Project Report of Assignment A for Project DarkWeb Analysis

Anirudh Bhupalarao(2018AAPS1242H)

Abstract :

For a given dataset, to develop a clustering algorithm that can divide nodes into different groups. These groups are formed such that nodes within a group are highly similar, whereas nodes between groups are dissimilar. This helps analyse the data clearly and give an idea as to how many pages come under each category and which is increasing in popularity. The algorithm was able to cluster the given data into four groups based on the page type i.e Government, Politician, TV Show and Company.

Approach Used :

As the project statement requires to arrange the given data as clusters, the best method for this is the machine learning clustering algorithms, particularly K-Means Clustering. The K-Means clustering algorithm is an unsupervised learning algorithm that separates data into K clusters fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a certain way because different locations cause different results. So, the better choice is to place them as far away from each other as possible. The next step is to take each point belonging to a given data set and associate it to the nearest center. Then we use the elbow method to determine how many clusters we need exactly. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

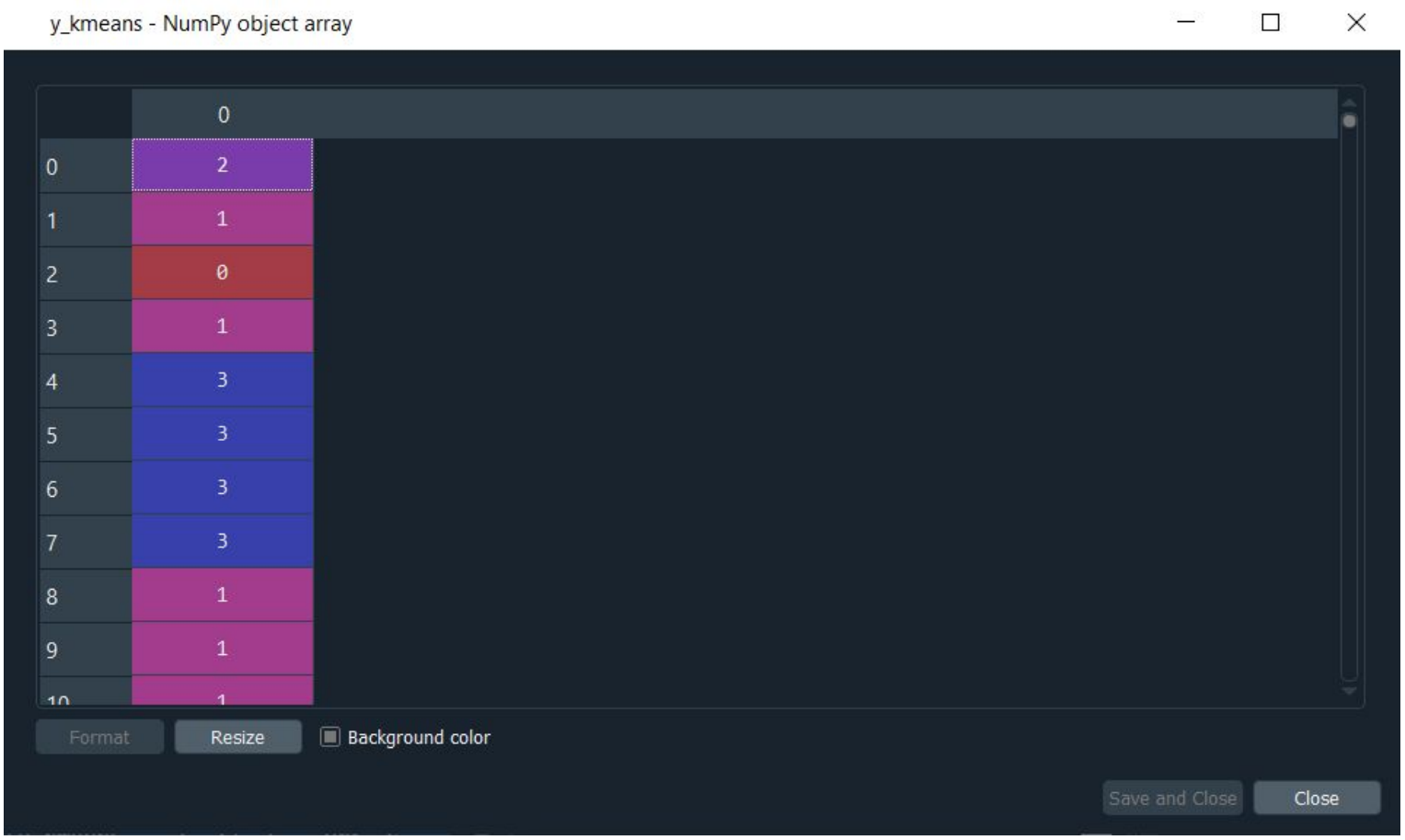
where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

