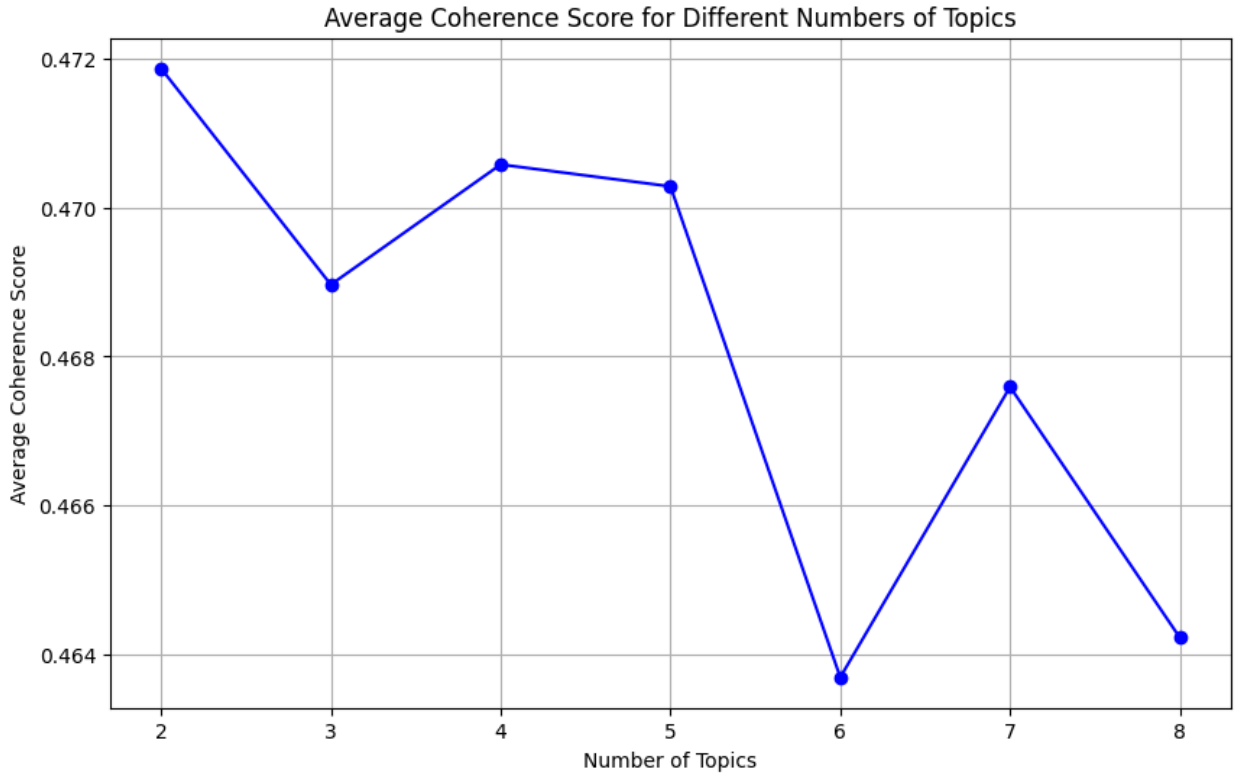
Sensitivity Analysis

This analysis aims to determine the optimal number of topics for the NYT dataset. By evaluating the coherence scores across different numbers of topics, we aim to identify the configuration that produces the most coherent and meaningful topics. This process helps in uncovering the underlying structure and themes within the dataset.



Observations

Average Coherence Score Trend

As the number of topics increases from 2 to 8, there is a slight fluctuation in the average coherence score. The average coherence score ranges from approximately 0.464 to 0.472, indicating a relatively consistent performance across different numbers of topics.

Optimal Number of Topics

The highest average coherence score is achieved with 2 topics, with a score of approximately 0.472. As the number of topics increases beyond 2, there is a slight decrease in the average coherence score, with scores ranging between 0.463 and 0.468 for 6 to 7 topics.

Interpretability

A lower number of topics (e.g., 2) may result in higher coherence scores, indicating more interpretable and distinct topics. However, increasing the number of topics beyond a certain threshold (e.g., 6 or 7) may decrease coherence scores, suggesting that the topics become more granular and potentially less interpretable.

Robustness

The LDA model demonstrates robustness to changes in the number of topics, as the average coherence scores remain relatively close to each other across different configurations. This suggests that the model is not overly sensitive to small variations in the number of topics within the specified range.

Considerations

The choice of the number of topics should consider a balance between coherence and granularity. A higher coherence score does not necessarily imply better topic quality if the topics become too broad or overlapping. Evaluating the coherence scores in conjunction with qualitative assessments of the topics is essential to ensure they are meaningful and interpretable.

