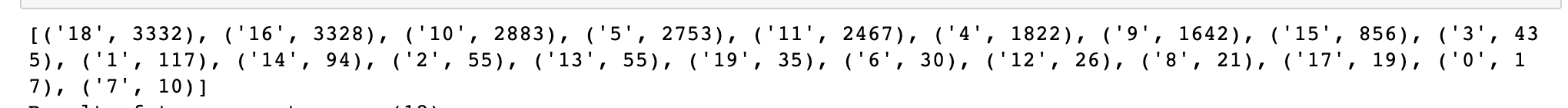
# CS777 Assignment6 Task4 report

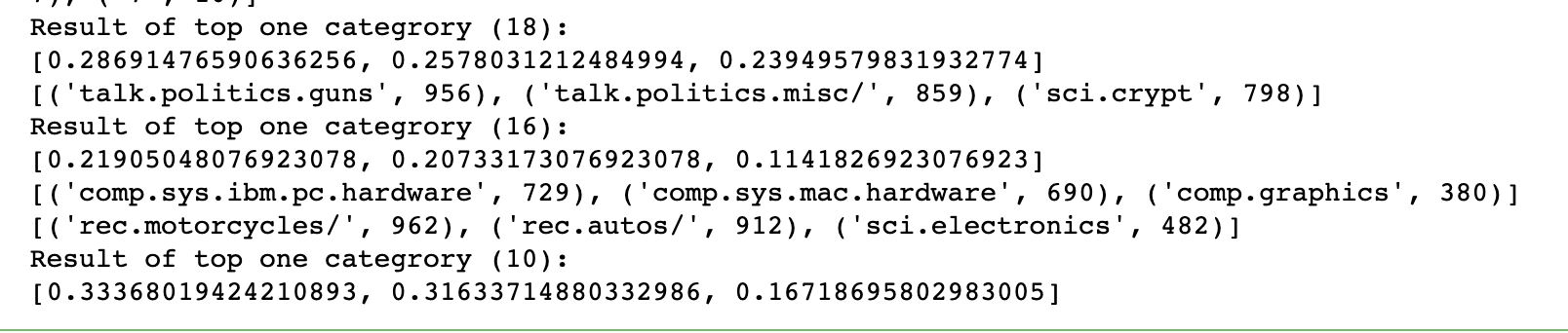
Group Member: Zhaopu Teng, Xiaoyang Wang, Lin Chen

Our clustering algorithms worked, firstly for each of the 20 mixture components we have got 20 clusters output in task2 and task3 name from ‘1’ to ‘20’. Then we print out the number of documents assigned to the cluster as follows:



So, we can get that ‘18’, ‘16’, ‘10’ these three clusters own the most documents. And the documents assigned to them is 3332, 3328, and 2883. And after that is the remaining 17 clusters and the number of documents assigned to them, the results have shown in the figure.

Then for these 3 clusters, we print the top three ‘real’ categories for these 3 clusters:



As shown in the figure, the top 3 categories for cluster ‘18’ is ‘talk.politics.guns’, and the percentage is 0.2869, the number of it is 956, and then is the ‘talk.politics.misc/’, and the percentage is 0.2578, the number of it is 859, then is ‘sci.crypt’, the percentage is 0.2395, the number of it is 798.

For cluster ‘16’ the first one is ‘comp.sys.ibm.pc.hardware’, and the percentage is 0.219, the number of it is 729, and then is the ‘comp.sys.mac.hardware’, and the percentage is 0.207, the number of it is 690, then is ‘comp.graphics’, the percentage is 0.114, the number of it is 380.

For cluster ‘10’ the first one is ‘rec.motorcycles’, and the percentage is 0.3337, the number of it is 962, and then is the ‘rec.autos’, and the percentage is 0.3163, the number of it is 912, then is ‘sci.electronics’, the percentage is 0.1672, the number of it is 482.

We think our results is very good and worked out , because almost each component will have a high prevalence of a particular category especially for the cluster ‘10’