Final Project Report - CSCI S-89B, Spring 2024

Analysis of Political Speeches Using NLP:

Case Study of the United Nations Biodiversity Conference

TO: Professor Kurochkin and TAs

FROM: Jaqueline Garcia-Yi

DUE: December 16, 2024

I. Background and Problem Statement

Biodiversity is the foundation of life on Earth. It supports not only ecosystems, but also the global economy, climate change resilience, and the food systems that sustain humanity. Currently, we are facing an unprecedented biodiversity crisis. Habitats are disappearing, and as many as one million species are close to extinction (IPBES, 2019). Species extinction is occurring at a rate 100 to 1,000 times higher than the natural rate (*idem*).

Among others, **biodiversity strengthens food security by supporting pollination.** Around 75 percent of crops depend on animal pollination, and losing biodiversity in pollinators would threaten the availability of fruits, vegetables, and nuts, affecting the nutrition of billions of people (FAO, 2018). Furthermore, fisheries support more than 10 percent of the world's population. Declining fish populations because of ecosystem and biodiversity loss directly threaten food and income sources for millions (FAO, 2020).

In addition, **biodiverse ecosystems help regulate diseases.** For instance, forests, which house around 80% of the world's terrestrial biodiversity, are vital for controlling zoonotic diseases. Deforestation increases the risk of diseases like malaria and COVID-19 (UNEP, 2020). Conserving biodiversity also supports pharmacology: 50% of drugs, including cancer treatments, come from or are derived from natural sources (WHO, 2019).

This pressing reality underlines the need for decisive action for biodiversity conservation. **The representatives of 196 countries gathered recently at the 2024 United Nations Biodiversity Conference (October 21 to November 1, 2024) in Cali, Colombia.** This United Nations Conference included the COP16 (the 16th Conference of the Parties to the Convention on Biological Diversity), CP-MOP11 (the 11th Meeting of the Parties to the Cartagena Protocol on Biosafety), and NP-MOP5 (the 5th Meeting of the Parties to the Nagoya Protocol on Access and Benefit-Sharing).

Conferences like this generate very large vast amounts of speech, as well as decision and declaration text data, which would be impractical to analyze manually. NLP offers the possibility to automatically process and analyze large volumes of text in a fraction of the time that it would take for any human researcher.

Specifically, NLP can help to extract key insights; highlight important global and regional trends on biodiversity conservation; compare the position of different geographical regions on biodiversity issues; and detect potential emerging patterns on topics related to biodiversity. By leveraging NLP, Governments, NGOs, and other interested stakeholders could

quickly obtain information about the current global and local discourse and trends surrounding biodiversity to help them shape future strategies for conservation and sustainability.

II. Data Availability and Data Sources

2.1. Current discourse of different geographic regions around the globe on biodiversity (2024)

For analyzing the current discourse of different geographic regions around the globe, the data consisted of **all the 315 official country (written) statements** from the meetings that occurred during the last UN Conference on Biodiversity (COP16), which took place on October/November 2024 in Cali, Colombia (available at https://www.cbd.int/conferences/2024/cop-16/documents):

