UN Activist Group Summary Statement Analysis

Project Overview

This project analyzes the summary statements from 20 activist groups who presented their positions at the United Nations (UN). The objective is to extract, summarize, and analyze the text from these PDF files to identify argumentation patterns and strategies employed by activist groups from different regions. Specifically, the goal is to understand the kinds of arguments made, identify regional trends in argumentation, and provide insights into how groups frame their positions on issues like human rights, environmental protection, and governance.

Problem Statement

Activist groups worldwide present summary statements at the UN, advocating for causes ranging from human rights to environmental protection. However, little is known about how these groups structure their arguments, and whether there are common patterns or strategies tied to their geographical location or cause. This project aims to answer the following questions:

1. What are the key themes and argumentation styles used by activist groups at the UN?
2. Are there noticeable regional patterns in the way these groups frame their arguments?
3. How can activist groups optimize their argumentation strategies to improve their advocacy at global forums like the UN?

Key Objectives:

- Extract textual content from 20 PDF summary statements.
- Summarize the core arguments presented by the activist groups.
- Discover and analyze patterns in argumentation styles based on the groups' geographical regions.
- Make recommendations for activist groups to improve their advocacy efforts at the UN.

Dataset

- **Source**: The dataset comprises 20 PDF documents of summary statements from activist groups that presented to the UN. Each PDF contains a detailed statement from a different group, with topics ranging from human rights to environmental concerns. The original pdf file can be found at https://github.com/shrekapoor99/Springboard/blob/main/NGOTextFile.pdf.
- **Data Contents**: Each document represents a different activist group, with content that includes appeals on human rights, environmental sustainability, and regional governance. The documents are structured in various formats, with some being longer and others more concise, but all focused on advocating for specific global or regional issues.

- **Regions Represented**: The dataset includes groups from regions like South Asia, Latin America, Africa, and Europe, offering a wide diversity of argumentation styles and issue focus.

Methodology

This section outlines the step-by-step process of how the data was processed, summarized, and analyzed to extract meaningful patterns. The methodology was divided into three core steps: data extraction, text summarization, and pattern discovery.

1. Data Extraction

The first task was to extract the raw text from the 20 PDF files. Given that each PDF was structured differently, the process required careful extraction and cleaning to ensure accuracy. The preprocessed segments can be found at https://github.com/shrekapoor99/Springboard/blob/main/preprocessed_segments.pdf.

Here's a detailed breakdown:

- **Libraries Used**:
  - I used the `PyPDF2` and `fitz` (PyMuPDF) Python libraries to open, parse, and extract the content from each PDF.
- **Text Extraction Process**:
  - A custom function was written to loop through each page of each PDF and extract the text. This function also dealt with variations in the PDFs (such as differences in formatting and layouts) to ensure that only the clean text was captured, while ignoring irrelevant content like headers, footers, or page numbers.
- **Error Handling**:
  - Some PDFs included images, tables, or non-textual elements that initially interfered with the text extraction process. These cases were handled by applying regular expression (regex) techniques to filter out noise and keep only the meaningful text.

2. Text Summarization

Once the text was extracted, the next step was to summarize the content of each activist group's statement to highlight the key points and themes.

- **Summarization Technique**:
  - The text-rank algorithm was used to generate summaries of each document. Text-rank is a graph-based ranking model for extracting key sentences by evaluating the importance of sentences based on word frequencies and relationships across the document.
- **Key Steps**:
  1. **Sentence Tokenization**: Each document was broken down into individual sentences.

    2. **Importance Ranking**: Each sentence was ranked based on the relevance of the words it contained in the context of the entire document.
    3. **Summarization**: Sentences with the highest importance scores were extracted to form a coherent summary of the statement.
- **Manual Adjustment**: After the automatic summarization, manual adjustments were made to ensure the summaries were accurate, concise, and reflective of the original intent of each activist group.

3. Pattern Discovery

The primary goal of the project was to discover patterns in the way different activist groups from various regions framed their arguments. This was achieved by categorizing the types of arguments and correlating them with geographical regions.

- **Argument Categorization**:
    - Arguments were categorized into several types:

1. Environmental Arguments

- **Description**: Appeals that focus on sustainability, climate change, conservation, and the responsible management of natural resources. These arguments often emphasize the need to protect the environment for future generations and address global challenges related to ecological preservation.
- **Keywords**: sustainability, environment, ecology, conservation, green, renewable, recycling, biodiversity, nature, organic, eco-friendly, sustainable development, climate change, carbon footprint, earth, ecosystem, natural resources, pollution, renewable energy, conservationism.

2. Reputation-Based Arguments

- **Description**: Arguments grounded in the reputation or status of a country, region, or group, often linked to human rights records, national prestige, or public recognition on a global scale. These arguments are used to assert moral or ethical superiority.
- **Keywords**: honor, fame, reputation, prestige, recognition, status, celebrity, dignity, glory, esteem, rank, notoriety, prominence, distinction, respectability, nobility, acclaim, veneration, admiration, legacy.

3. Domestic Issue Advocacy

- **Description**: Arguments centered on national or local governance, social issues, or domestic policies. These arguments often focus on maintaining cultural traditions, fostering community bonds, and addressing problems that directly affect the country or region.

- **Keywords**: tradition, family, loyalty, heritage, trust, kinship, roots, ancestry, home, lineage, nurturing, patriarchy, matriarchy, fidelity, domesticity, respect, belonging, elders, customs, genealogy, family tree, family history.

4. Inspirational Arguments

- **Description**: Arguments that focus on creativity, innovation, and artistic expression. These appeals are often rooted in inspiration, passion, and the idea of pushing beyond boundaries to achieve transcendence or artistic excellence.
- **Keywords**: creativity, inspiration, art, innovation, imagination, originality, intuition, passion, spirit, vision, aesthetic, muse, genius, artistic, spontaneity, expression, idealism, transcendence, freedom, inventiveness.

5. Civic Arguments

- **Description**: Arguments related to social responsibility, democracy, and community welfare. These arguments emphasize the collective good, equality, justice, and participation in civic life. They often advocate for unity and solidarity to achieve common goals.
- **Keywords**: solidarity, collective, social, public, community, welfare, common good, democracy, civic, participation, equality, justice, citizenship, public spirit, social responsibility, unity, altruism, cooperation, commonwealth, public interest, social contract, social cohesion, social capital, social justice.

6. Market-Based Arguments

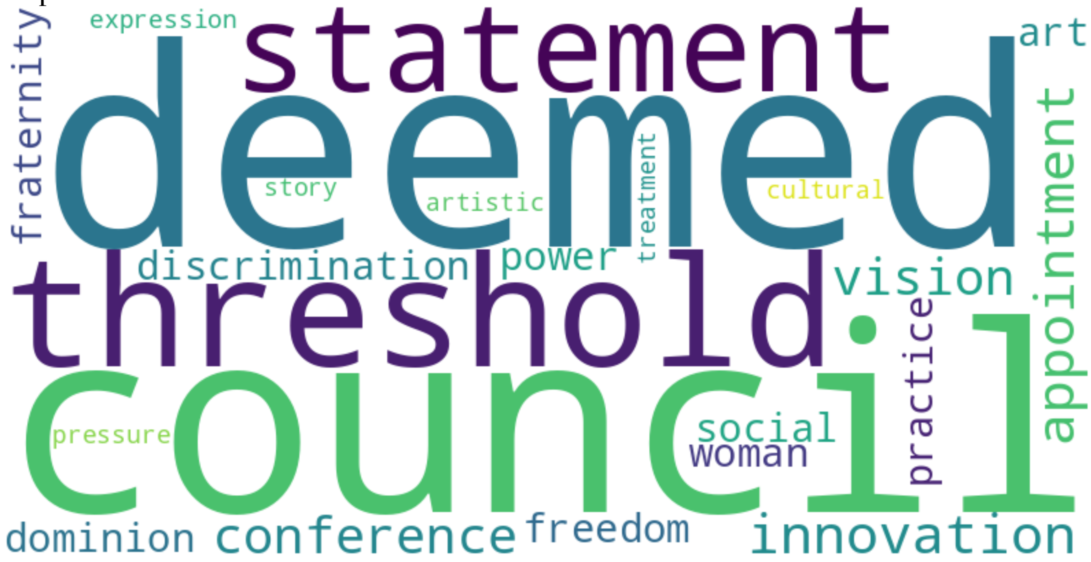
- **Description**: Arguments that revolve around economic principles, including market competition, trade, and consumer behavior. These appeals often highlight the benefits of free markets, investment, and capitalism, with a focus on economic growth and profitability.
- **Keywords**: competition, market, price, value, profit, trade, consumer, business, capitalism, commerce, transaction, demand, supply, monetary, investment, bargaining, enterprise, economic, marketing, sales, retail, exchange, commercial.

