**Movie Dialogues Topic Modeling**

This project implements a movie dialogue topic recommendation system that identifies similar movies based on the topic similarity within conversations. The system uses natural language processing techniques to analyze movie conversations and generate movie similarities based on topics modeled using Latent Dirichlet Allocation (LDA).

**Core Model Components**

* **LDA (Latent Dirichlet Allocation) Topic Model**
  + Model Name: sklearn’s LatentDirichletAllocation
  + Architecture: A generative probabilistic model that uses two levels of Dirichlet priors: one for document-topic distributions and one for topic-word distributions
  + Dimensions: The topic-word matrix is adjusted to (10, 20670) to match the vectorizer
  + Training: The model is trained using the training subset of the data after vectorization. Key training metrics include perplexity (evaluating generalization) and log-likelihood (evaluating the fit of the model).
  + Purpose: To recommend movies based on the similarity of their extracted topic distributions.
* **Text Preprocessing Pipeline**
  + Tokenization: The raw text was split into individual tokens (words).
  + Lowercasing: All text was converted to lowercase to ensure uniformity.
  + Stopword Removal: Common, uninformative words were removed using a predefined stopword list.
  + Lemmatization: Words were reduced to their base form to consolidate variants of the same term.
  + Punctuation and Special Character Cleanup: Punctuation and extraneous characters were removed to reduce noise.
* **Similarity Computation**
  + Method: Cosine similarity between feature vectors representing movies.
  + This similarity score is then used to rank or filter movies for recommendation purposes.

**Performance Metrics**

* Coherence: 0.86 Average for 10 Topics (Higher values indicate more interpretable topics)
* Perplexity: 4382.17 on Test Data (Lower Values indicate better generalization)
* Log-Likelihood: -1722882 on Test Data (Indicates relatively poor fit)
* Precision: 0.59 (59% of recommended movies are relevant based on genre overlap)
* Recall: 0.59 (59% of relevant movies appear in the top 5 recommendations)
* F1 Score: 0.59 (Harmonic mean of precision and recall)

**Analysis and Insights**

**Topic Quality Variation**: There is significant variation in topic coherence scores (0.0749 to 2.1470), indicating that some topics are much more interpretable than others. Topics 2 and 4 demonstrate high coherence, suggesting they capture meaningful patterns in the movie dialogue data.

**Low Coherence Topics**: Topics 8 and 9 have very low coherence scores, suggesting they may be capturing noise rather than meaningful patterns. These topics might benefit from refinement or could be candidates for removal in a revised model.

|  |  |  |
| --- | --- | --- |
| **Topic** | **Coherence Score** | **Relative Coherence** |
| Topic 0 | 0.289 | Low |
| Topic 1 | 1.165 | Medium |
| Topic 2 | 2.147 | High |
| Topic 3 | 0.385 | Low |
| Topic 4 | 2.111 | High |
| Topic 5 | 1.417 | Medium-High |
| Topic 6 | 0.462 | Low |
| Topic 7 | 0.429 | Low |
| Topic 8 | 0.075 | Very Low |
| Topic 9 | 0.122 | Very Low |
| Average | 0.8601 |  |

Additionally, the coherence score was found to be optimal at 10 topics aligning to the distribution of movie genres within the corpus. This was also the same for perplexity where the increase in topics dramatically increased the perplexity score.

A graph of blue bars with names

AI-generated content may be incorrect.

A comparison of a graph

AI-generated content may be incorrect.

**Balanced Precision and Recall**: The identical values for precision and recall (0.5903) suggest that the recommendation approach is balanced in terms of relevance and coverage.

**Genre-Based Relevance**: The F1 score of 0.5903 indicates moderate performance in recommending movies with similar genres. This is promising for a content-based approach using only topic distributions derived from dialogue.

**Error Analysis**

**Low Coherence Topics**: Topics 8 and 9 have very low coherence scores (0.0749 and 0.1220), indicating they may not represent meaningful patterns. Analysis of the top words in these topics might reveal overly generic terms or rare words that don't form coherent themes.

**Genre Overlap Limitations**: Using genre overlap as the sole criterion for relevance may not capture the nuanced similarities between movies. Two movies might share genres but have very different tones, themes, or target audiences.

**Topic Interpretability**: The wide range of coherence scores suggests that some topics may be capturing noise or very specific patterns that don't generalize well.

**Improvement Strategies**

**Optimize Number of Topics**: The current model uses 10 topics, but further experimentation with different numbers could improve overall coherence and interpretability.

**Vocabulary Refinement**: Review and potentially expand the stopword list to remove common but uninformative terms from the analysis.

**Document Preprocessing**: Enhance the text preprocessing pipeline with more sophisticated lemmatization, named entity recognition, or phrase detection.

**Weighted Genre Matching**: Instead of binary genre overlap, implement a weighted approach that considers the prominence of each genre in a movie.