Clustering problem

The task was about parsing xml file, extracting some interesting features from it and performing clusterization.

Firstly, I examined dataset and decided to extract following features:

- ➔ number of contributors which I took as a sum of editors and authors,
- ➔ number of pages,
- ➔ year of publication

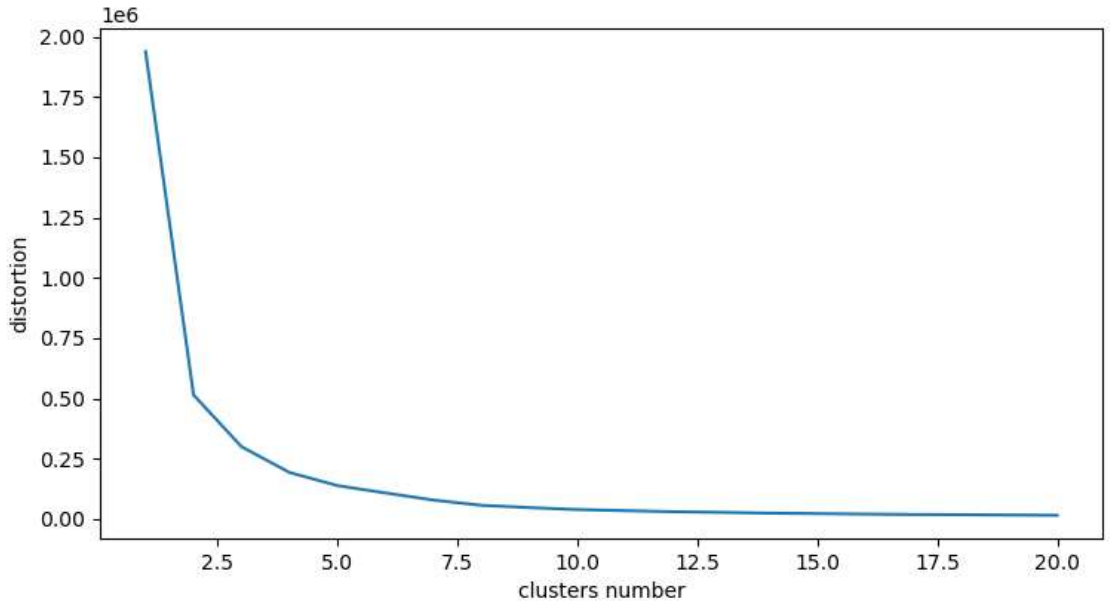
Then I read given xml file into python script using *xml.etree* python built in module and gathered mentioned features into 3 lists each for one feature. Using python *pandas* library I created DataFrames which are simply containers for datasets, it uses numpy underhood so it is very efficient and provides handy methods to manipulate the data and extract specific rows and columns from it.

In next step I performed PCA reduction analysis (*sklearn* python library) by doing fitting and transforming my data frame in order to see variance ratio of each feature. Basically PCA is used for dimension reduction and is based on covariance matrix and its eigenvectors computing. Output vector is presented below:

$$[0.972,\ 0.027,\ 0.001]$$

Regarding vector above, it shows that second and third feature doesn't bring too much information so it is pointless to perform clustering with 3 features. I decided to keep 2nd feature and present clusterization results for each of 3 pair combinations made from the set of 3 features.