7. Industrial Arguments

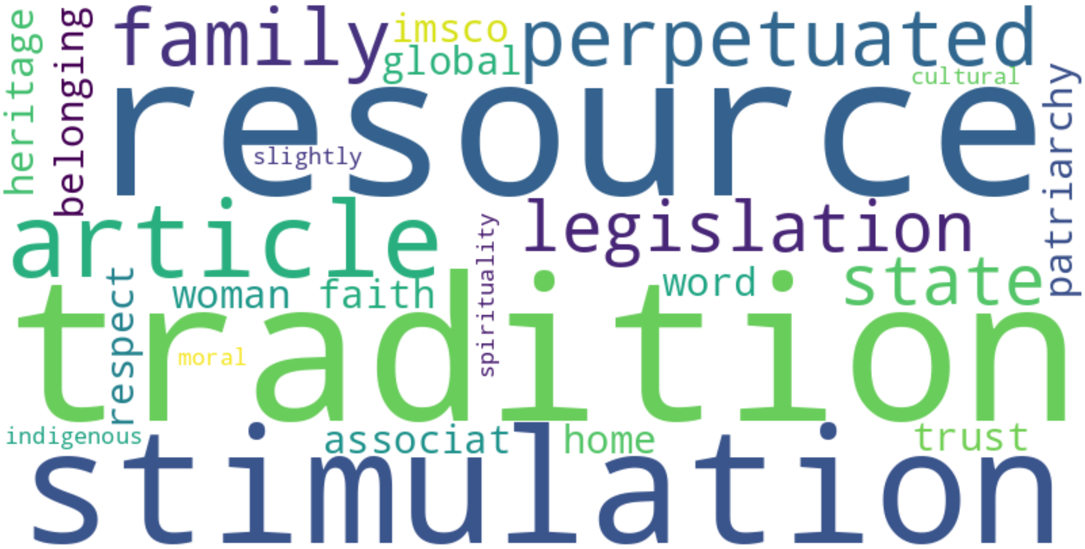
- **Description**: Arguments focusing on efficiency, productivity, and the advancement of technology. These appeals are grounded in the benefits of industrialization, automation, and technical expertise, often advocating for innovation and optimization in manufacturing and craftsmanship.
- **Keywords**: efficiency, productivity, technology, expertise, process, skill, method, automation, industry, standardization, precision, proficiency, workmanship, specialization, mechanism, optimization, craftsmanship, technical, output, machinery, manufacturing, fabrication, mechanization, industrialization.

Wordcloud: Here are lists of worldclouds with updated word for each cloud beyond the initial target terms I had set (while some of these original target terms were, as discussed earlier, dropped based on the lack of data)

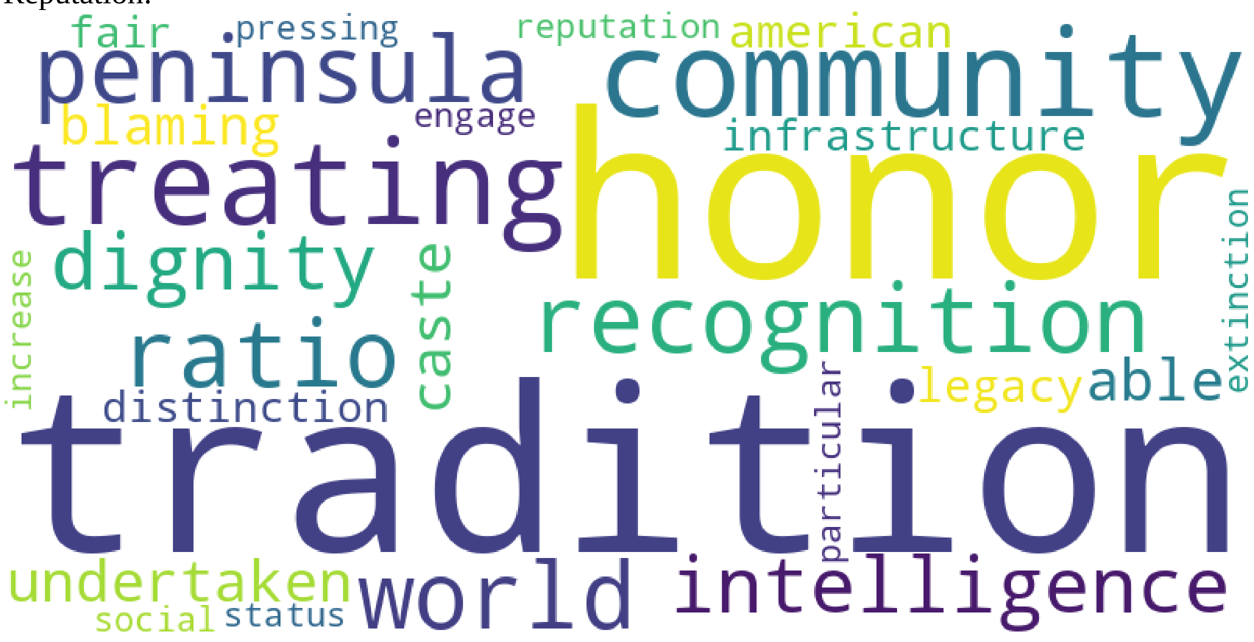
Inspiration:



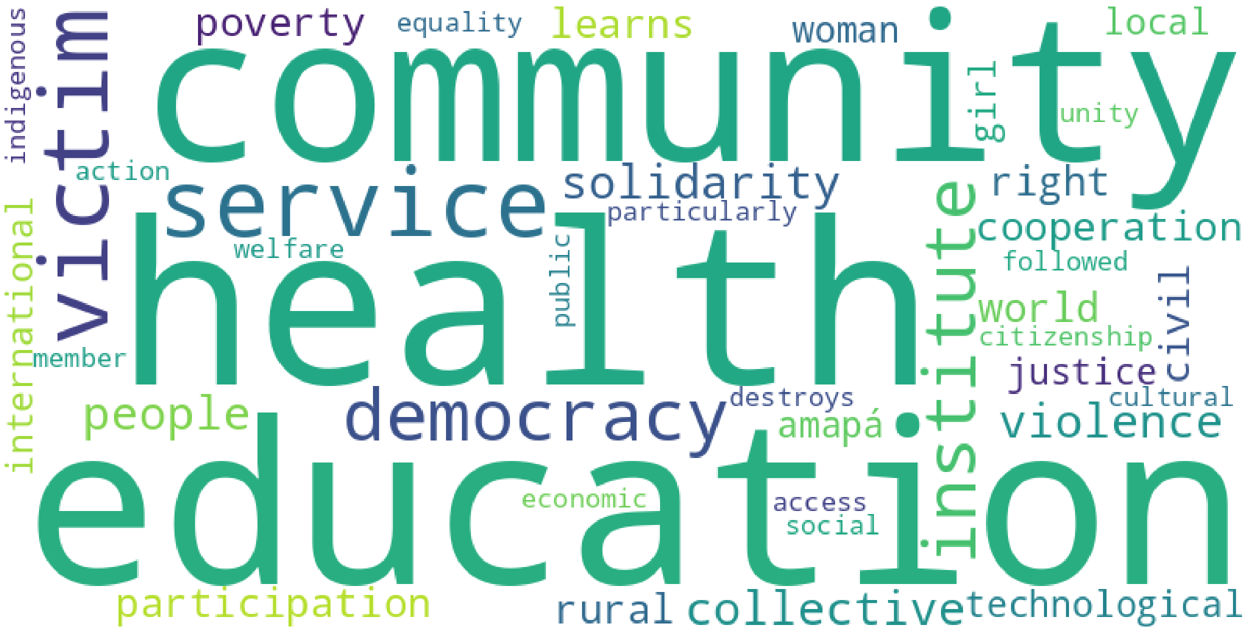
Domestic:



