Clustering Article

In this process, we give the tf-idf file and we obtain a list of cluster with the list of their article and the coordinate for each article.

At first, we shrink the number of dimension so we can make sensible calcul.
For the clustering, it’s a unsupervised learning using the module HdbScan and optuna.

We define a number of Trials to accomplishe and we create a study with optuna, bound to our database. The step of optimization is parallelized by using the thread thanks to arguments in optuna functions. But it’s deprecated so warning ! The good way would be to create many new process and make some multiprocessing. But, in optuna, the threading is managed by OpenMP and there is some incompatibility when OpenMP is used in a child process create by fork().

When the number of trials is accomplished, we get the best result and give coordinates to each article.

We give to the cluster a list of the most words that appear in its articles and we identified the cluster by this list of word we name «Topic». This is a string of words separated by comma.