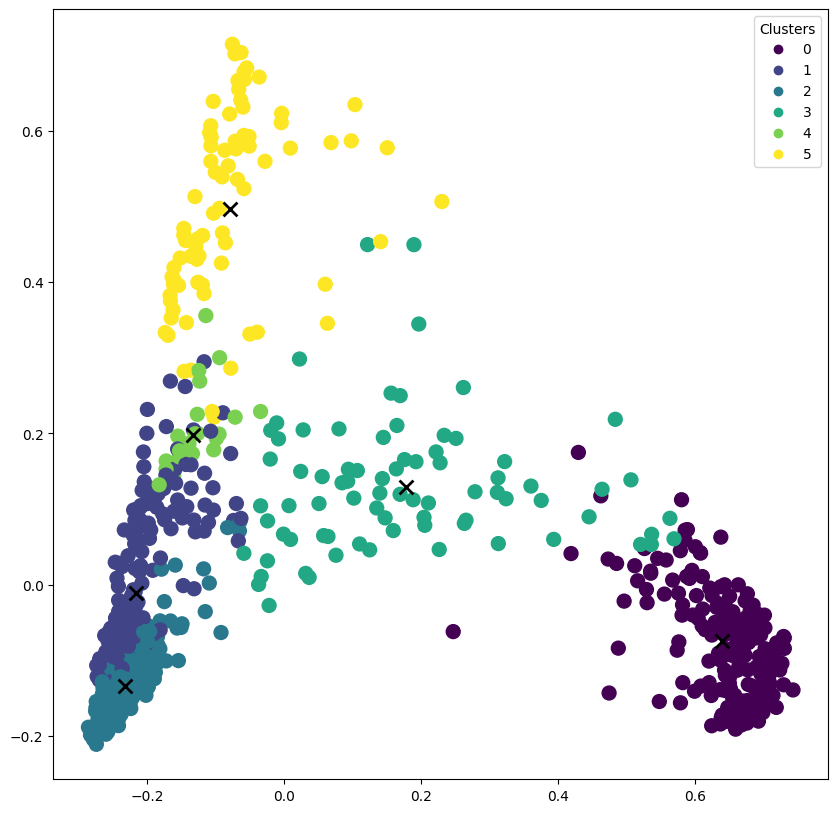
TP3 : Clustering et word2vec

1. Clustering avec le Notebook s2\_clustering.ipynb

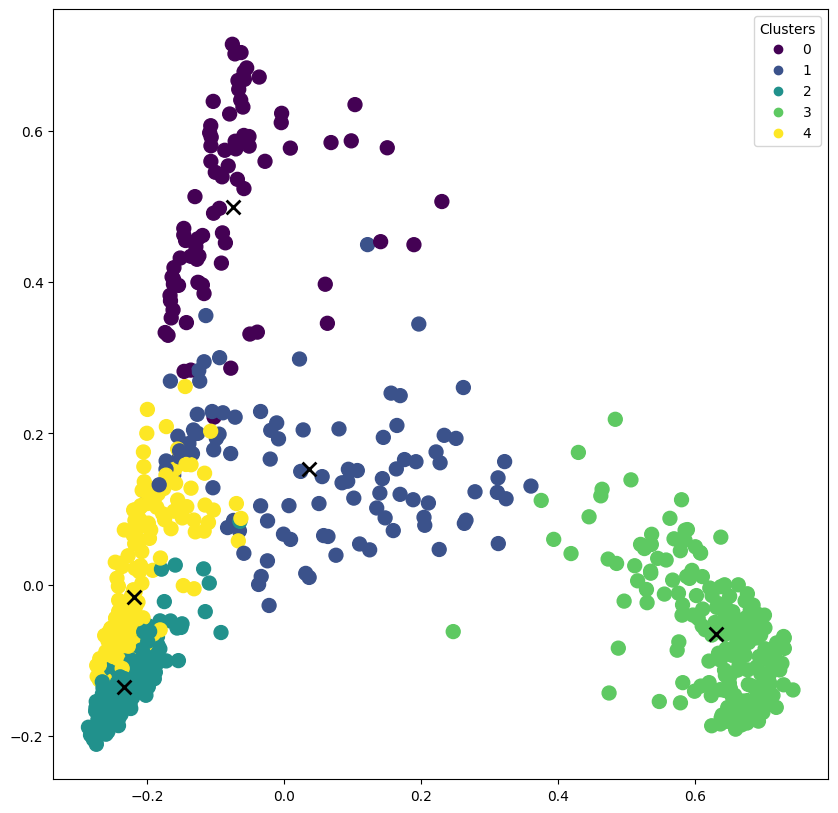
La décennie choisie est celle de 1890 – 1899

Notre décennie contient 900 fichiers en tout et nous allons faire des différents clusters afin de trouver la plus adaptés.

