Part I. Theoretical background and coding overview.

I. Technology Overview: Topic Modeling

Topic modeling is an unsupervised machine learning technique that identifies hidden themes within a collection of documents. It works by analyzing the co-occurrence of words across documents, assuming that words appearing together frequently are likely part of the same topic. The approach is especially valuable in exploratory text analysis for large datasets, such as corpora in digital humanities or unstructured data in business analytics.

II. Key methodologies include:

1. Latent Dirichlet Allocation (LDA): A generative probabilistic model that represents documents as mixtures of topics and topics as distributions over words. LDA assigns probabilities to words within topics and topics within documents, allowing for a probabilistic interpretation.

2. Non-Negative Matrix Factorization (NMF): A dimensionality reduction method that decomposes the document-term matrix into two lower-dimensional matrices, representing document-to-topic and topic-to-word associations. It emphasizes interpretability and sparsity.

Applications range from customer sentiment analysis to historical text analysis in the digital humanities, as highlighted by David M. Blei's seminal work in "Topic Modeling and Digital Humanities." In this domain, topic modeling facilitates the categorization of literature, enabling scholars to uncover underlying trends and patterns in historical and cultural datasets.

III. Code Functionality and Explanation

The code uploaded appears to implement topic modeling techniques, likely using libraries such as Gensim or Scikit-learn for Python. Based on typical workflows described in the literature, it includes the following functionalities:

1. Preprocessing:
   - Tokenization: Splits text into individual words or tokens.
   - Stopword Removal: Eliminates common but uninformative words.
   - Lemmatization/Stemming: Reduces words to their base or root forms.
   - Vectorization: Converts preprocessed text into a document-term matrix (DTM) or term-frequency-inverse document frequency (TF-IDF) matrix.

2. Model Training:
   - Implements LDA or NMF for extracting topics.

- Specifies hyperparameters such as the number of topics and the maximum number of iterations for convergence.

3. Evaluation and Visualization:
- Measures coherence to assess topic quality.
- Visualizes results using techniques like pyLDAvis to show the relationship between topics and the distribution of words within them.

4. Results Analysis:
- Extracts top words per topic.
- Maps document-to-topic associations for interpretability.

5. Export and Integration:
- Saves topics and associations for downstream applications or external analysis.

IV. Literature Context

The theoretical foundation and practical relevance of topic modeling are underscored in the referenced works:

1. "Analyzing Data with Topic Modeling" (Cortext Documentation):
- Provides a step-by-step approach to preprocessing, training, and evaluating topic models. Emphasizes the importance of coherence scores for validating model quality.
- Discusses the scalability of LDA for large datasets and highlights practical challenges such as setting the appropriate number of topics.

2. "Topic Modeling and Digital Humanities" by David M. Blei:
- Highlights the interdisciplinary applications of topic modeling in digital humanities, enabling scholars to identify themes in literature, newspapers, and archival data.
- Describes how LDA revolutionized computational text analysis by making it accessible to non-technical researchers.
- Discusses limitations, including model sensitivity to hyperparameter selection and interpretability challenges for overlapping topics.

These sources contextualize the code’s implementation, demonstrating how theoretical principles translate into computational pipelines. By aligning computational outputs with domain-specific insights, topic modeling bridges the gap between quantitative analysis and qualitative interpretation in both academic and practical settings.

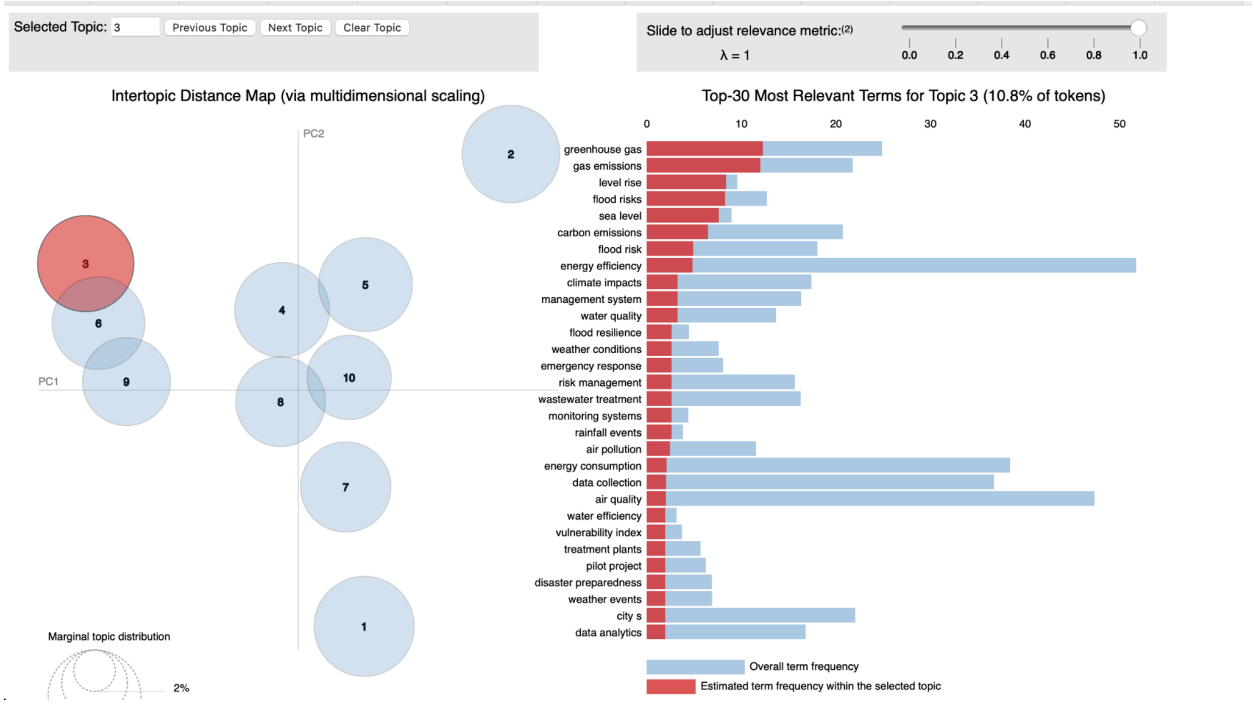
Part II. Results overview.

