', 895),
("d'alost", 252),
('d’alost', 153),
('caricatures', 104),
('juifs', 101),
('#alost', 98),
('belgique', 80),
('belge', 77),
('alost', 61),
('antisémites', 59),
('accusé', 53),
('antisémite', 50),
('#carnaval', 50),
("d'alost,", 41),
('polémique', 39),
('d’alost:', 36),
('#belgique', 35),
('#aalst', 35),
('chars', 34),
('ville', 34),
('européenne', 34),
('d’alost,', 32),
("d'#alost", 32),
('défilé', 32),
('commission', 32),
("l'antisémitisme", 31),
('monde', 31),
('@rglucks1', 30),
('#aalstcarnaval', 30),
("d'alost.", 29),
('caricature', 28),
('alost,', 28),
('dénonce', 28),
("d'antisémitisme", 26),
('fête', 26),
('rire', 25),
("d'alost:", 25),
('belgique,', 25),
('persiste', 25),
('belges', 23),
('flamands', 23),
('antisémitisme', 22),
('place', 22),
('orthodoxes', 21),
('@lesoir', 21),
('bourgmestre', 20),
('liberté', 20),
('2020', 20),
('honte', 19),
('droit', 18),
('@rtbfinfo', 18),
('@lemondefr', 18),
('nazis', 18),
('patrimoine', 18),
('charlie', 17),
("d'antisémitisme,", 17),
('d’#alost', 17),
('prétexte', 16),
('carnaval,', 16),
('@unia', 15),
('moque', 15),
('jamais', 15),
('@libe', 15),
('signe', 15),
('oublie', 15),
('renversement', 15),
('#antisémitisme', 15),
('flamand', 14),
('@marcelsel', 14),
('carnaval.', 14),
('rôles', 14),
('puissants', 14),
('char', 14),
('faut', 13),
('"honte"', 13),
('#antisemitisme', 13),
('@marechalboulon', 13),
('nez', 13),
('face', 13),
('l’humiliation', 13),
('attise', 13),
('honte"', 13)]

Annexe 4 : Liste exhaustive des hashtags utilisés en espagnol et catalan.

Espagnol	Catalan
[('ENMundo', 9), ('Aalst', 7), ('Alost', 3), ('Carnavales2020', 3), ('Carnaval', 3), ('Bélgica', 2), ('aalstcarnaval', 2), ('antisemitismo', 2), ('ascoracistas', 1), ('humor', 1), ('carnaval', 1), ('patrimonio', 1), ('UNESCO', 1), ('Carnaval2020', 1), ('cartooning', 1), ('Israel', 1), ('antisemita', 1), ('España', 1), ('Unesco', 1), ('Racisme', 1), ('Belgique', 1), ('Bogotá', 1), ('Colombie', 1), ('FelizDomingo', 1), ('23Feb', 1), ('aalstcarnival', 1), ('Antisemitism', 1), ('racismo', 1), ('AalstCarnaval', 1), ('CriptanaEspaña', 1)]	[('Carnaval', 2), ('ViscaBèlgica', 2), ('Antifascisme', 1), ('Société', 1), ('PCI', 1), ('Unesco', 1), ('cridanacional', 1), ('Bèlgica', 1), ('aalstcarnaval', 1), ('Aalst', 1), ('Antisemitism', 1), ('VivaFlandes', 1), ('VivaBelgica', 1)]

Annexe 5 : Mots les plus utilisés en espagnol et catalan

Source de la liste de mots-stop en catalan : http://latel.upf.edu/morgana/altres/pub/ca_stop.htm
Source de la liste de mots-stop en espagnol : https://github.com/stopwords-iso/stopwords-es

es word :	CA words:
[('carnaval', 97), ('aalst', 43), ('alost', 22), ('bélgica', 22), ('ciudad', 15), ('antisemitas', 14), ('antisemita', 13), ('antisemitismo', 13), ('unesco', 13),	[('carnaval', 21), ('aalst', 10), ("d'aalst", 8), ('bèlgica.', 8), ('ciutat', 7), ('flamenca', 7), ('n-va,', 7), ('Puigdemont', 7), ('governada', 6),

