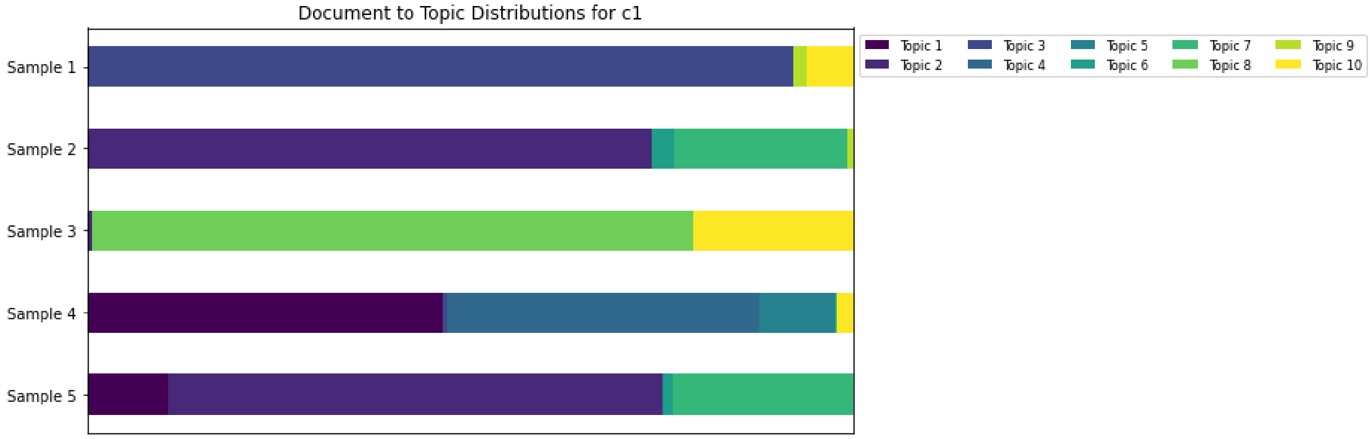
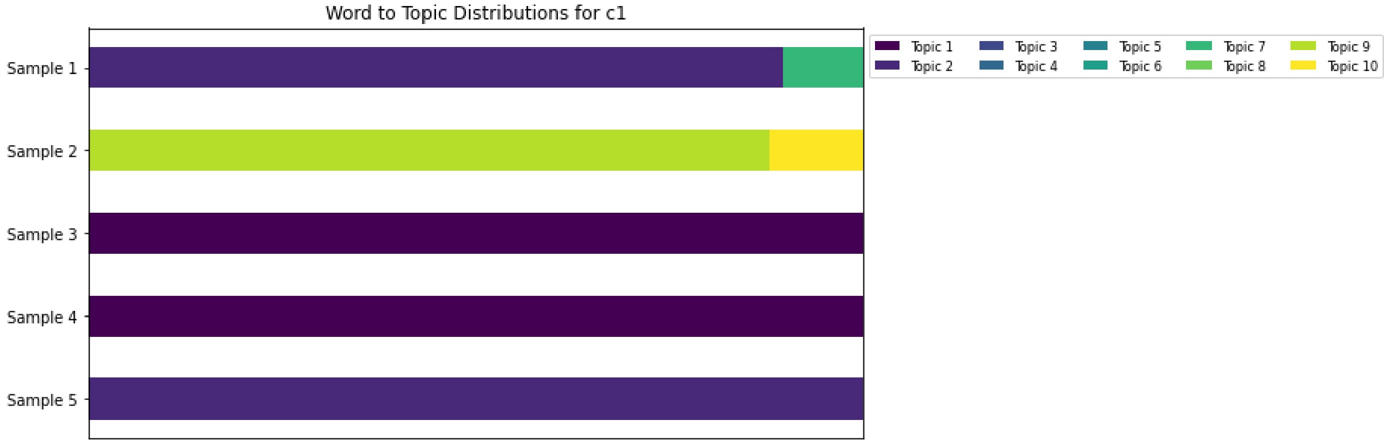
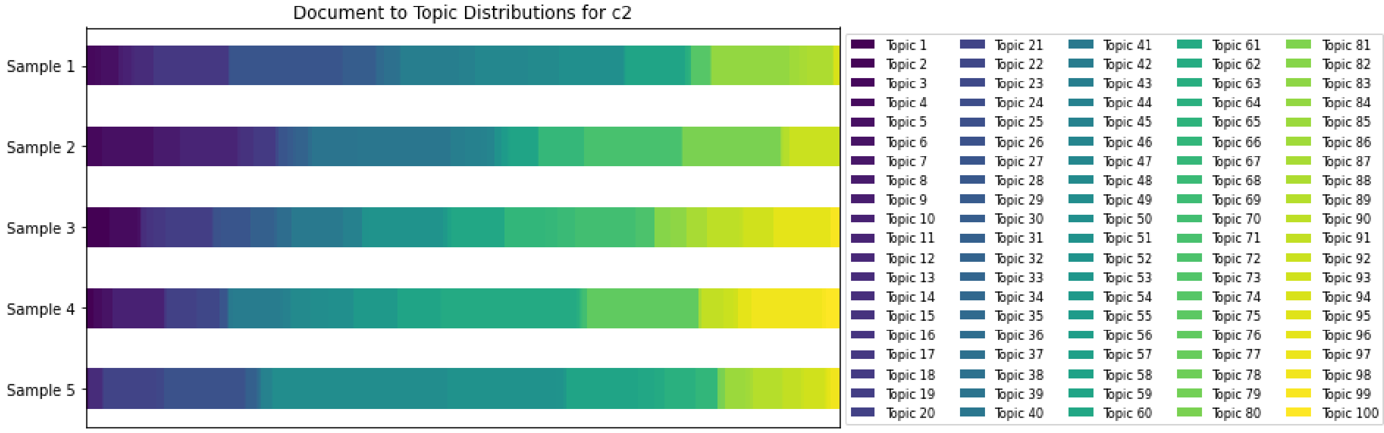
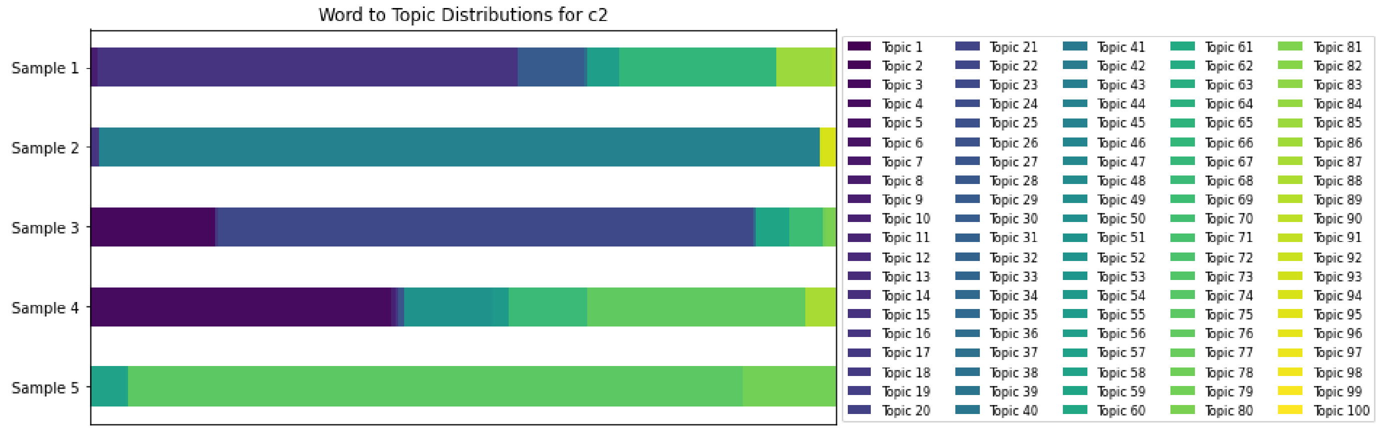
Report