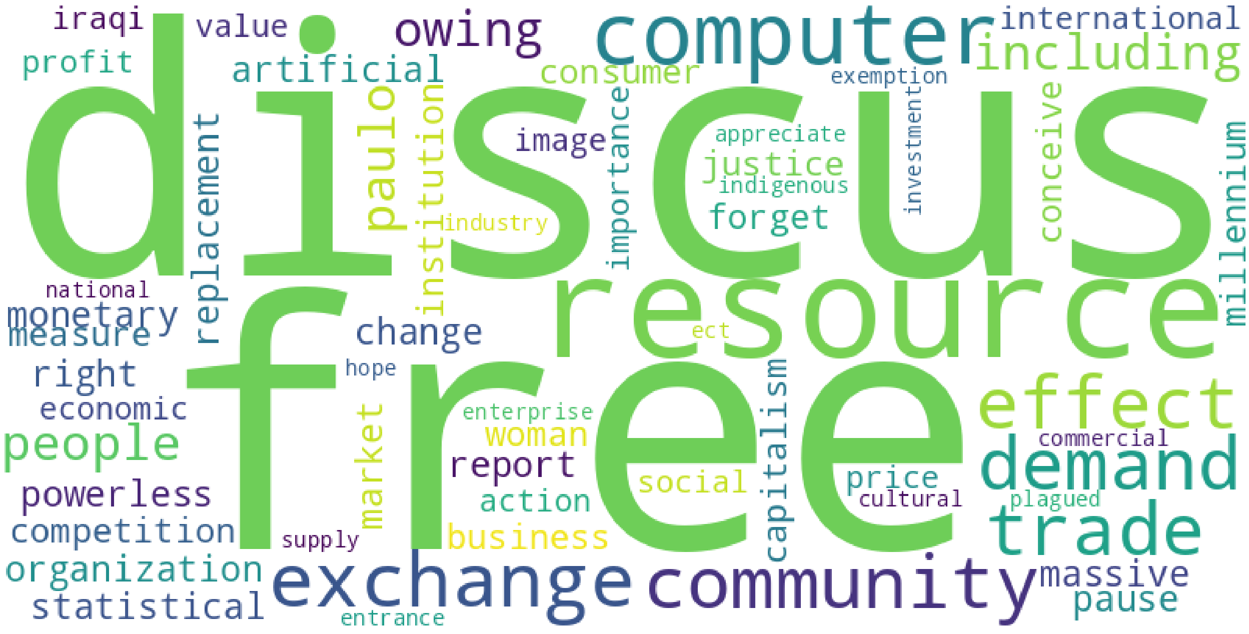
Reputation:



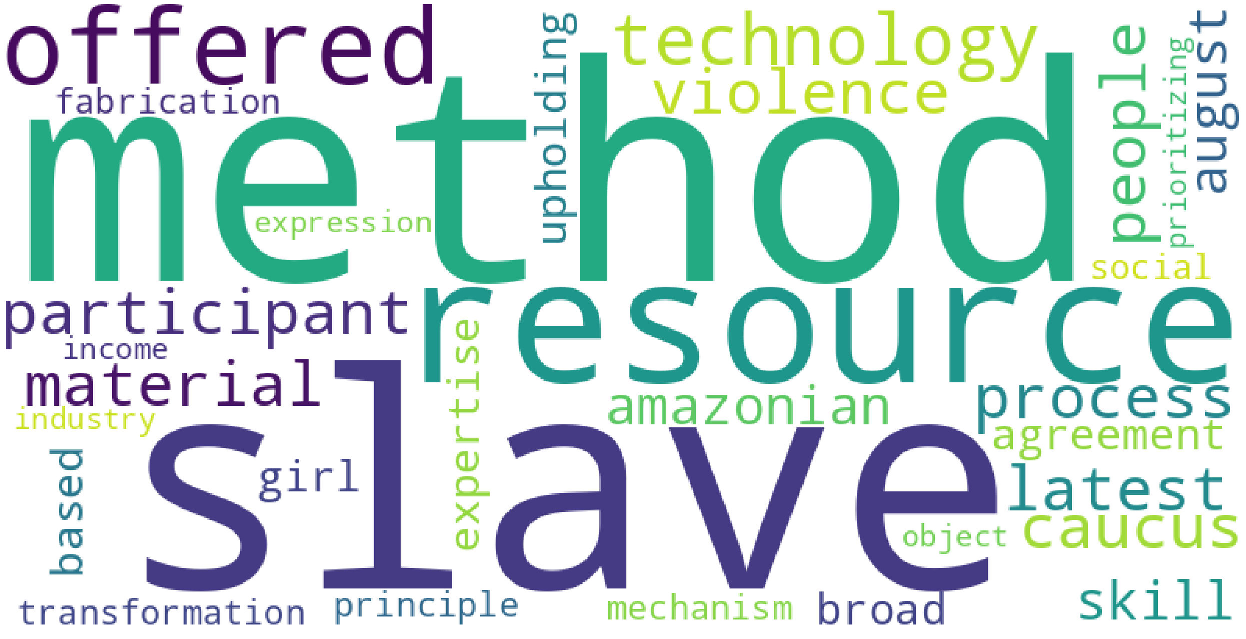
Civic:



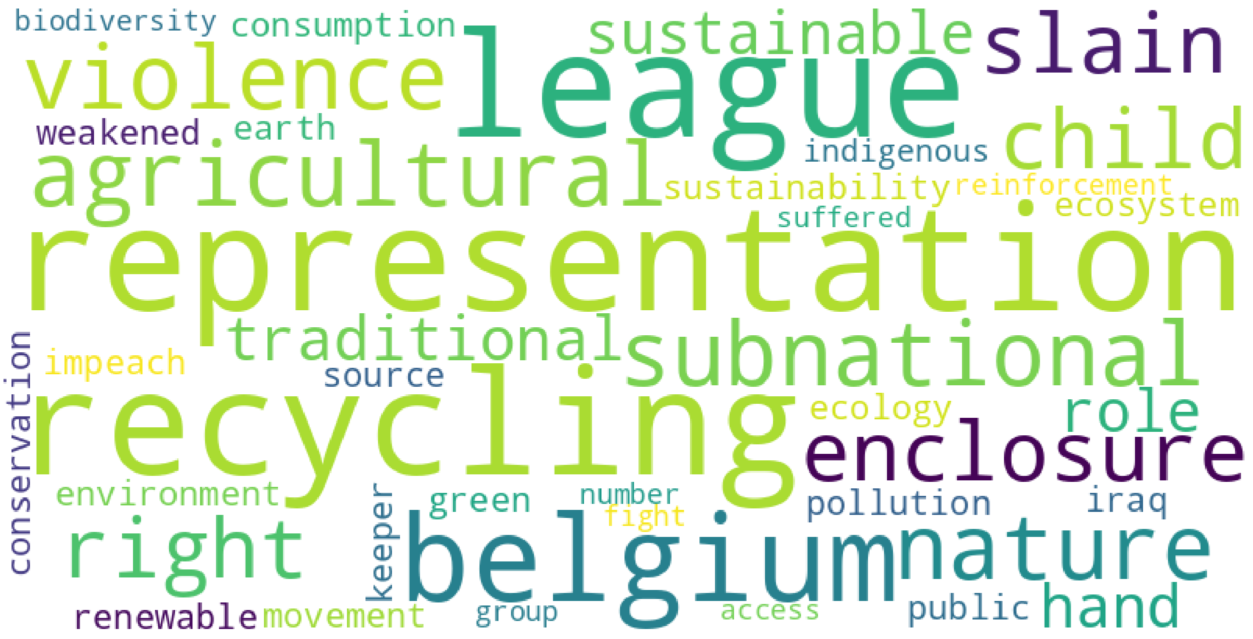
Market:



Industrial:



Green:



- **Region-Based Analysis**:
    - After categorizing the arguments, I analyzed them based on the geographical regions from which the activist groups originated. This allowed me to uncover trends such as:
        - **South Asian Groups**: Tended to rely heavily on domestic issues and reputation-based arguments.
        - **Latin American Groups**: Demonstrated a strong tendency to frame their arguments in terms of environmental sustainability, even when their primary focus was on other issues like social justice.
- **Text Analysis**:
    - I employed text analysis techniques, such as **term frequency-inverse document frequency (TF-IDF)**, to quantify the importance of terms and phrases in each region’s statements. This helped reveal dominant themes and recurring concepts in the text.
- **Correlation Discovery**:
    - By correlating the argumentation styles with the regions, I uncovered some clear trends:
        - South Asian groups linked local challenges to global governance.

- Latin American groups favored global themes like environmental protection, even in cases where environmental concerns were not the primary issue being discussed.

Results

Statistical Tests: ANOVA and Chi-Square Analysis

Chi-Square Test of Independence

To assess whether the distribution of argumentation styles varied significantly across regions, I conducted a **Chi-Square Test of Independence**. The objective was to test whether the choice of argumentation style was independent of the region from which the document originated.

Objective:

- To evaluate whether there is a significant association between the **region** (e.g., Africa, Asia, Latin America) and the **argumentation styles** (e.g., Civic, Domestic, Industrial) employed in the activist group statements at the UN.

Methodology:

- Each document was categorized by region, and the frequency of argumentation styles was recorded.
- The **Chi-Square test** compared the observed frequencies of argumentation styles across regions to the expected frequencies, assuming independence.
- The **degrees of freedom (df)** for the test were 18, based on the number of regions and categories of argumentation styles.

Results:

- **Chi-Square Statistic**: 0.09
- **P-Value**: 1.0
- **Degrees of Freedom**: 18
- Since the **p-value** is well above the threshold of **0.05**, the results indicate that the choice of argumentation style is **independent** of the region. This means there is no significant difference in the overall distribution of argumentation styles between regions.

Conclusion:

- The Chi-Square test provides strong evidence that activist groups from different regions employ **similar argumentation styles** when addressing the UN. The lack of significant association suggests that activist messaging is fairly consistent, regardless of regional differences.

ANOVA (Analysis of Variance)

In addition to the Chi-Square test, I conducted an **ANOVA** to compare the mean frequency of specific argumentation styles (e.g., Civic, Domestic, Green) across regions to identify any significant differences.

Objective:

- To determine if the frequency of specific argumentation styles (such as Green or Industrial modes) differs significantly between regions.

Methodology:

- For each argumentation style, a one-way **ANOVA** was performed, with the region being the independent variable and the frequency of the argumentation style being the dependent variable.
- The test checked for differences in the **mean frequency** of each style across regions.

Results:

- None of the argumentation styles showed significant differences in mean frequency across regions, with all **p-values** being greater than **0.05**.
- The **Reputation mode** showed a **slight exception** (F = 1.40, p = 0.28), which indicates that there might be some regional variation in how frequently reputation-based arguments are used, particularly in **Asia**. However, this result was not statistically significant.

Order	F-Statistic	P-Value	Conclusion
Inspirational	0.36	0.78	Not significant
Domestic	0.057	0.98	Not significant
Reputation	1.40	0.28	Slight regional difference in Asia
Civic	0.83	0.50	Not significant
Market	0.17	0.92	Not significant
Industrial	0.51	0.68	Not significant
Green	0.34	0.79	Not significant

Conclusion:

- The ANOVA results suggest that **none** of the argumentation styles differ significantly in frequency across regions. The slight exception observed for the **Reputation mode** in

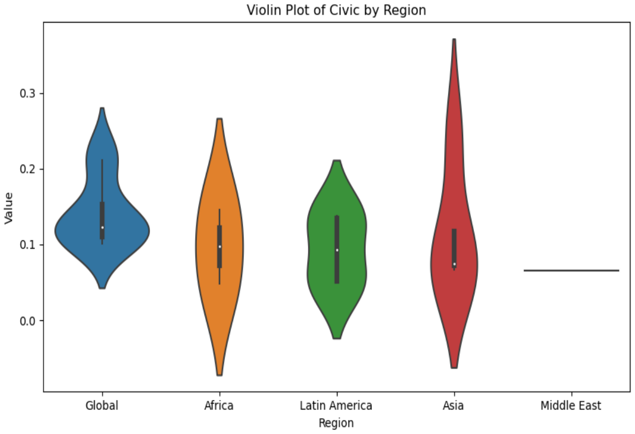
Asia was further explored using **Tukey's HSD Test**, confirming no statistically significant difference.

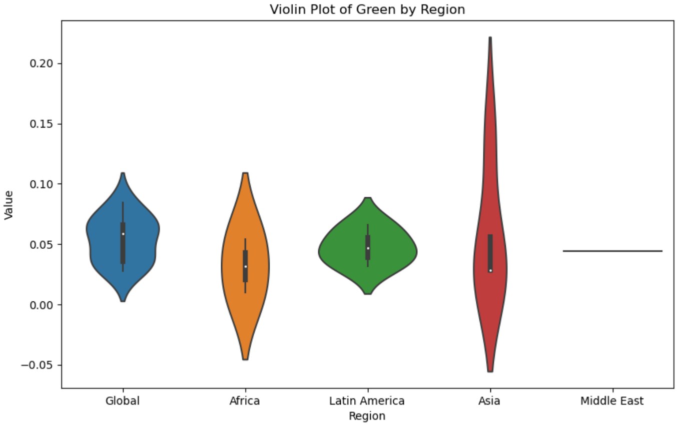
Violin Plots for Distribution Analysis

In addition to the ANOVA and Chi-Square tests, **violin plots** were used to visualize the **distribution** and **variance** of argumentation styles by region. These plots provide a more granular view of how the frequencies of certain argumentation styles (e.g., Civic and Green modes) vary across regions, allowing us to examine the **spread** and potential **skewness** in the data.

- **Civic Argumentation (Left Plot)**:
  - The **Global** and **Africa** regions have a relatively **broad distribution** of civic argumentation frequencies, indicating greater variance in how often civic arguments are employed.
  - In **Asia**, the distribution shows a **narrower spread**, suggesting that civic arguments are used more consistently across documents from this region.
  - The **Middle East** region shows only a single data point, which limits interpretability for this region.
- **Green Argumentation (Right Plot)**:
  - **Africa** and **Global** groups display a similar **spread** of frequencies, with a slight skew towards lower frequencies.
  - **Asia** shows the greatest **variance** in the use of green arguments, with the distribution being **heavily skewed**, indicating some documents focus heavily on green arguments while others barely touch on them.
  - **Latin America** has a relatively uniform but **narrow distribution**, showing that green arguments are employed somewhat consistently across documents, but not as frequently compared to other regions.

These visualizations complement the **ANOVA** results by illustrating how argumentation modes are distributed within regions, confirming that while the overall use is consistent across regions, certain modes (like Civic and Green) exhibit different patterns of variance and skew across regions.