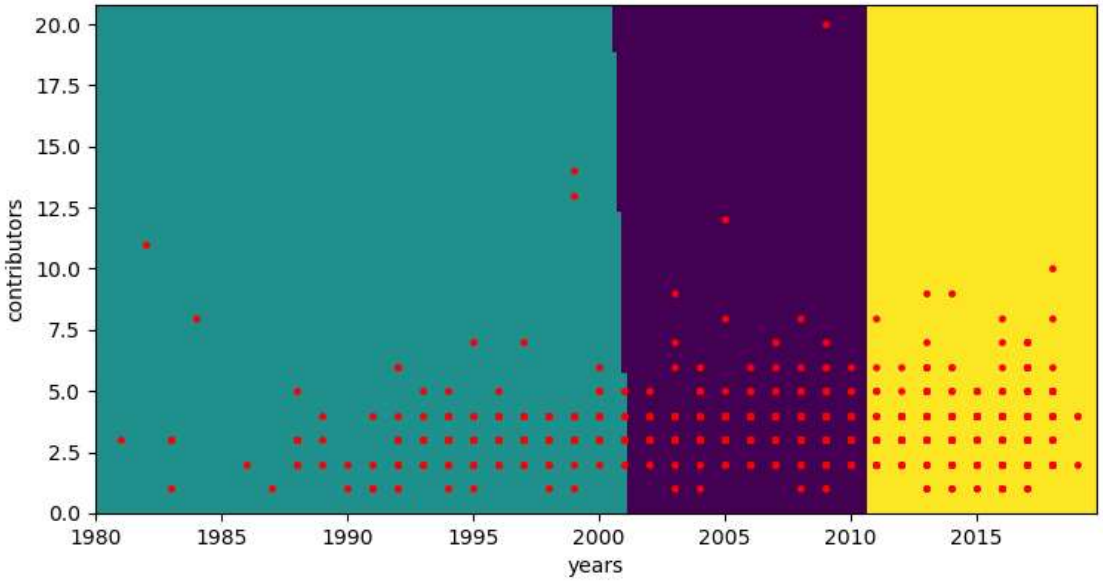
I examined the proper number of clusters that I should split the data into using *elbow method*. This method is all about calculating distortion parameter for every clusterized dataset in a loop while looping over number of clusters (1 to 20 in my case). Then I created a plot presenting number of clusters against distortion value. According to used method the optimal number of clusters is for the point where elbow would be if the plot would be considered as a human arm. For this case it can be seen that number 3 is optimal one.



Plot [1]. Cluster number against distortion value.

On the charts below there are presented results of KMeans (*sklearn* library) clusterization for every pair combined from 3 features took in account in this exercise. Because of high value of variance for first element in a PCA result vector it could be expected that clusters will be similar to simple vertical or horizontal stripes.

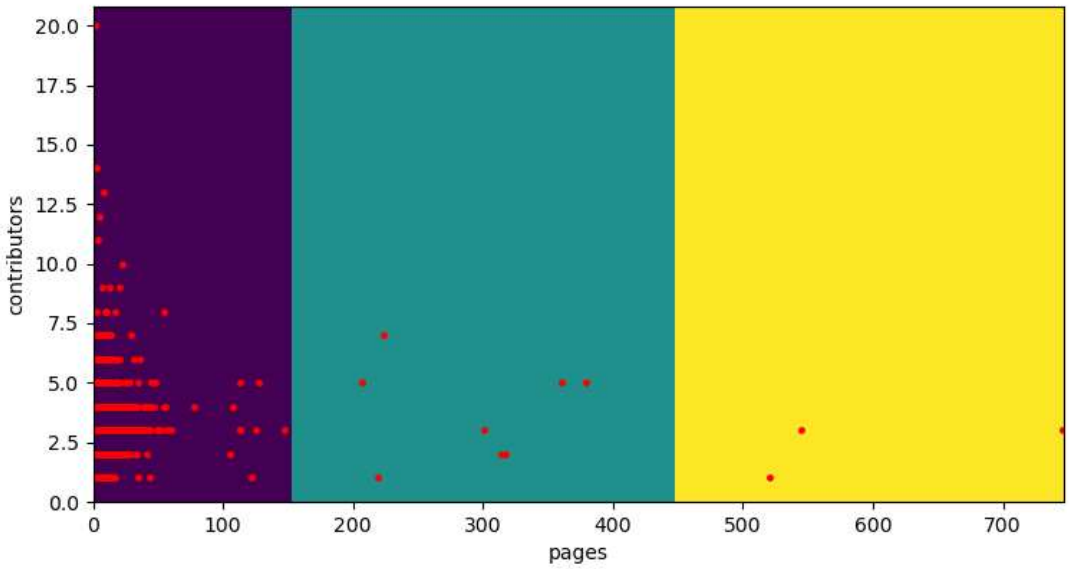
1. publication year and number of contributors



Plot [2]. Clusters for 2 dimensions - publication year and contributors number.