- 3 written statements: COP-16 / CP-MOP-11 / NP-MOP-05 - Plenary, Friday, 1 November 2024 - 22:00 America/Bogota
- 1 written statement: COP-16 / CP-MOP-11 / NP-MOP-05 - Working Group I, Friday, 1 November 2024 - 14:30 America/Bogota
- 6 written statements: CP-MOP-11 - Plenary, Wednesday, 30 October 2024 - 20:00 America/Bogota
- 19 written statements: COP-16 - High Level Segment, Wednesday, 30 October 2024 - 15:30 America/Bogota
- 1 written statement: COP-16 / CP-MOP-11 / NP-MOP-05 - Working Group I, Wednesday, 30 October 2024 - 10:00 America/Bogota
- 30 written statements: COP-16 - High Level Segment, Wednesday, 30 October 2024 - 10:00 America/Bogota
- 45 written statements: COP-16 - High Level Segment, Tuesday, 29 October 2024 - 14:30 America/Bogota
- 6 written statements: COP-16 - High Level Segment, Tuesday, 29 October 2024 - 10:30 America/Bogota
- 2 written statements: COP-16 / CP-MOP-11 / NP-MOP-05 - Plenary, Friday, 25 October 2024 - 19:30 America/Bogota
- 18 written statements: COP-16 / CP-MOP-11 / NP-MOP-05 - Working Group I, Friday, 25 October 2024 - 10:00 America/Bogota
- 12 written statements: COP-16 / CP-MOP-11 / NP-MOP-05 - Working Group II, Wednesday, 23 October 2024 - 10:00 America/Bogota
- 39 written statements: CP-MOP-11 / NP-MOP-05 - Working Group II, Tuesday, 22 October 2024 - 15:00 America/Bogota
- 54 written statements: COP-16 / CP-MOP-11 / NP-MOP-05 - Working Group I, Tuesday, 22 October 2024 - 10:00 America/Bogota
- 23 written statements: COP-16 / NP-MOP-05 - Working Group I, Monday, 21 October 2024 - 15:00 America/Bogota
- 21 written statements: COP-16 / NP-MOP-05 - Working Group II, Monday, 21 October 2024 - 15:00 America/Bogota
- 35 written statements: COP-16 / CP-MOP-11 / NP-MOP-05 - Plenary, Monday, 21 October 2024 - 09:00 America/Bogota

2.2. Data for analyzing the global trends over time on biodiversity conservation (1994-2024)

For analyzing the global trends over time on biodiversity conservation, the data was obtained from the **all the 498 Decision Reports of the United Nations Biodiversity Conferences (from the first one in 1994 to the latest one in 2024)**, which were available at:

- 13 Decision reports from the 1994 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-01
- 23 Decision reports from the 1995 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-02
- 27 Decision reports from the 1996 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-03
- 19 Decision reports from the 1998 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-04
- 29 Decision reports from the 2000 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-05
- 32 Decision reports from the 2002 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-06
- 36 Decision reports from the 2004 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-07
- 34 Decision reports from the 2006 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-08
- 36 Decision reports from the 2008 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-09
- 47 Decision reports from the 2010 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-10
- 33 Decision reports from the 2012 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-11
- 35 Decision reports from the 2014 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-12
- 34 Decision reports from the 2016 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-13
- 38 Decision reports from the 2018 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-14
- 35 Decision reports from the 2022 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-15
- 27 Decision reports from the 2024 United Nations Biodiversity Conference, available at https://www.cbd.int/decisions/cop?m=cop-16

TOTAL **498 decision reports**

There were two extraordinary meetings, one in 1999 and one in 2019, but they covered mostly administrative matters, and therefore, they were excluded from the analysis. In addition, in 2020, there was not UN Biodiversity Conference due to COVID-19.

III. Data Pre- Processing

3. 1 Data loading and cleaning

3.1.1. Data for analyzing the current discourse of different geographic regions around the globe on biodiversity (2024)

The code for loading and cleaning the data is available in the jupyter notebook named *FinalProject-DataCleaningRegion*. All the country statements were downloaded manually from the web-page indicated above in 3.1. In total, 9 of the 315 statements were damaged or not available. Therefore, the final number of statements used in this project was 306. Those country statements were classified by geographic region, as follows:

-Africa	70 country statements
-Arab States	13 country statements
-Asia	38 country statements
-Australia and Oceania	9 country statements
-The Caribbean and Small Islands	19 country statements
-Eastern Europe	5 country statements
-Latin America	48 country statements
-Middle East	9 country statements
-Organizations	45 statements
-USA and Canada	5 country statements
-Western Europe	45 country statements
TOTAL	**306 country statements**

When a country's official language was one of the six official UN languages (Arabic, Chinese, English, French, Russian, or Spanish), its statement was delivered in that language. To analyze these statements, I had to translate them into English. I used Google Translate. Additionally, since some statements were provided in PDF format, I first converted the PDF documents into DOCX format. This process may have affected the semantic coherence of the model's results, as discussed in Section V of this report.