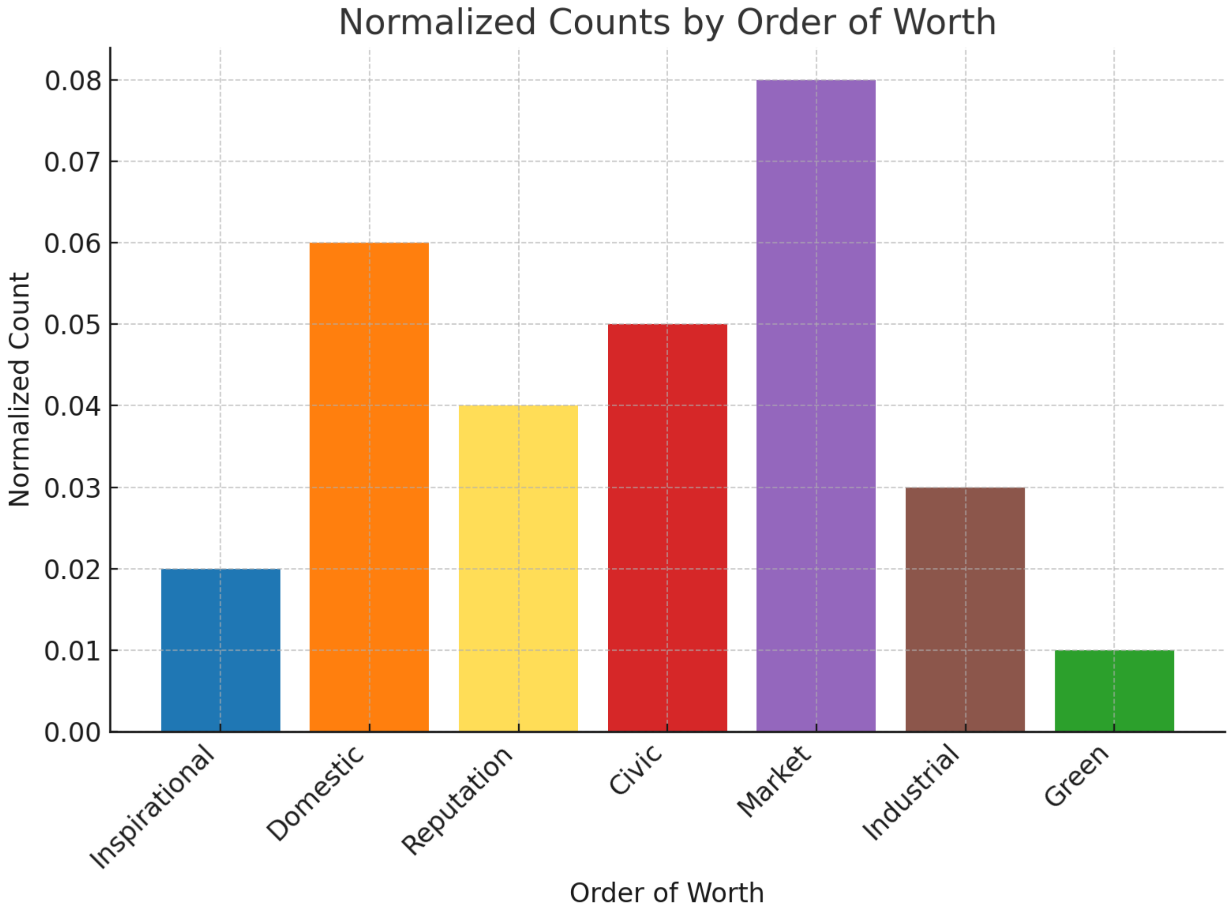
Conclusion from Statistical Tests:

Both the **Chi-Square Test** and **ANOVA** indicate that the distribution and frequency of argumentation styles are **relatively uniform across regions**. This reinforces the idea that activist groups, regardless of their geographical origin, tend to use similar strategies when addressing the UN. This finding suggests a level of **global convergence** in advocacy messaging, potentially influenced by the universal nature of issues like human rights and sustainability.

Normalized distribution for each order:

> This metric presents a bar chart that visualizes the normalized frequency of each "Order of Worth" category across the entire set of documents. Each category represents a distinct value system, such as environmental, civic, or market-based arguments. By normalizing the counts, we ensure that the differences in the length of the documents do not skew the results, allowing for an accurate comparison of how much each value system is emphasized across the dataset as a whole.
>
> This bar chart helps us understand which value systems are most frequently used by activist groups and how these systems are distributed throughout the collection of summary statements. It provides a clear overview of the dominant argumentative

frameworks employed in the activism at the UN.



Correlation/Covariance Metrics:

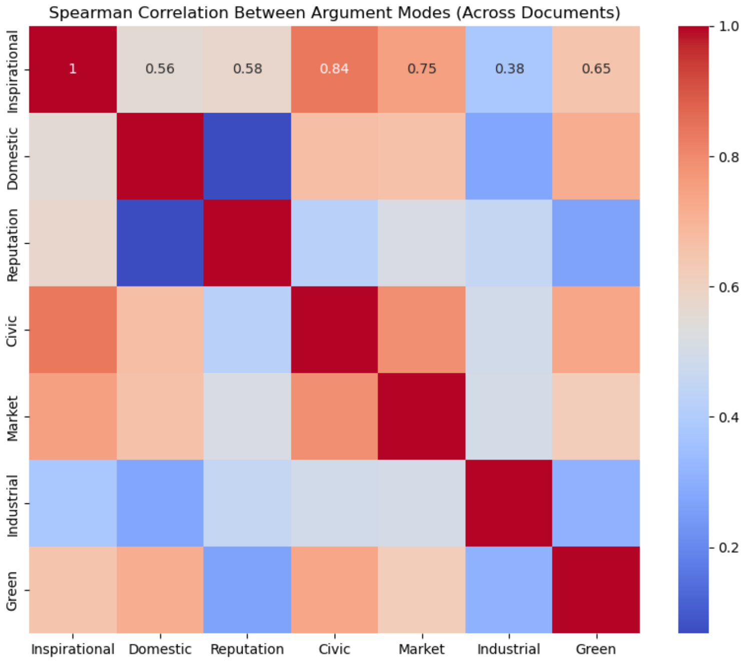
Spearman Correlation Analysis

The Spearman correlation analysis was employed to assess the relationships between different argumentation modes used by activist groups. This method was chosen due to its ability to capture **rank correlations** and **monotonic relationships** between argumentation styles, making it less sensitive to extreme values and outliers compared to other linear measures like Pearson's correlation.

Findings:

- **Inspirational** and **Civic** argument modes have a relatively **strong positive correlation (0.84)**, suggesting that documents emphasizing inspirational values often invoke civic participation.
- **Reputation** and **Domestic** argument modes exhibit a **negative correlation (-0.56)**, indicating that when one of these styles is used more frequently, the other tends to be used less.

- Other correlations of interest include a **moderate positive correlation** between **Inspirational** and **Market** argumentation modes, potentially reflecting a strategic overlap in framing market-based policies with aspirational rhetoric.



Interpretation:

These correlations suggest **underlying patterns** in how different argumentation styles are linked:

- The positive relationship between **Inspirational** and **Civic** modes suggests that activist groups invoke **civic duties** and **social participation** when making idealistic or visionary arguments.
- The **negative association** between **Reputation** and **Domestic** modes hints at differing priorities between these logics, potentially showing that groups prioritizing **global reputation** may downplay more domestically-rooted concerns.

Hierarchical Clustering / Dendrograms

Using **hierarchical clustering (Ward's method)**, the argumentation modes were grouped based on their similarities across documents. Ward's method minimizes the variance within clusters, creating **compact, homogeneous clusters**, making it easier to visualize how argumentation modes are grouped and "genetically" related across regions.

Global Analysis:

- **Civic** and **Market** modes cluster closely together, indicating a **strong similarity** in how frequently these modes are used across regions.