* En essayant l’algorithme avec 6 clusters au départ, nous avons une structure assez complexe, les données sont superposées surtout les bleus et les deux verts qui sont presque les unes sur les autres, les centres vides sont éloigner. Le cluster mauve semble assez éloigné des autres comme pour dire qu’il n’y a pas grande similarité entre les mots de ce cluster et les autres.

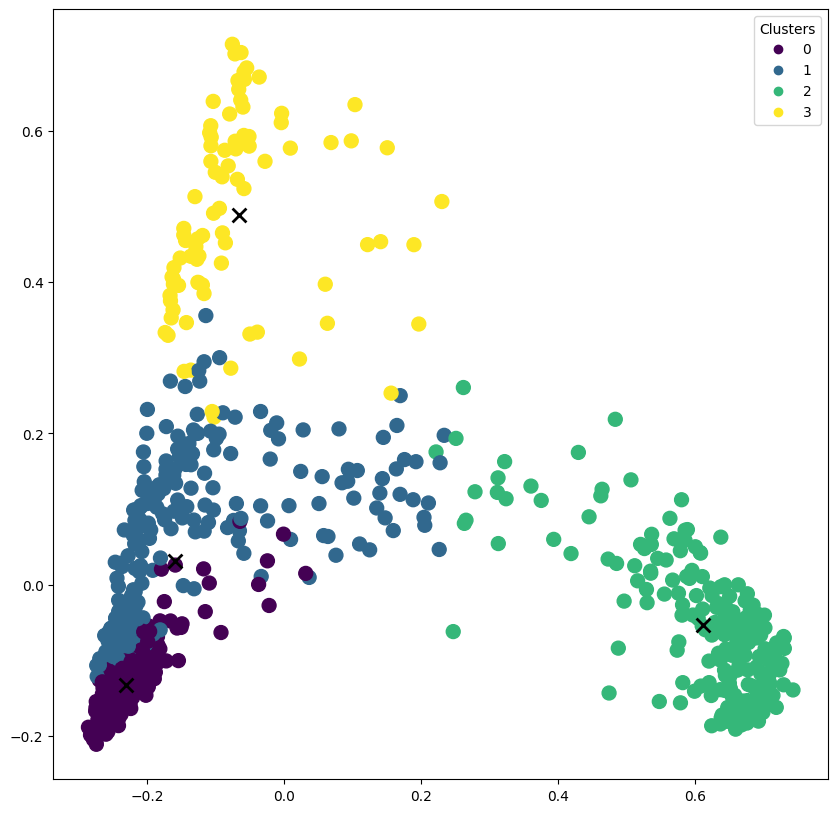


* En réduisant à cinq clusters, il n y’a toujours pas une meilleure séparation qu’avec six clusters, les clusters continuent à être proche, notamment le bleu, jaune et le vert. Les clusters mauve et vert citron semblent n’avoir pas de grande similarités avec les autres du leurs distances.