The classified and translated statements were organized into folders named according to their corresponding geographical regions and then batch-loaded into the Jupyter Notebook. Statements from the same geographical region were consolidated into a single text file for each region. Subsequently, each text file was processed by converting all text to lowercase and removing the following elements:

- numeral markers

- numbers

- punctuation and special characters

- non-ASCII characters

- country names

- multiple blank spaces

- words with length smaller than 2 and stop words

The process of converting the PDF documents into Word format and then translating them resulted in uneven paragraph structures, with many paragraphs differing significantly from their original lengths. To address this, I divided the text into chunks of 50 words, ensuring uniform paragraph lengths across each text file. However, this process also impacted the semantic coherence of the model's results, as detailed in Section V of this report.

Finally, the regional text files were combined into a single consolidated text file, structured as a dataframe with one column containing the text and the corresponding geographic region listed in the adjacent column as metadata.

3.1.2. Data for analyzing the global trends over time on biodiversity conservation (1994-2024)

The code for loading and cleaning the data is available in the Jupyter Notebook named *FinalProject-DataCleaningTrends*. A total of 368 out of 498 decisions, available in Word format, were manually downloaded. These decisions correspond to the years 1994 and 2004–2024 (see webpages referenced in Section 3.2) and were organized into folders by year. The remaining 130 decisions, available in HTML format, were downloaded by the code directly.

The code loads all the files either in Word of HTML format into the Jupyter Notebook. Documents from 2018 to 2024 were in DOCX format, while those from 1994 to 2016 used the older DOC format. To handle this, a custom function was implemented to recognize file extensions before loading the data.

After loading, the documents from each year were merged into a single text file per year. Finally, these yearly text files were combined into a consolidated text file, structured as a dataframe. The dataframe includes one column containing the text and an adjacent column listing the corresponding year as metadata.

IV. Data Analysis and Modeling

I conducted two separate data analysis and modeling for: (a) the current discourse of different geographic regions around the globe on biodiversity (2024), and (b) the global trends over time on biodiversity conservation (1994-2024).

The analyses were conducted in R's STM, including the following commands[1]:

- **textProcessor()** for pre-processing the raw textual data for STM, including tokenization; lowercasing; stemming; and removing stop words, numbers, and punctuation.

[1] The links of the R Documentation per command are included in the Reference section of this report.

In addition, common words such as biodiversity, conservation, framework, global, convention, cop, cbd, and statement were added as "customstopwords". They were not in the default stopword list but were overly frequent in the specific dataset used in this project. The main objective of adding these specific words as stop words was to improve sematic coherence of the STM model.

- **preDocuments()**, which processes the output of textProcessor() and create inputs for the STM algorithm. It organizes the tokenized and cleaned text into three structured components (Documents, Vocabulary, Metadata) for the STM model.

  I included lower.thresh = 5 and upper.thresh = 0.8 x length(documents) to remove words that appear in fewer than 5 documents and in more than 80 percent of all documents, respectively. The main objective of using these lower and upper thresholds was to improve semantic coherence of the STM model.

  In addition, in this final project, the init.type used was "Spectral" that would help the subsequent STM models to converge more quickly during the iterative steps of the expectation maximization (CRAN R Project, 2023).

- **searchK()** for selecting the optimal number of (K) topics for the STM model, which is critical for producing coherent and interpretable results. It evaluates the performance of STM models across a range of candidate values of K.

  For this final project, the Ks evaluated were in the range between 2 and 10. A preliminary evaluation also considered a broader range (for instance, up to 20 or 50), these higher values resulted in low semantic coherence, making them unsuitable.