Although the clustering wasn't too effective in our case, some interesting relation might be seen. Ignoring alone points that are away from the main group we can conclude that people were more likely to cooperate with the course of time. When in the 80s and 90s there were 1 to 5 individuals involved in publication in the second decade of 2000 year this number ranged from 1 up to 10. The possible reason of such relation might be in technology becoming more and more complicated and complex so it became hard for 1 person to prepare complete publication about some topic because it overlaps with other disciplines or sectors and other experts could helps in this kind of situations.

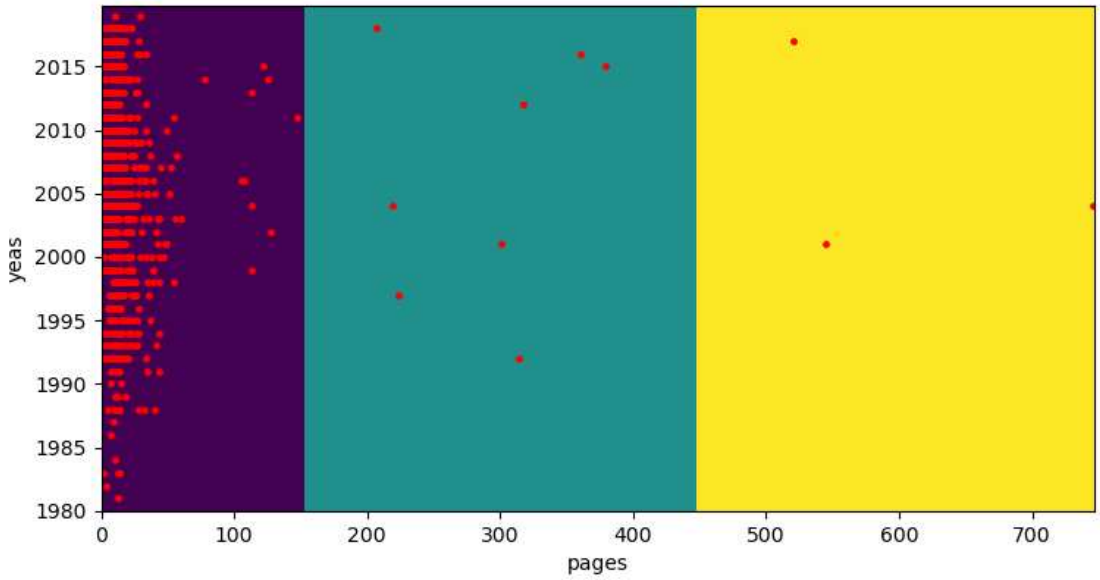
2. Number of pages and contributors number



Plot [3]. Clusters for 2 dimensions - pages and contributors numbers.

In this case it can be clearly noticed that greater publications in terms of number of pages are written by smaller groups of specialists. It could be related to fact that it is more difficult and to organize any kind of work in more numerous groups so larger projects are more often prepared by only a couple of people. I imagine how challenging it must be to write a coherent and uniformed article by lets say 10 people when it should have around 200 pages.

3. Number of pages and publication year



Plot [4]. Clusters for 2 dimensions - page number and publication year.

In the last case there are no evident relation at first glance. Maybe the only conclusion could be that with time articles and publications became more extensive.

Conclusions

To sum up I can say that clustering for this specific dataset and features that I have chosen wasn't that fenomenal. Selected features would probably present better on one dimension charts with simple grouping by years and counting for example. However it was possible to spot some interesting patterns thanks to 2 dimensional data presentation so whole work was definitely worth it.