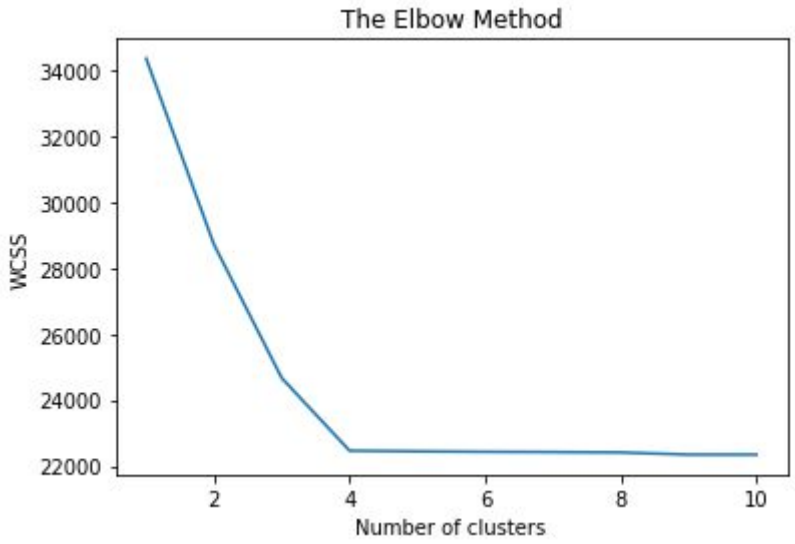
'c' is the number of cluster centers.

Results and Discussions :



C1 page_name

	A	B	C	D	E	F
1	id	facebook_id	page_name	page_type		
2	0	1.45647E+14	The Voice of China 中国好声音	tvshow		
3	1	1.91483E+11	U.S. Consulate General Mumbai	government		
4	2	1.44761E+14	ESET	company		
5	3	5.687E+14	Consulate General of Switzerland in Mont	government		
6	4	1408935539376	Mark Bailey MP - Labor for Miller	politician		
7	5	1.34465E+14	Victor Dominello MP	politician		
8	6	2.82657E+14	Jean-Claude Poissant	politician		
9	7	2.39338E+14	Deputado Ademir Camilo	politician		
10	8	5.44818E+14	T.C. Mezar-ı Şerif Başkonsolosluğu	government		
11	9	2.85156E+11	Army ROTC Fighting Saints Battalion	government		
12	10	2.95295E+14	NASA Student Launch	government		
13	11	8.37707E+14	Eliziane Gama	politician		
14	12	1.89778E+11	Socialstyrelsen	government		
15	13	1.53345E+14	Brisbane Water LAC - NSW Police Forc	government		
16	14	3.74623E+11	NASA's Marshall Space Flight Center	government		
17	15	1.35948E+14	Municipio de Lomas de Zamora	government		
18	16	2.18961E+14	Die Techniker (TK)	company		
19	17	6.03089E+14	Digvijaya Singh	politician		
20	18	1.07219E+11	1st Armored Division Sustainment Brigad	government		
21	19	2.88891E+11	Shapeways	company		
22	20	2.0075E+14	Françoise Guégot	politician		
23	21	3.12379E+14	Hydro Coco	company		
24	22	1.97258E+14	Embassy of the Netherlands in Uganda	government		



After predicting the number of clusters required for the data, we fit the model to the dataset and predict the nodes in each cluster. As shown in the above figures the model is accurately fitted and shows which cluster each data point is assigned to. As python indexing starts from 0, the clusters start from 0 (0,1,2,3).
Here :
Cluster #0 - Company
Cluster #1 - Government
Cluster #2 - TV Show
Cluster #3 - Politician

Conclusions and Recommendations :

Since this is a relatively simple dataset, such a simple unsupervised learning algorithm is more than sufficient to give results with almost 0% error. However if there are more parameters to be analysed and used to cluster the data we can implement neural networks to train the model and then pass it through the clustering algorithm to group the data. Another method that can be used instead of K-means clustering is Hierarchical clustering which uses a dendrogram to find the clusters with maximum distance between each other. Hierarchical clustering typically works by sequentially merging similar clusters, as shown above. This is known as agglomerative hierarchical clustering. However, both the methods mentioned above have shown similar accuracies so using either one is advisable.