Summary:

Improving Coherence Scores in topic modeling on clinical text

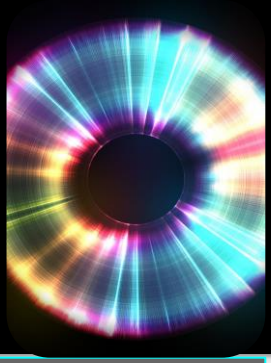
Topic modeling using Latent Dirichlet Allocation (LDA) can be a powerful tool in identifying semantic groups in vast unstructured textual data like discharge notes. **One of the ways to evaluate the degree to which learned topics are meaningful and interpretable is to calculate the Coherence Score under varying ranges of the model hyperparameters.**

In this case, **Coherence scores have been calculated and visualized for different combinations of key hyperparameters** from topic modeling of clinical text. The results can be used for

1. **Further tuning of hyperparameters.**
2. **Checking model performance under varying rules of text pre-processing.**
3. **Validation of top topics, their distribution per note, and distribution of words associated with each topic.**

Repository on GitHub: https://github.com/olga12kz-DS/LDA-Coherence-Scores/tree/main

Results of Hyperparameter Tuning and its Impact on Model Coherence Scores



As seen in the graph, the best performance is found with:

- **Number of topics: 20** (lowest of the two options). Setting the number of topics to a lower value encourages the model to cluster information more broadly rather than fragmenting it into overly granular topics.
- **Alpha: 0.5 (lowest of the two options).** This hyperparameter controls how evenly topics are distributed across notes. A moderate value like 0.5 allows for a balanced mix per note.
- **Eta: 0.8 (highest of the two options).** Eta controls the sparsity of topic-word distributions. A higher eta value here leads to denser topic-word distributions, appropriate for clinical notes, where many words are likely to be shared across multiple topics.

In conclusion, this corpus of 2,000 discharge notes can be characterized as covering a moderate number of broader topics, with a balanced topic mix per note, and substantial word sharing across topics, a pattern consistent with the overlapping and comprehensive nature of clinical narratives.

Testing Varying Hyperparameters Ranges