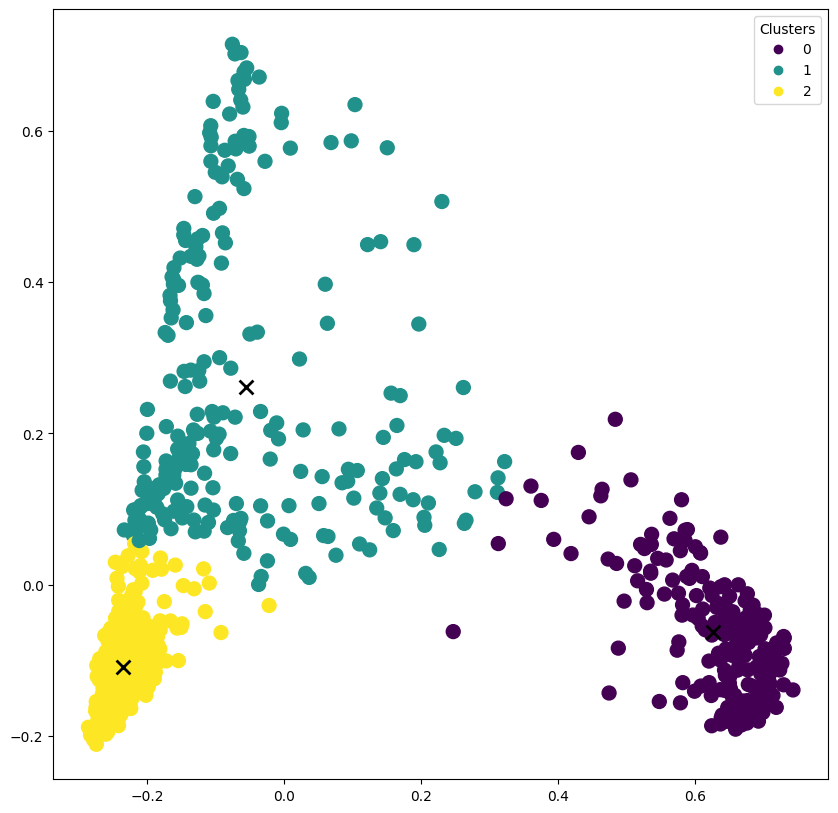


La segmentation ici, n’est toujours pas très claire. Y’a par contre une forte concentration des clusters jaune et vert au point ou leurs centre vide se touche pratiquement.

* En essayant avec quatre clusters, on a déjà une assez bonne segmentation. Les clusters jaune et vert citron sont bien éloignés des autres, chaque centre des clusters est éloigné les uns des autres. Néanmoins, les centres vides des clusters mauve et bleu sont proches et les clusters sont pratiquement les uns sur les autres.



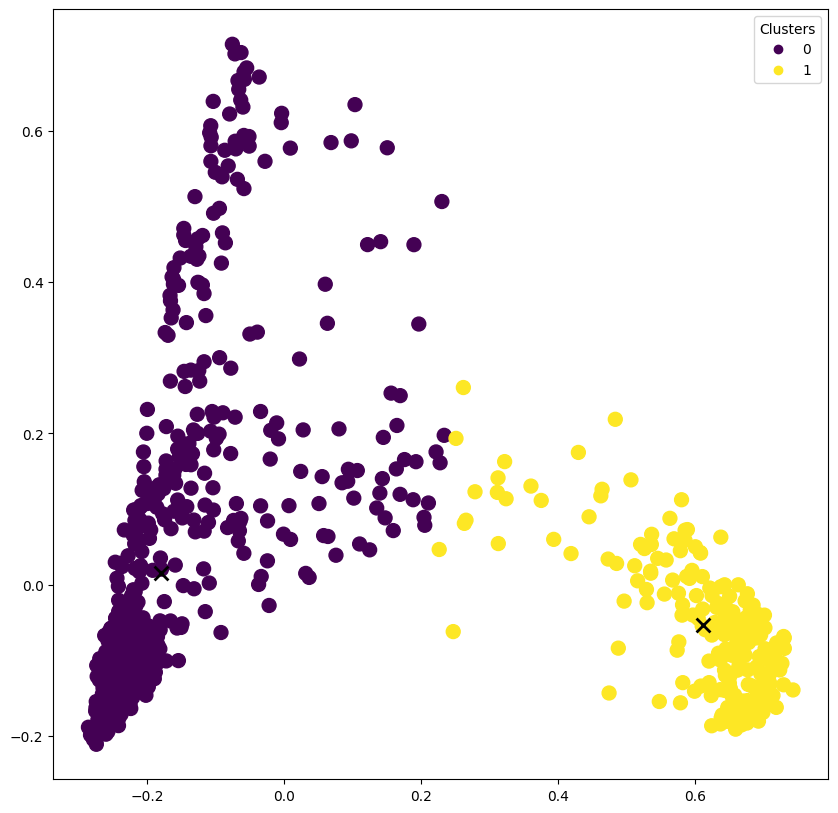
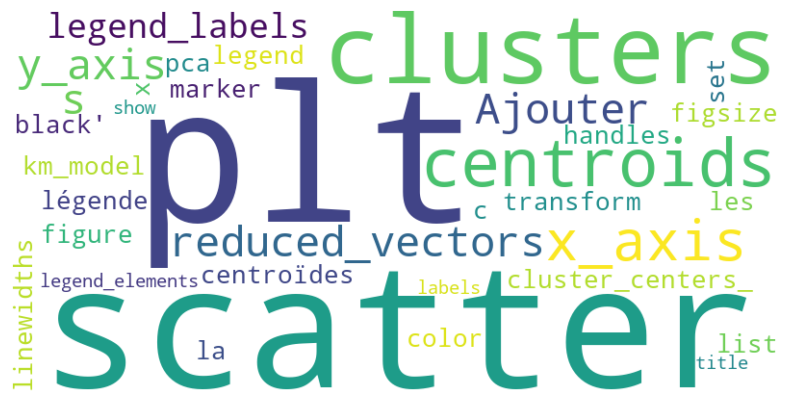
* En procédant cette fois avec trois clusters, la structure semble ne plus être complexe, les groupes sont bien distincts. Les centres bien visible et chacun éloignés de l’autre ; les éléments de chaque cluster gravite autour de son centre, la segmentation est clair.