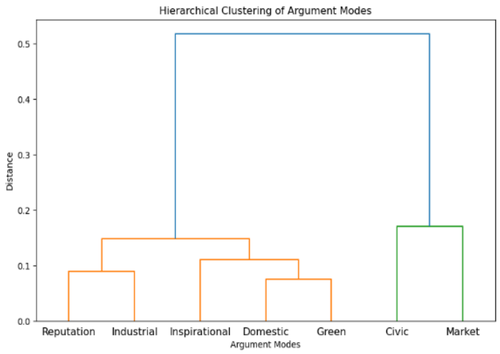
- **Reputation**, **Domestic**, and **Industrial** argument modes form another distinct cluster, suggesting that these modes tend to co-occur or be invoked in a similar way.

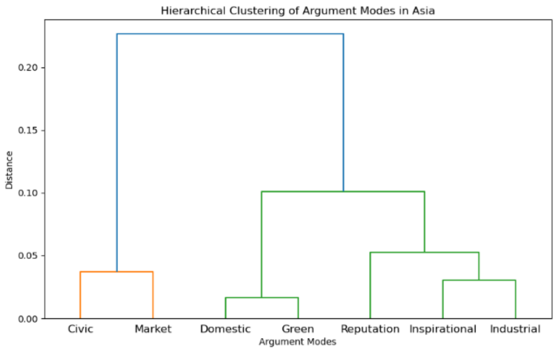
Regional Analysis (Example from Asia):

- In **Asia**, **Civic** and **Market** arguments remain closely linked, but **Green** and **Inspirational** modes form another unique cluster. This indicates that **environmental advocacy** and **visionary narratives** are used together more frequently in this region compared to others.

Interpretation:

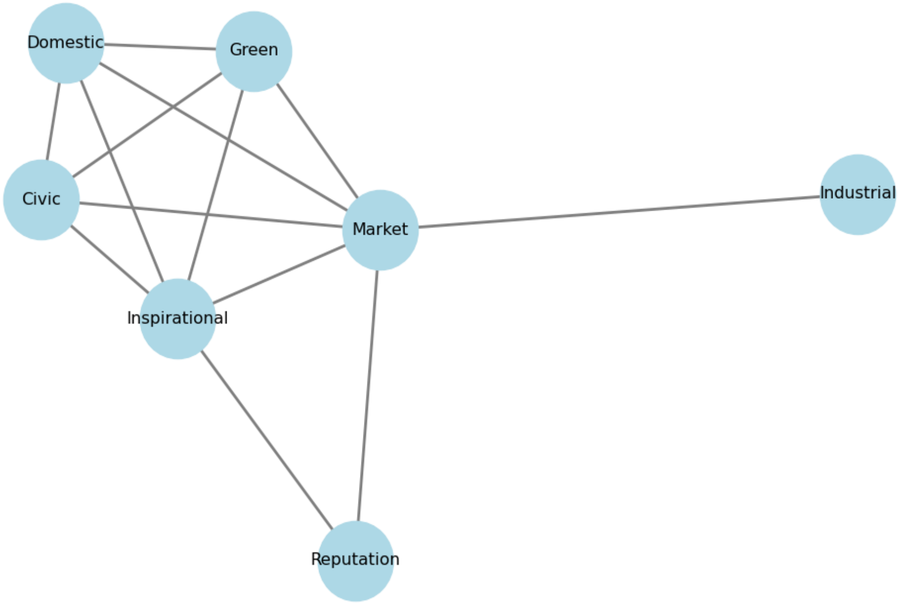
The hierarchical clustering provides a **clear structure** of argumentation relationships, revealing which modes are more commonly paired in activist statements. The similarity patterns vary slightly by region, with **Green** arguments having a stronger link to **Inspirational** rhetoric in Asia, while this relationship might be less pronounced in other regions.





Network Graph of Argument Modes

The **network graph** was constructed from the Spearman correlation matrix to further explore the relationships between argumentation modes. **Only strong correlations** (greater than **0.5** or less than **-0.5**) were used to create edges between nodes (representing argumentation modes).



Findings:

- **Civic**, **Domestic**, and **Green** are central nodes in the network, often linked to other modes like **Market** and **Inspirational**, reflecting their frequent co-occurrence in documents.
- The **Industrial** mode, however, appears **isolated**, indicating that it does not share strong correlations with other argument modes. This suggests that when **Industrial** arguments are employed, they stand out as a **more independent logic** rather than being intertwined with other styles of argumentation.

Interpretation:

The network graph highlights key argumentation **linkages**, with **eco-economic** and **eco-interpersonal** ties standing out. For example, the connections between **Green** and **Market**, and **Green** and **Civic**, suggest that activist groups often **interweave environmental concerns** with economic development and social engagement, reinforcing their narratives across different contexts.

Conclusion

These analyses—Spearman correlation, hierarchical clustering, and the network graph—paint a **comprehensive picture** of how argumentation modes relate to one another. The **Spearman correlation matrix** reveals specific **positive and negative relationships**, while the **dendrograms** illustrate which modes are most **genetically similar** in their usage patterns. Finally, the **network graph** highlights the **centrality of certain argumentation modes**, particularly **Civic** and **Green**, in forming the backbone of activist group discourse across the UN statements.

Additional Insights on Methodology:

1. **Spearman Correlation**: The **Spearman correlation** was selected over Pearson's correlation due to its ability to handle **non-linear relationships** and **outliers** more robustly. The **rank-based** nature of Spearman’s correlation made it more suitable for assessing how frequently argument modes appear in tandem, regardless of the exact magnitude.

2. **Ward’s Method in Hierarchical Clustering**: We chose **Ward’s method** for clustering because it focuses on minimizing **within-cluster variance**, which helps in identifying **coherent groups** of argumentation modes. This method is effective when the goal is to uncover **distinct but internally consistent clusters**.
3. **Threshold of 0.5 for Network Graph**: A **threshold of 0.5** was used in the network graph to capture only **strong correlations**, ensuring that the focus remained on **meaningful** relationships. This choice allowed for a more **interpretable graph**, highlighting the most **significant connections** between argumentation modes, while ignoring weaker ties.
4. **Spring Layout for Network Graph**: The **spring layout** algorithm was used to visualize the graph. This layout places nodes that are **more connected** (i.e., more highly correlated) closer together, intuitively representing the **strength of the relationships** between argument modes. The isolated position of **Industrial** argumentation in the graph provides a clear indication of its relative **independence** from other modes.

Corex Topic Modelling Analysis:

Methodology: Choosing Corex Over LDA

In this analysis, the **Corex (Correlated Topic Model)** was chosen over the more commonly used **Latent Dirichlet Allocation (LDA)** because Corex allows for a more flexible approach to topic discovery. While LDA rigidly assigns words to topics, Corex enables the use of **anchor words**, allowing the model to remain flexible while ensuring that certain topics remain grounded in key thematic areas. Anchor words help guide the topic discovery process, making Corex particularly suited for this analysis where argumentation modes need to be mapped to meaningful topics while still allowing room for fluidity in word assignments.

The document-word matrix generated through **CountVectorizer** helped convert the text data into a numerical format for topic modeling. The sparsity of this matrix (i.e., the percentage of non-zero elements) was relatively low, indicating that most words did not appear frequently across documents, a common trait in text data. Despite this sparsity, Corex was able to find meaningful topic structures by focusing on the relationships between the most informative words.

The model was set to identify **seven topics** (based on the number of predefined argumentation modes) to match the existing categories in the data. The flexibility of Corex allowed the model to adjust and refine the topic structure dynamically based on both anchor words and underlying patterns in the data, ensuring the topics captured relevant dimensions of the argumentation.

The results of the Corex model were visualized by grouping topics by region. This allowed for an exploration of how the importance of certain topics varied across geographic regions, providing deeper insights into the regional relevance of argumentation modes. This was particularly useful for seeing how argumentation differed between global and region-specific documents.

This methodology offered a balance between pre-set theoretical concepts and data-driven topic discovery, enabling a more nuanced understanding of argumentation modes within activist group discourse across regions.