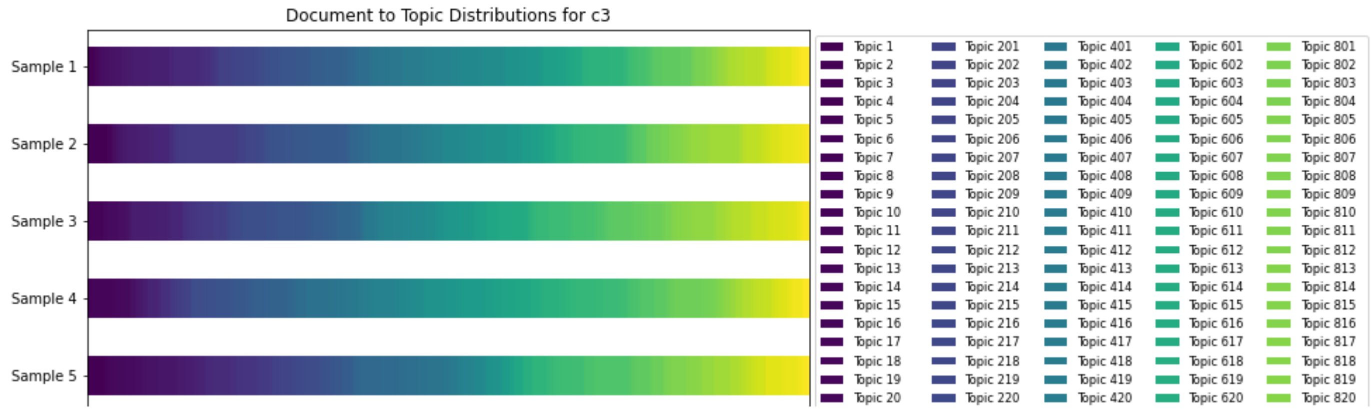
1.