On a une vision de chaque catégorie bien distincte, visible et défini.

* En essayant tous de même avec deux clusters, cela semble être la bonne car chacun des clusters est de son coté, même si au milieu ils semblent avoir quelques données similaires car quelques point semblent se rencontrés. Chaque centre est bien défini l’un de l’autre. Les textes constituant ces derniers sont très bien concentrés autour du centre vide.

Et nous avons donc générer un nuage de mot de ce cluster qui ne semble vraiment pas faire sens.

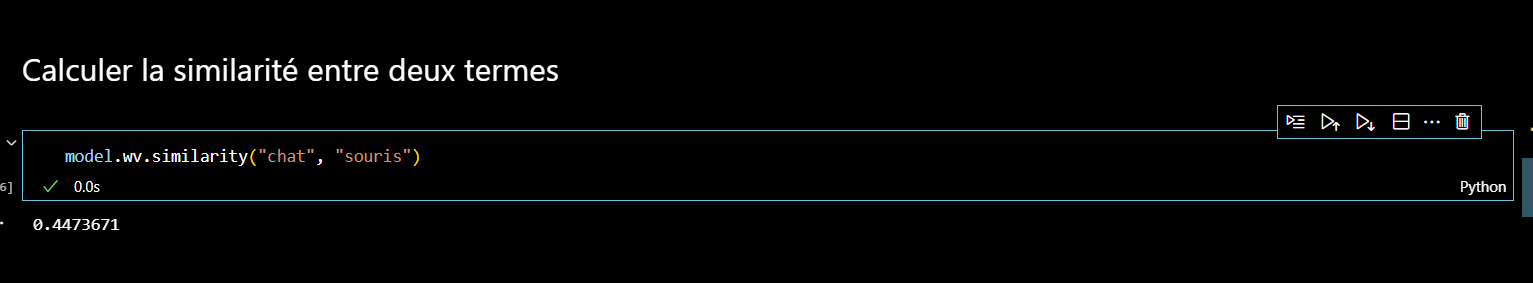
**CONCLUSION**: La meilleure configuration au vu des différents cas étudiés est celle ou on dispose de 2 clusters et 3 car elle permet une distinction beaucoup plus claire de celle-ci. Les centres vide sont bien distinct et surtout bien distant d’où le meilleur cluster parmi les six qui ont été analysés. Les données sont assez bien segmentées.

1. Modèle word2vec

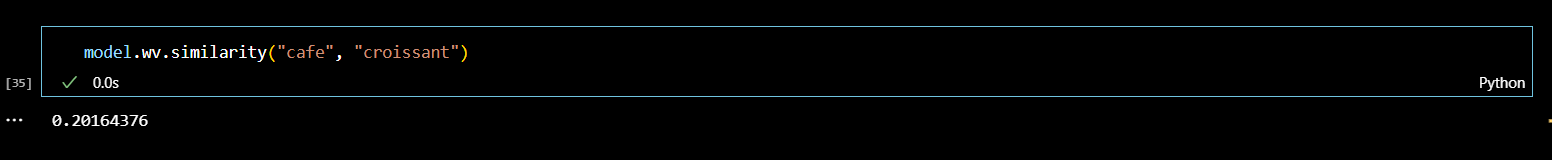
Afin d’entrainer le modèle word2vec, nous allons adapter la taille du vecteur à 52 et mettre le Windows à 7.

