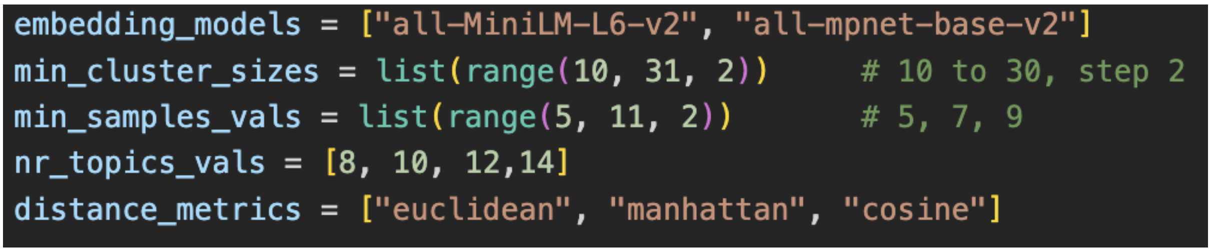
BERTopic was used to find the best clustering of topics, which in turn are used for the one-shot LLM to classify a given post under a certain topic. The hyperparameters were chosen based on the results yielded by a grid search consisting of 594 unique combinations. The following parameters were tuned. An exhausted list can be found in *bertopic_grid_log.csv*.

! Sampled posts ! ← FINISH HERE

0.99 yes climate labeled posts + 60 chars or more



```
embedding_models = ["all-MiniLM-L6-v2", "all-mpnet-base-v2"]
min_cluster_sizes = list(range(10, 31, 2))     # 10 to 30, step 2
min_samples_vals = list(range(5, 11, 2))       # 5, 7, 9
nr_topics_vals = [8, 10, 12,14]
distance_metrics = ["euclidean", "manhattan", "cosine"]
```

The 594 different combinations of topics were evaluated based on the following criteria:

- n_topics: A balance between the clusters being too inclusive and too exclusive. The topics should be coherent while remaining general.
- outlier_pct: The percentage of outliers should be low as to include most of the data in the clusters. High outlier percentage indicates poor clustering.
- Keyword coherence: The top words should be semantically similar
- Topic Distinctiveness: Each topic should be clearly different from the rest. Similar topics are to be merged into one.
- Size distribution: The clusters should be proportionally similar in size. The samples should not be represented by one large cluster and many small ones.

We found X X X X X X X X X X X X X to yield the best topic clusters.

These topics are used for the one-shot LLM to classify a given post into a subtopic.