Document to Topic Distributions for c1

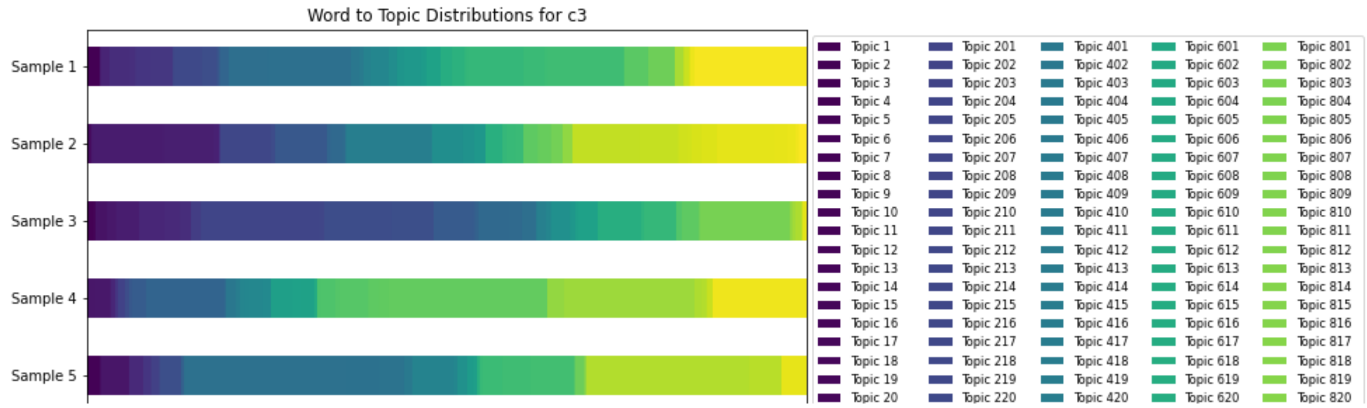

Word to Topic Distributions for c1


Document to Topic Distributions for c2





(Topic labels reduced due to large size)



(Topic labels reduced due to large size)

2.