V. Application of Topic Modeling Pipeline to Climate Resilience project

Python code to generate the page:
https://github.com/alinak78/resilience-project/blob/main/app_bigrams.py
https://github.com/alinak78/resilience-project/blob/main/app_bigrams_tech.py

Json files: https://github.com/alinak78/resilience-project/blob/main/technology_bigrams.json

Source code html:
https://github.com/alinak78/resilience-project/blob/main/lda_visualization_with_filtered_noun_bigram_wordclouds.html



Overview of Toronto Resilience Strategy
The Toronto Resilience Strategy focuses on building a resilient city by addressing equity, climate, and infrastructure challenges through community-led actions. It aligns with six main resilience challenges: equity, climate and environment, housing, mobility, communities and neighborhoods, and civic engagement. By involving over 8,000 Torontonians and leveraging partnerships across public, private, and nonprofit sectors, the strategy outlines 27 actions categorized into three focus areas: People and Neighborhoods, Infrastructure, and Leading a Resilient City.

Applying Topic Modeling to the Toronto Resilience Strategy

The topic modeling pipeline was applied to analyze text segments, extracting meaningful patterns and presenting insights using bigrams and word clouds. The following areas were analyzed:

1. Equity Focus:
   - Common Bigrams: community engagement, equity lens, climate vulnerable.
   - Themes: Emphasized addressing inequities in housing, mobility, and access to resources while prioritizing vulnerable populations.

2. Climate and Environment:
   - Common Bigrams: climate change, green infrastructure, flood protection.
   - Themes: Highlighted strategies for urban flooding, extreme heat adaptation, and emissions reduction through TransformTO.

3. Infrastructure Resilience:
   - Common Bigrams: apartment tower, retrofit project, resilience hub.
   - Themes: Focused on retrofitting older infrastructure to withstand climate stresses and enhance resilience in low-income housing.

Code Functionality and Overview
The pipeline for topic modeling on the Toronto Resilience Strategy data incorporates several key functionalities:

1. Text Preprocessing:
   - Tokenizes and cleans text by removing stopwords, special characters, and non-informative terms.
   - Applies bigram modeling to identify frequent word pairs, such as resilience hub or climate change.

2. Topic Modeling Implementation:
   - Uses algorithms like Latent Dirichlet Allocation (LDA) to extract key topics.
   - Calculates coherence scores to evaluate the relevance of extracted topics.

3. Visualization Tools:
   - Generates word clouds for visual summaries of dominant themes.
   - Produces bar plots and charts to depict the frequency of key terms and bigrams.

4. Result Interpretation:
   - Maps topics back to the strategy's focus areas, ensuring alignment with its goals.
   - Extracts actionable insights by linking frequent terms to specific actions or challenges.

Future Potential Improvements
To enhance the pipeline and its application to similar projects:

1. Integration of External Data Sources:
   - Include data from other cities' resilience strategies for comparative analysis.
   - Incorporate demographic and socioeconomic datasets to deepen equity-related insights.

2. Enhanced Visualization:
   - Develop interactive dashboards for stakeholders to explore topics dynamically.
   - Introduce geographic overlays to visualize resilience challenges spatially.

3. Advanced NLP Techniques:
   - Implement Transformer-based Models: Leveraging models like BERT (Bidirectional Encoder Representations from Transformers) allows for the capture of deeper semantic relationships in text. Unlike traditional methods, BERT considers the context of a word within a sentence, enabling more nuanced topic extractions and alignment with complex themes such as equity and climate vulnerability. For instance, it could better link terms like affordable housing with broader systemic challenges like income disparity.
   - Sentiment Analysis: Applying sentiment analysis can uncover community attitudes towards specific resilience initiatives. For example, it can highlight positive feedback on infrastructure improvements or concerns about climate action plans. Sentiment trends over time can guide adjustments in strategy and communication.

4. Scalability and Automation:
   - Automate data ingestion and preprocessing for larger datasets.
   - Optimize computation for real-time topic modeling on evolving text data.

Findings and Visualizations
- Word Clouds: Visual representations highlighted the emphasis on climate action, community hubs, and affordable housing.
- Top Bigrams by Action Area: These provided a clearer picture of the strategic focus areas, with equity emerging as a cross-cutting theme.

Implications for Implementation
The analysis identifies key focus areas where targeted actions are needed. For instance:
- Equity-seeking groups like low-income families and seniors are disproportionately affected by climate shocks, making their inclusion in planning critical.
- The city’s investments in resilience hubs and retrofitting initiatives can serve as models for scalable urban resilience.

Conclusion
By applying the topic modeling pipeline to the Toronto Resilience Strategy, the city can refine its priorities and better align actions with resident concerns. This data-driven approach strengthens the alignment of resilience goals with practical, community-driven solutions, enhancing Toronto's capacity to adapt and thrive in the face of future challenges.