### Part 1 Evaluation

There are a total of 1928 texts in this sample.

The contingency table is used to measure the similarity between clusterings of different algorithms. In Figure 1, the rows represent the clusters created by my implementation of k-means (T1…T5), and the columns (C1…C5) represent the clusters created by SciKitLearn’s k-means. Each cell denotes the number of points that are common to both clusterings. C3 and T1 share many points, so we can induce there is a significant similarity between items in that group and the C3 seems to align with the T1 cluster. The next largest share of points between the two clustering is C3 (again) and T4, with 514 points. This shows a lack of accuracy in my k-means algorithm, or possibly a lack in the number of iterations when partitioning the points. The threshold was set to 1e-5 which led to only 3 iterations of the algorithm.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *C1* | *C2* | *C3* | *C4* | *C5* |
| *T1* | *29* | *0* | *1159* | *0* | *0* |
| *T2* | *0* | *0* | *1* | *0* | *0* |
| *T3* | *2* | *0* | *50* | *0* | *0* |
| *T4* | *41* | *2* | *514* | *4* | *34* |
| *T5* | *0* | *0* | *3* | *0* | *0* |

Figure : Contingency Table. Rows is my k-means implementation. Columns is SciKitLearn’s .

The top 10 most frequent words in each cluster are:

|  |  |
| --- | --- |
| T1 | bill, state, thank, would, California, think, member, need, year, work |
| T2 |  |
| T3 | Aye, think, state, would, know, thank, get, thing, issue, Mitchell |
| T4 | Thank, aye, okay, support, bill, witness, senator, yes, file, much |
| T5 | One, way, first, meeting, grower, look, stress, real, collaborative, process |

A lot of the clusters share common words. Furthermore, they don’t seem to have an even distribution on the amount of words. T1 has *a lot* of points and the frequency of its top words is very high (the top one having a value of about 23.0). Then, the third cluster T3 has far less points and its top word, “aye”, is common among the other clusters as well and only has a frequency of 1.68. Cluster T2 ended up without any points or points with empty data.

I believe my implementation has some issues with how it’s calculating the nearest clusters. The results I am getting give little information to how the information has been grouped and classified.