1. Similarity

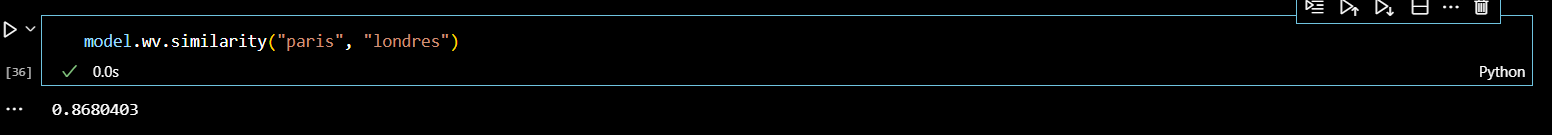
* Pour cette fonction, nous essayons d’interpréter la similarité entre les termes ‘chat’ et ‘souris’ et nous constatons qu’elle donne une valeur neutre. On peut donc conclure que les termes ici sont loin d’être proche, mais ne sont pas aussi distants l’un de l’autre surement dû à la manière donc au contexte spécifique du modèle.



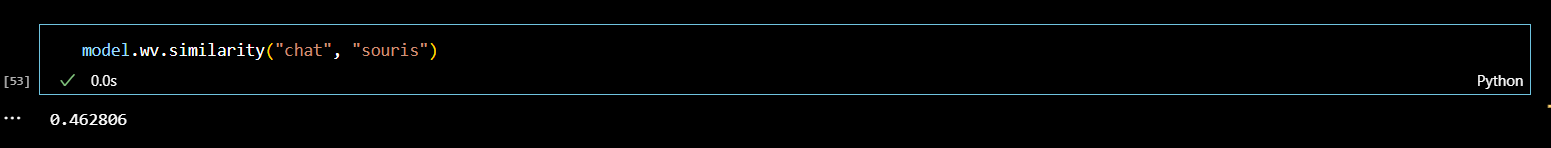
* Pour les termes ‘café’ et ‘croissant’, on constate que la similarité se rapproche et a une valeur positive. Par contre, la valeur montre que ces termes malgré tout n’ont pas une bonne similarité or, café et croissant sont deux termes qui devraient être assez proche, étant donné qu’ils sont tout le temps utilisés dans le même contexte.



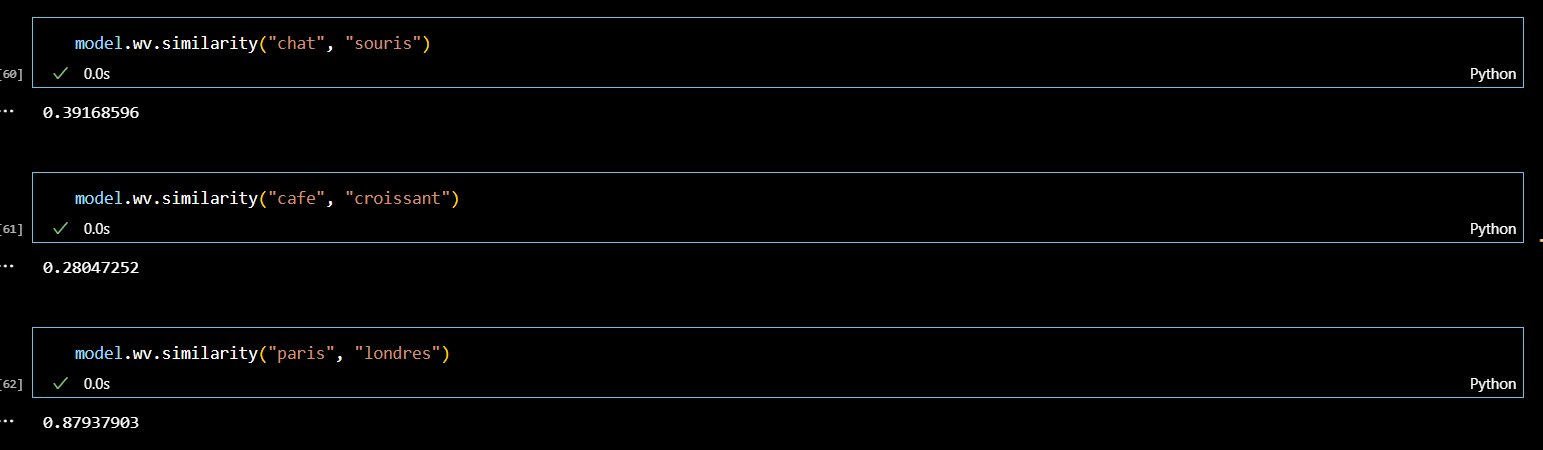
* Pour les termes ‘Paris’ et ‘Londres’, on a une similarité assez bonne, ce qui implique que les termes sont assez proche.