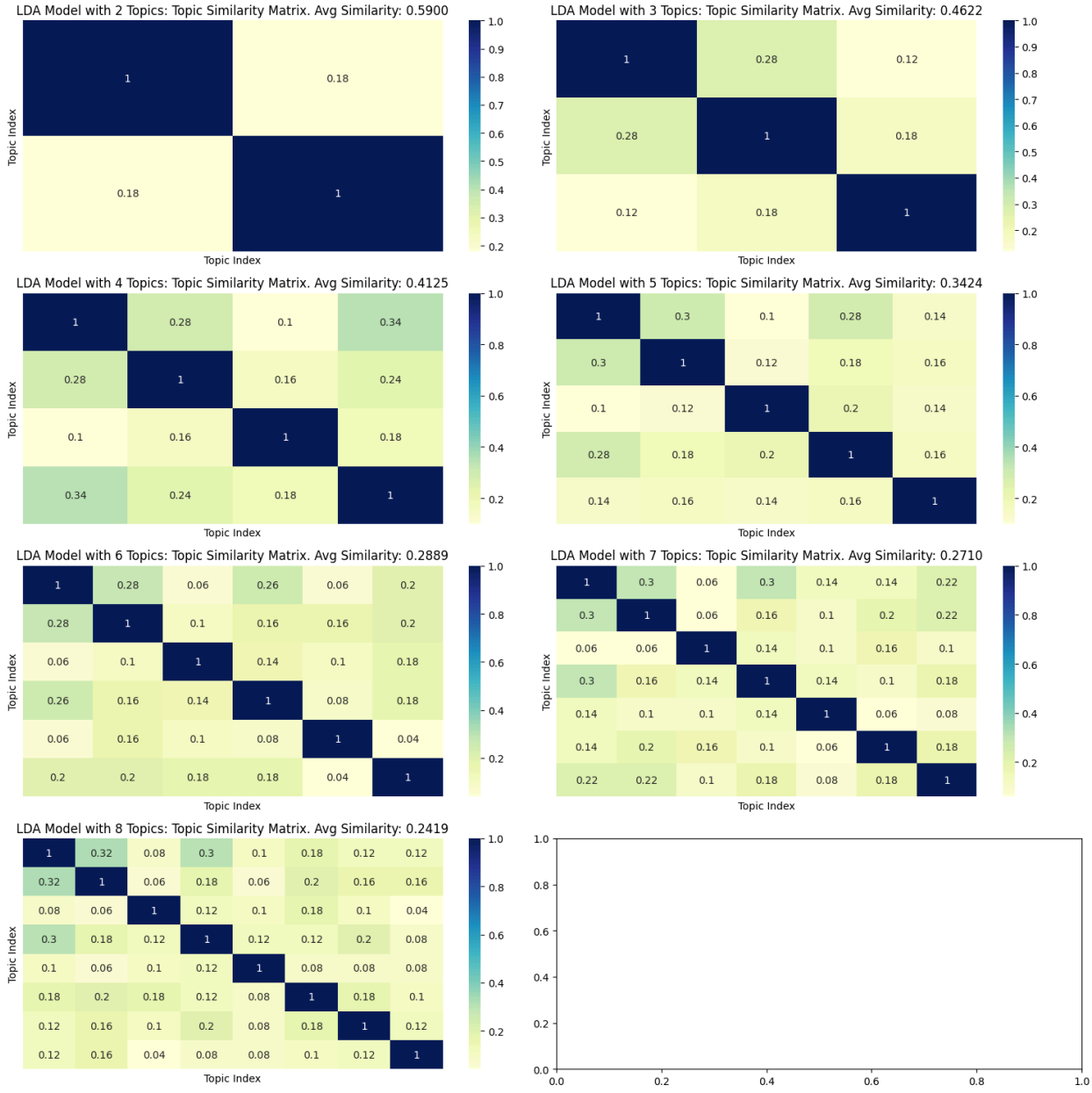
So, is n_components=2 optimal?

Using an LDA model with only 2 topics for classifying articles may not be the most effective approach for several reasons:

1. With only 2 topics, the LDA model may struggle to capture the diverse range of topics in NYT articles effectively. This lack of granularity can result in topics that are too broad or overlapping, making it challenging to classify articles into meaningful categories accurately and resulting in the lack of topic differentiation.
2. Limiting the number of topics to 2 can lead to a significant loss of information. NYT covers many topics spanning politics, business, sports, entertainment, science, and more. A binary classification of articles into just two broad topics would not adequately capture this content richness. This results in a loss of information
3. A binary classification model may lack the discriminative power to distinguish between closely related topics or subtopics within the same category. This can result in misclassification of articles and reduced performance in tasks such as recommendations.

Topic Similarity

We want to quantify the similarity or dissimilarity between topics identified by different LDA configurations. This allows for a direct comparison of topics generated by LDA models with different numbers of topics. By visualising the similarity matrices, we can identify patterns and trends in how topics evolve as the number of topics varies.



Observations

The LDA models were constructed with varying numbers of topics ranging from 2 to 8.

Average Similarity Trend

As the number of topics increases, the average similarity score decreases consistently. This suggests that as the LDA model becomes more granular with a higher number of topics, the topics become less similar to each other on average.

Interpretability

Higher similarity scores indicate greater overlap or coherence between topics, suggesting that the topics are more closely related or represent similar themes. Conversely, lower similarity scores imply that the topics are more distinct or diverse from each other.

A higher similarity score may indicate a more cohesive set of topics, but it could also imply a lack of granularity. Conversely, a lower similarity score may suggest more distinct topics but could lead to potential overlaps or ambiguity in topic interpretation.