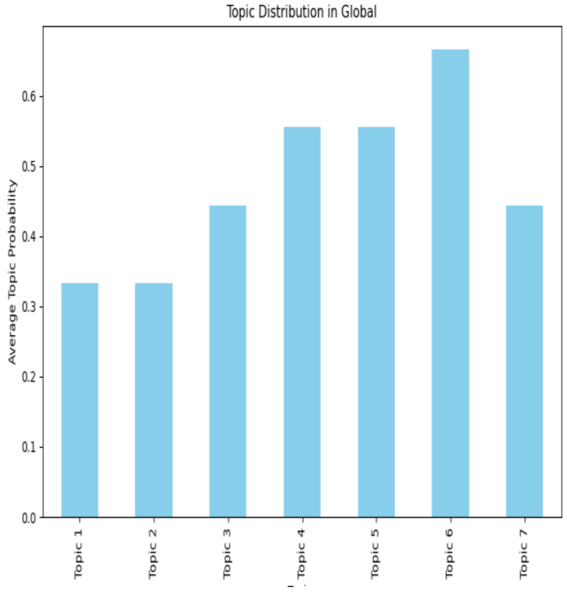
Identified Topics

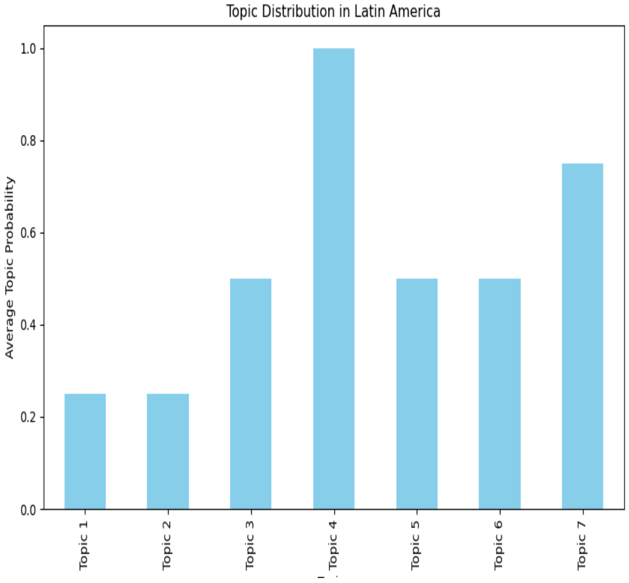
Seven distinct topics emerged from the model, each representing a cluster of frequently co-occurring terms across the activist statements. These topics highlight the central themes used by different regions:

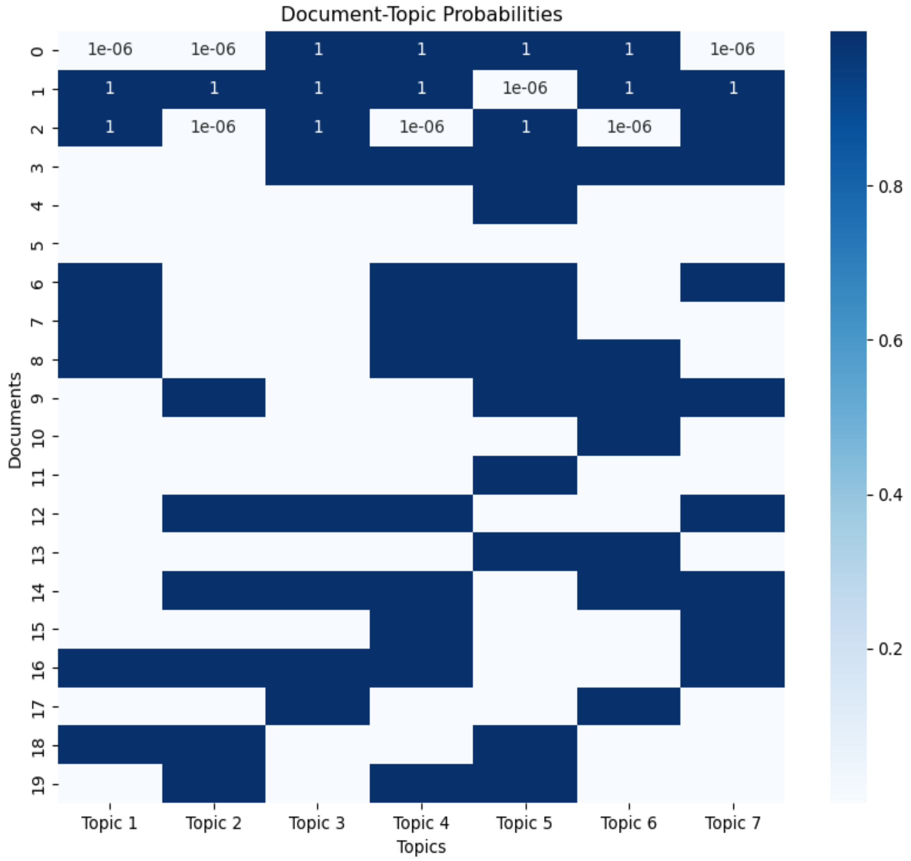
- **Topic 1** relates to principles of **justice, freedom, and accountability**, often invoking terms related to exclusion and recognition of rights.
- **Topic 2** touches on themes of **family, African sectors, and property**, suggesting a focus on issues relevant to social structures and regional contexts.
- **Topic 3** captures arguments centered on **employment, extreme life conditions, and prevention**, framing a discourse around economic hardship and protective measures.
- **Topic 4** emphasizes the **environment, progressive politics, and water**, indicating a strong link between environmentalism and political action.
- **Topic 5** addresses **indigenous rights, obligations, and development**, pointing to arguments around land and territorial issues.
- **Topic 6** revolves around **poverty, labor, and official relationships**, underscoring the ongoing struggles in global poverty alleviation.
- **Topic 7** focuses on **support, diversity, and regional gaps**, showcasing the global concern for disparities and representation.

Regional Distributions

The bar charts on Slide 2 illustrate how these topics are distributed across different regions. For example, **Latin America** demonstrates a strong association with **Topic 4** (environment and politics), highlighting the region's focus on ecological and social justice issues. Meanwhile, the **Global** region has a more even topic distribution, indicating a broader engagement with multiple concerns. The **heatmap** further visualizes how specific documents align with different topics, showing that while some documents strongly favor one or two topics, others span across multiple themes.







Interpretation

The **Corex model** provides a nuanced understanding of the argumentation patterns used by activist groups in the UN. The flexible nature of the model, combined with anchor words, allowed for the discovery of core themes that reflect both broad and region-specific concerns. The strong presence of topics around **justice, environmentalism, and social inequality** suggests that these are dominant issues across the board, while regionally specific topics like **land rights** and **family issues** highlight the particular discourses relevant to certain areas.

Overall, the use of Corex provided a rich, structured insight into the thematic landscape of the activist statements, offering a clearer view of how argumentation modes align with key topics and regions.

Machine Learning: Clustering and Supervised Models

Clustering Models

1. **K-Means Clustering**:

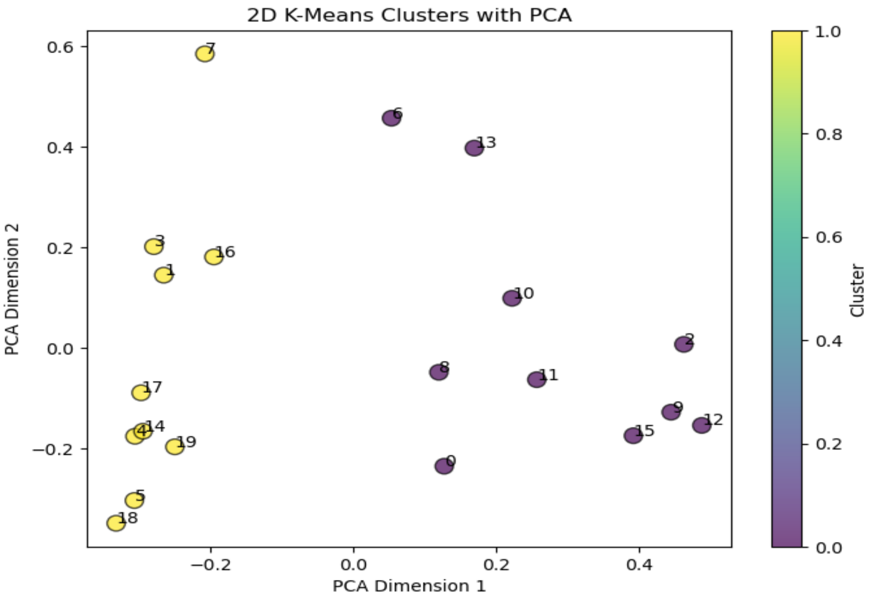
  - **K-Means** was used to cluster the PCA-reduced TF-IDF matrix. By standardizing the data with **StandardScaler** and applying **Principal Component Analysis (PCA)** to reduce dimensionality, the optimal number of clusters was determined using silhouette scores.
  - **Evaluation**: Various cluster configurations were tested (from 2 to 10 clusters), with the best performance found using **2 clusters** after reducing the data to **2 dimensions** using PCA. This setup produced the highest silhouette score.
  - **Key Findings**: The clusters primarily focused on themes related to **indigenous issues and justice** (e.g., land rights, governance) and **socio-economic challenges** (employment, poverty).