* En changeant les paramètres pour entrainer le modèle, donc en mettant la taille du vecteur à 32 et le Windows à 5, nous obtenons un résultat plus proche donc similaire malgré un pas très grande différence des résultats.



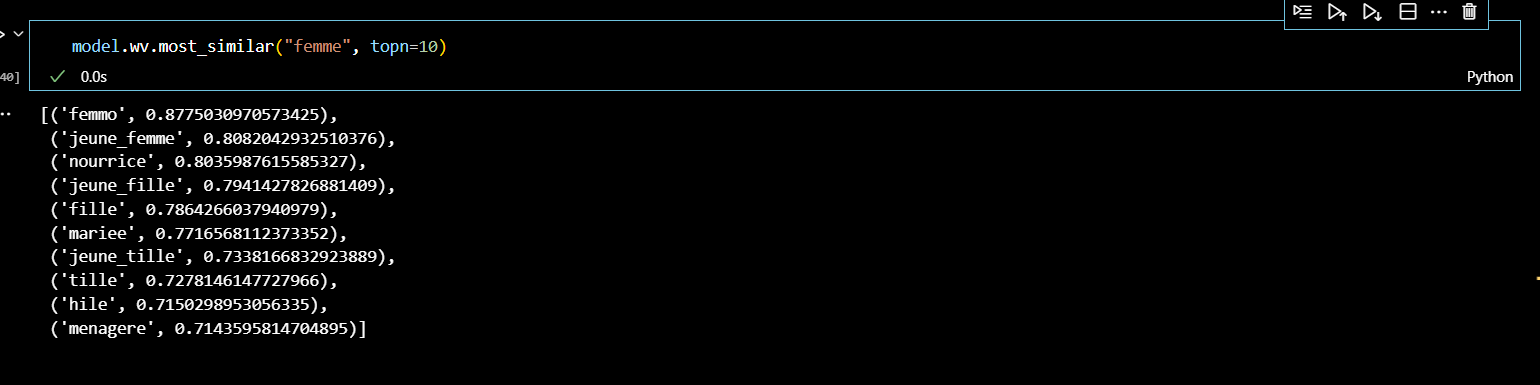
* Une fois de plus, afin d’entrainer notre modèle, nous avons changé de paramètres en mettant cette fois ci la taille du vecteur à 52, le Windows à 7 et le Mint\_Count à 7, nous avons obtenus les résultats suivant :



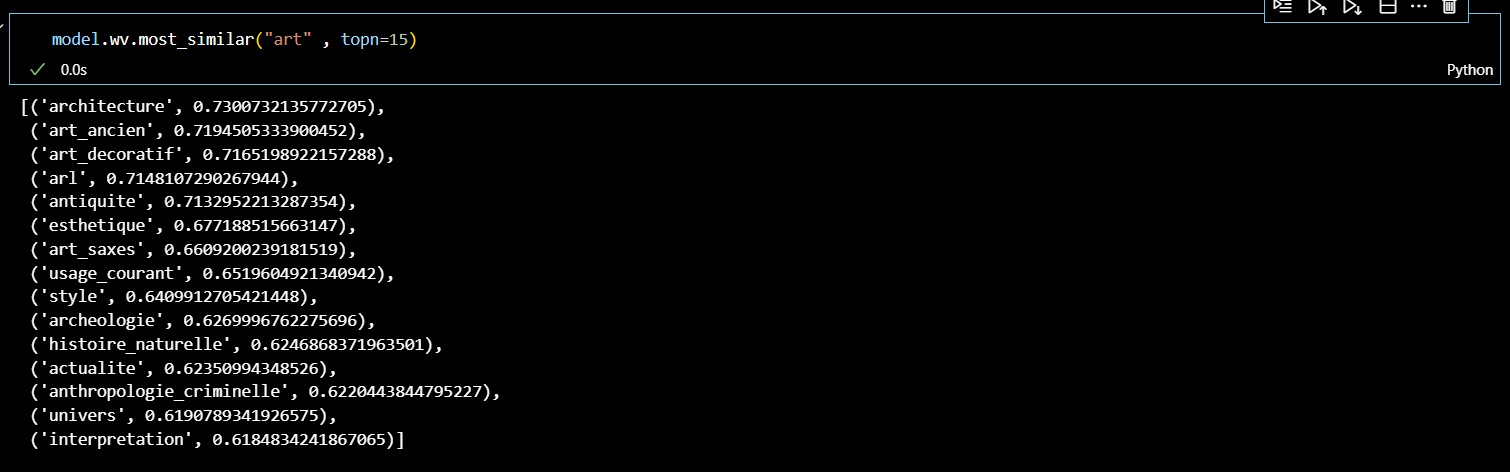
Et nous remarquons que les degrés de similitude ont changes voir même progresser. Par contre le degré de similitude entre «  chat et souris » a plutôt régressé.