('belga', 12), ('elementos', 12), ('polémica', 11), ('año', 11), ('bélgica,', 10), ('patrimonio', 10), ('aalst,', 10), ('judíos', 9), ('#enmundo', 9), ('https://t.co/phmxmamxav', 9), ('flamenca', 9), ('desfile', 8), ('humanidad', 7), ('más', 7), ('lista', 6), ('también', 6), ('n-va', 6), ('(bélgica)', 6), ('puigdemont', 6), ('amiguitos', 6), ('alost,', 5), ('conocido', 5), ('ue', 5), ('qué', 5), ('bélgica:', 5), ('#aalst', 5), ('comisión', 4), ('años', 4), ('después', 4), ('nazis', 4), ('#carnaval', 4), ('inmaterial', 4), ('alost.', 4), ('retiró', 4), ('antisemita.', 4), ('aberrante', 4), ('título', 4), ('aalst.', 4), ('gobernada', 4), ('(los', 4), ('tropa)', 4)]	('pel', 6), ('tropa', 6), ('(prop', 6), ('gant)', 6), ('antisemita', 6), ('és', 5), ('visca', 5), ('(bèlgica),', 4), ('amiguets', 4), ('això', 4), ('del', 4), ('😾😻😻', 4), ('amics', 3), ('flamencs', 3), ('any', 3), ('aalst,', 3), ('belgues.', 3), ('al', 3), ('lleva', 3), ('cabo', 3), ('@amcorominas', 2), ('@de_censurado', 2), ('https://t.co/oeqbogtvx3', 2), ('israel', 2), ('exigeix', 2), ('bèlgica', 2), ('prohibeixi', 2), ('exhibeix', 2), ('repugnant', 2), ('antisemitisme', 2), ('https://t.co/uoilycyxgk', 2), ('@324cat', 2), ('@montser45122141', 2), ('#carnaval', 2), ('(bèlgica)', 2), ("d'aalst,", 2), ('#viscabèlgica', 2), ('fan', 2)]

Annexe 6 :

Source de la liste de mots-stop en anglais : https://github.com/stopwords-iso/stopwords-en

[('aalst', 293),	('#aalst', 42),	('carnaval,', 33),
('carnaval', 228),	('#carnaval', 41),	('jews', 31),
('people', 75),	('jewish', 40),	('understand', 30),
('carnaval.', 60),	('#aalstcarnaval', 38),	('belgium', 28),

('#belgium', 28),
('unesco', 27),
('#oilsjt', 26),
('@emmanuelnahshon', 26),
('@belgiummfa', 26),
('antisemitism', 20),
('carnival', 19),
('#carnival', 19),
('@belgiuminisrael', 19),
('aalst.', 19),
('parade', 18),
('@israel_katz', 18),
('laugh', 17),
('@raphaelahren', 16),
('belgian', 16),
('@johnhyphen', 16),
('@adammouchtar', 16),
('antisemitic', 16),
('fun', 16),
('#belgique', 15),
('aalst,', 14),
('caricatures', 14),
('#carnival2020', 13),
('year.', 13),
('mocked', 13),
('time', 13),
('#antisemitism', 12),
('city', 12),
('hate', 12),
('read', 12),
('israel', 12),
('&', 11),
('picture:', 11),
('flemish', 11),
('@holocaustmuseum', 11),
('about.', 11),
('carnaval!', 10),
('explain', 10),
('year,', 10),
('racist', 10),
('mock', 10),
('float', 10),
('learn', 10),
('dresscode', 10),
('bad', 9),
('jews,', 9),
('everything.', 9),
('satire', 9),
('dialect', 9),
('word', 9),
('@beholdisrael', 9),
('@eu_commission', 8),
('day', 8),
('live', 8),
('antisemitism.', 8),
('@larryoneill', 8),
('media', 8),
('tradition', 8),
('mocking', 8),
('parade.', 8),
('single', 8),
('about,', 8),
('jokes', 8),
('that.', 8),
('mayor', 8),
('@dschwammenthal', 8),
('nazi', 8),
('love', 8)]