2. **Agglomerative Clustering**:
    - **Hierarchical/Agglomerative Clustering** was explored as an alternative, which forms clusters based on the recursive merging of smaller clusters.
    - **Silhouette Scores**: When compared to K-means, the silhouette scores for agglomerative clustering were slightly lower, which suggested less cohesion between the clusters. Despite testing different numbers of clusters, K-means provided better-defined clusters.
3. **Gaussian Mixture Model (GMM)**:
    - The **Gaussian Mixture Model (GMM)** was also applied, assuming that data points are generated from a mixture of several Gaussian distributions.
    - **Silhouette Scores**: GMM’s performance was lower compared to K-means, with less distinct cluster boundaries and a tendency to assign overlapping probabilities for certain data points. As a result, K-means was preferred due to its clearer separation of clusters.

Dimensionality Reduction and Clustering

- **Principal Component Analysis (PCA)** was applied to reduce the dimensionality of the TF-IDF matrix to 2 or 3 components. This dimensionality reduction step helped visualize clusters and boosted the performance of the models by eliminating noise and irrelevant features.
- **Evaluation**: The silhouette scores indicated that **2 clusters** provided the most stable grouping for this dataset when combined with 2D PCA.

By leveraging the above techniques, the analysis showed clear separation of the clusters and highlighted distinct thematic focuses within the documents, including topics around **indigenous issues**, **justice and freedom**, and **socio-economic challenges**.



Supervised Learning Models

Once the cluster labels were obtained, we tested several **supervised learning models** to predict cluster assignments. After performing a **train-test split** with 60% of samples as our training set vs 40% as our test set to reduce overfitting, we used a **Stratified 3-Fold Cross-Validation** strategy to evaluate model performance, ensuring that the class distributions were preserved across the folds. The models included:

1. **Logistic Regression**:
   - A simple yet effective classification model used to predict cluster membership based on linear relationships.
   - **Hyperparameters tuned**: The regularization parameter (C) and penalty type (l2).
   - **Best Parameters Found**: C=10, penalty='l2'.
2. **Random Forest**:
   - An ensemble model that builds multiple decision trees and averages their predictions, robust to overfitting due to its bagging approach.
   - **Hyperparameters tuned**: The number of trees (n_estimators), maximum depth (max_depth), and minimum samples required to split a node (min_samples_split).
   - **Best Parameters Found**: n_estimators=50, max_depth=None, min_samples_split=2.
3. **Support Vector Machines (SVM)**:
   - A powerful algorithm that aims to maximize the margin between data points, using kernel functions to capture non-linear patterns.
   - **Hyperparameters tuned**: Regularization parameter (C), kernel type (linear or rbf), and kernel coefficient (gamma).

    - **Best Parameters Found**: C=1, kernel='linear', gamma='scale'.

4. **K-Nearest Neighbors (KNN)**:
    - A non-parametric algorithm that classifies a point based on the majority label of its nearest neighbors.
    - **Hyperparameters tuned**: The number of neighbors (n_neighbors), distance metric (p), and weighting scheme for neighbors (uniform or distance).
    - **Best Parameters Found**: n_neighbors=3, p=1, weights='uniform'.

Evaluation Metrics

For each model, the **accuracy, precision, recall**, and **F1-score** were computed. Cross-validation was performed using **StratifiedKFold**, and **GridSearchCV** was employed to fine-tune hyperparameters for optimal model performance.

- **Random Forest** emerged as the best model with a 100% accuracy during cross-validation and 87.5% accuracy on the test set, due to its strong generalization ability, performance stability across different configurations, and capacity to handle small datasets well without overfitting
- **Logistic Regression** also performed well, offering excellent accuracy and interpretability with simpler tuning and computational efficiency.
- **SVM** and **KNN** models, while showing good performance, exhibited more sensitivity to the number of clusters and dimensions, making them more prone to overfitting.

Output for the chosen model, Random Forest:

```
Random Forest Stratified 3-Fold CV Accuracy: 1.0000

Random Forest Test Accuracy: 0.8750

Classification Report for Random Forest:

          precision   recall  f1-score   support

       0       1.00      0.75      0.86         4

       1       0.80      1.00      0.89         4

  accuracy                         0.88         8

 macro avg       0.90      0.88      0.87         8

weighted avg       0.90      0.88      0.87         8
```

Recommendations and Future Work

Recommendations:

Argumentation Styles: The analysis revealed that there were intriguing links made between ecological and interpersonal violence amongst many of the activist groups. Activist groups from South Asia frequently used arguments related to their reputation issues. This often involved linking national issues with global frameworks, especially in the context of human rights. Interestingly, the domestic governance trend was not favored by these groups while the reputation trend was for these issues, indicating that caste issues are by activists are conceived of more in terms of the rank and to distinction than to kinship or genealogy.

On the other hand, many Latin American groups, but even an Iraqi one, conceived of linkages between ecological and interpersonal violence, with ecological issues commonly mediating between economic and domestic issues, as per our network graphs. From our results, we also found that there was a distinct environment + land rights + indigeneity grouping, both in our CorEx and in our clustering models for machine learning, indicating that this mode of argumentation and style of document was particularly distinct compared to many others. It would be interesting to view a more systematic linkage of these connections, but the small size of the data set makes any conclusions tentative at this point.

Based on the analysis of the activist groups' argumentation strategies, I recommend the following approaches for groups seeking to improve their advocacy at the UN:

1. **Employ Multimodal Argumentation**: Activist groups should consider adopting multiple strategies in their argumentation. For example, blending local issues with global concerns could result in more persuasive advocacy. Groups that framed their arguments from both a regional and global perspective were more successful in making their case.
2. **Avoid Stereotypical Framing**: Some groups tend to rely on stereotypical argumentation patterns associated with their region. Instead, groups should experiment with less traditional or expected argumentation frameworks, such as focusing on environmental sustainability even if their primary focus is human rights.
3. **Research Successful Strategies**: Groups that diversified their argumentation styles tended to achieve better advocacy results. Analyzing and learning from the argumentation strategies of more successful groups could offer insights into effective presentation and advocacy.

Future Work:

This project provided valuable insights into the argumentation strategies of activist groups presenting at the UN. However, there are several areas where further research could be beneficial:

- **Temporal Analysis**: Future research could explore how argumentation strategies evolve over time in response to changing global priorities or issues, and assess how major events potentially affect and lead to shifts in argumentation rhetorics.
- **Success Metrics**: A deeper dive into the outcomes of each activist group's advocacy could help to better understand the link between argumentation strategy and policy impact. If more places were selected than finer tuned comparisons would be more statistically valuable.
- **Global Comparison**: Expanding the dataset to include activist groups from more regions would allow for a broader comparison of argumentation strategies across the globe, and would allow more inter-regional based analyses (which I did not attempt here due to the small sample size).