1. Most\_Similar

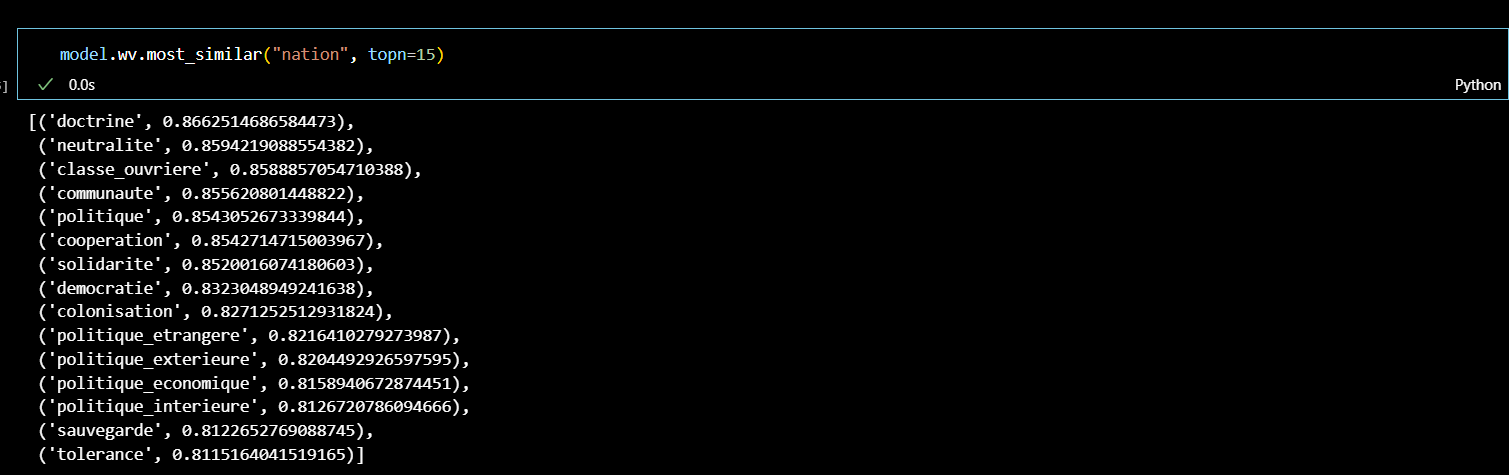
* Mot similaire pour le terme « femme » : la majorité des mots ici se rapprochent assez bien du terme FEMME et forme un bon champ lexical de mot. Par contre, les mots tels que « jeune-tille », « tille », « hile » n’ont vraiment rien avoir avec femme mais ont ici une très forte similarité avec celui-ci, le modèle a du se baser sur les valeurs ou la distance grammatical entre femme et ces autres mots.



* Mots similaires pour le terme ‘art’ : Pour la plupart des mots trouvés, la similarité est bonne car très proche. Par contre, on rencontre comme précédemment le terme ‘arl’ qui n’a pas de réel sens mais y figure sûrement à cause de sa proximité sur le plan grammatical.



* Mots similaires pour le terme « Nation » : le calcul est assez réaliste car la similarité est assez évidente ici ; tous les 15 mots ont une assez bonne similarité avec le mot de base Nation.



En entrainant le modèle, on se rend que le degré de similitude varie à chaque fois