How the Code Answers Key Research Questions Using LDA

This document explains how the provided Python code addresses three research questions about children's career development using Latent Dirichlet Allocation (LDA), and why LDA is effective for extracting topics from empirical literature.

1. What specific topics have emerged from the empirical literature on children's career development?

The code applies Latent Dirichlet Allocation (LDA) to a corpus of academic texts related to children's career development. LDA is a probabilistic model that assumes each document is a mixture of various topics, and each topic is represented as a distribution over words. By training an LDA model, the code uncovers hidden topic structures by identifying clusters of words that frequently co-occur across documents. These clusters are interpreted as coherent topics. The code then displays the top words in each topic, revealing what each topic is about, thereby answering the first part of the question.

Additionally, for each topic, the most relevant terms (those with the highest probabilities in that topic's word distribution) are listed. These represent the specific terms that emerged under each topic in the empirical literature.

2. How do the topics of career development of children change over time?

To understand topic evolution, the code associates each document with its year of publication. After the LDA model assigns topic probabilities to each document, the code aggregates these topic distributions by year. It then calculates the mean presence of each topic per year, which reflects how prominent each topic was in that time period. This data is visualized with time series plots, showing the temporal dynamics of each topic. These visualizations help answer how the prominence of specific topics changes over time in the literature.

3. What are the topics with increasing/decreasing trends, and which topics remain consistently popular during this period?

To quantify topic trends, the code performs linear regression on each topic's yearly average prevalence. By computing the slope of the topic's trend line, it determines whether the topic is increasing, decreasing, or stable over time. Topics with positive slopes are identified as growing in importance, while those with negative slopes are declining. Topics with near-zero slopes are considered consistently popular. This statistical approach provides a rigorous way to detect and describe long-term patterns in topic popularity.

Why LDA Can Extract Topics

1. Assumption About How Text is Generated

LDA assumes that:
- Each document is a mixture of latent topics.
- Each topic is a probability distribution over words.
- Each word in a document is generated by:
  1. Sampling a topic from the document's topic distribution.
  2. Sampling a word from the chosen topic's word distribution.

This structure allows LDA to reverse engineer the hidden (latent) topic structure in the data based on the observed words in documents.

2. Unsupervised Learning via Inference

LDA uses algorithms like:
- Variational inference or
- Gibbs sampling

...to infer the hidden topic structure:
- What topics exist.
- What words belong to which topic (topic-word distribution).
- What topics are present in each document (document-topic distribution).

Even without any labels, LDA can group co-occurring words into coherent topics because they tend to appear together across documents.

3. Topic = Cluster of Co-occurring Words

For example:
- If words like 'career', 'aspiration', 'motivation', 'future' often appear together, LDA might create a topic called 'Career Motivation'.
- If 'parent', 'influence', 'support', 'role' co-occur, LDA might extract a 'Parental Influence' topic.

The algorithm identifies these patterns statistically, not semantically.

4. Intuition Recap (Simplified)

Imagine a giant bag of documents:
- LDA thinks: 'These documents are made by mixing a few hidden topics.'
- It tries to find those topics and how much each topic contributed to each document.
- Each topic is just a collection of words that often occur together.

5. Technical Foundations

- Dirichlet distributions are used as priors:
  - For document-topic distribution (θ) — how much each topic contributes to a document.
  - For topic-word distribution (φ) — how much each word contributes to a topic.
- The model maximizes the likelihood of observing the documents under this generative process.

6. Summary

Feature	Explanation
Latent topics	Hidden groups of co-occurring words
Document-topic mixture	Each document is composed of several topics
Topic-word distribution	Each topic favors certain words
Learning method	Probabilistic inference (e.g., Gibbs sampling)
Why it works	Finds statistical regularities in word co-occurrence patterns