A document is said to be more ‘multi-topical’ if it has a broader discussion involving many topics instead of a single topic (relates to document-topic distribution). From this definition, c3 exhibits the most ‘multi-topical’ nature, due to its large number of topics (1000), followed by c2 (100) and then c1 (10). This statement is clearly highlighted by the *Document to Topic Distribution* charts, where in c1, the majority of each document sample is made up of about three clearly distinguishable topics, whereas c2 and especially c3 display much more broad (multi-topical) distributions. Specifically for c3, the bar charts appear nearly continuous due to the number of topics in each sample, even whilst having a Dirichlet distribution, which is non-uniform.

Therefore, I would quantify the degree of 'multi-topicality' by how easily unique topics can be distinguished in a sample's *Document to Topic Distribution* bar chart. The easier it is, the less ‘multi-topical’ the collection is. This is mainly affected by the number of topics a collection has. Multi-topicality is also highlighted by how different each sample's distribution is to each other. As can be observed, samples in c1 have very different bar charts, whereas they are close to indistinguishable in c3.

3.

C1 document pair average cosine similarity: 1.143e-3

C1 word pair average cosine similarity: 9.349e-4

C2 document pair average cosine similarity: 1.059e-06

C2 word pair average cosine similarity: 5.822e-07

C3 document pair average cosine similarity: 1.024e-09

C3 word pair average cosine similarity: 1.130e-09

When observing the average cosine similarities of both document and word vectors, it is apparent that c1 has the highest cosine similarity, followed by c2 and then c3. This is because the number of topics per document increases from c1 to c2 to c3 (10, 100, 1000), where a pair of documents with less topics is likely to be more similar than a pair with more topics due to the decreased spread of the distribution. In the same way, words that are distributed across less topics have higher similarity, since there is a greater probability of two words belonging to the same topic. All this being said, even the highest similarity (1.143e-3) is not very high with 1 representing identical vectors, so these results actually show that all document and word pairs are quite different in their distributions.

4.

Purities for c1:

K=20: 0.9794

K=100: 0.9008

K=500: 0.9365

K=5000: 0.9792

Purities for c2:

K=20: 0.19582

K=100: 0.96388

K=500: 0.91648

K=5000: 0.74206

Purities for c3:

K=20: 0.02023

K=100: 0.1006

K=500: 0.42263

K=5000: 0.785

When looking at purities for c1, they are all relatively similar at above 0.9, meaning that all the choices in K result in clusters which are quite pure. This is likely because the number of topics is only 10 – the K-means algorithm executes on 10–dimensional values – meaning that all the values of K are greater than the dimensionality of the problem, which seems to be a requirement for the algorithm to produce pure results.

A more apparent example can be seen in c2 (100 topics), where the purity for K=20 is very low at 0.19582, since the number of clusters are simply not enough to describe the number dimensions in which documents are represented. The purity becomes much better at K=100 with 0.96388, since now the number of clusters and topics are equal. A similarly good result can be found at K=500, however the purity drops to 0.74206 at k=5000. This is likely because creating too many clusters has resulted in artificial boundaries within real clusters, resulting in data points being assigned to incorrect clusters, indicating that K has an optimal value. It is also important to note that c2 and c3 have 100K documents instead of 10K in c1, causing more inaccuracies in the clustering algorithm.

The hypothesis that the number of clusters must be greater than the number of dimensions is further supported by the purities found for c3. Since c3 has 1000 topics, the first three values of K (20, 100, 500) all report very poor purities for their generated clusters. K=20 being extremely low at 0.02023, shows that the lower K is compared to the dimensionality of the problem, the worse the outcome gets. However, even at K=5000 (greater than dimensionality), the cluster still only has a purity of 0.785, which is marginally lower than some of the more optimal results that can be found in c1 and c2. This is likely because working with such large dimensions (1000) introduces a lot of spread in the distribution as well as uncertainty, causing the algorithm to produce less pure clusters.