- **ggplot()** for plotting the coherence versus exclusivity scores obtained from searchK(). The optimal K was selected based on this visualization and on the maximum composite score obtained from the following formula: Composite Score = 0.5 x Sematic Coherence + 0.5 * Exclusivity Score.

- **stm()** for running the STM model with the optimal K (selected as indicated above). The prevalence used for the stm model of the current 2024's discourse of different geographic regions on biodiversity was Region, while for the stm model of the global trends over time from 1994 to 2024 on biodiversity conservation was year.

  In addition to using *int.type = "Spectral"*, as during the *preDocument()* phase, the *max.em.its* was set to 500. This means that the algorithm used a maximum of 500 iterations, which helps to control the trade-off between computation power and model accuracy. That number seems to be suitable for most medium-sized datasets of fewer than 10,000 documents (CRAN R Project, 2023).

  The *gamma.prior* of the discourse of different geographic regions was set to *"L1"*, which works well with categorical prevalence variables such as region; and of the global trends over time to *"Pooled"*, which is more suitable for continuous prevalence variables such as years. L1 refers to Lasso regulation (L1 regulation), which avoids overfitting by reducing complexity. While "Pooled" assumed a shared distribution of topic proportions across

documents, allowing for smooth variations across the continuous variable (year), and producing topics that are more general among documents (CRAN R Project, 2023).

- **exclusivity()** and **semanticCoherence()** for obtaining the exclusivity and semantic coherence of the fitted models, which provide information about the quality and interpretability of the topic generated by the STM model.

  Exclusivity refers to how distinct (or exclusive) a topic is relative to other topics. A higher exclusivity value means that the words in a topic are unique to that topic, while lower exclusivity means that the words are more shared across topics.

  While semantic coherence measures how semantically related the top words of a topic are to one another. Coherent topics are those where the words within the topic make sense together in a meaningful way (for example, words that often appear together in real-world texts). A higher semantic coherence score indicates that the words in a topic tend to co-occur frequently, suggesting that the topic is meaningful and interpretable. Low coherence suggests that the words in the topic are loosely related and may not form a meaningful cluster.

- **summary()** for providing a comprehensive overview of the topics generated by the STM model. It outputs key metrics such as: (a) *"highest prob"* or the words that have the highest probability of being associated with each topic; (b) *"top words"* or words that are most frequent within a topic; (c) *"FREX"* or frequency-weighted exclusivity of words, which combines both frequency and exclusivity of words in the topic (they occur often within a topic but not much across other topics); (d) *"Lift"* which helps in identifying words that are significantly more likely to appear in one topic versus others, which can help in distinguishing topics that are similar but with slight differences, and (e) "Score" which indicates words that are central to the content of each topic.

- **cloud()** for cloud words per topic. It visually represents the most frequent and significant words associated with each topic. It provides a quick, intuitive way to see which words dominate each topic.

- **estimateEffect()** for evaluating the effects of year or region/country on the topics generated by the STM models. The function helps to understand the external factors (such as geographic region or year) that drive the emergence of specific topics, which can provide deeper insights into the data.

  The prevalence used for analyzing the current 2024's discourse of different geographic regions was "Region", while the prevalence for analyzing the global trends over time on biodiversity was "Year".

Additional visualizations were obtained by using the following:

- **plot(stm_model)** was used for visualizing the distribution of topics across documents, as calculated by the STM model.

- **plot(effects)** was used for visualizing the estimated effects of geographic regions and year on topic prevalence, as calculated by estimateEffect().

Finally, **stm$theta** was used for obtaining the topic proportions, and then obtaining the top documents for each topic. Each theta represents the proportion of topic k in document i. In other words, it informs of the weight of topic k in document i. Therefore, it can be used for identifying the documents with the highest proportions of each topic (meaning the documents that most strongly associate with each topic).

V. Model Results

5.1. Current 2024's discourse of different geographic regions on biodiversity

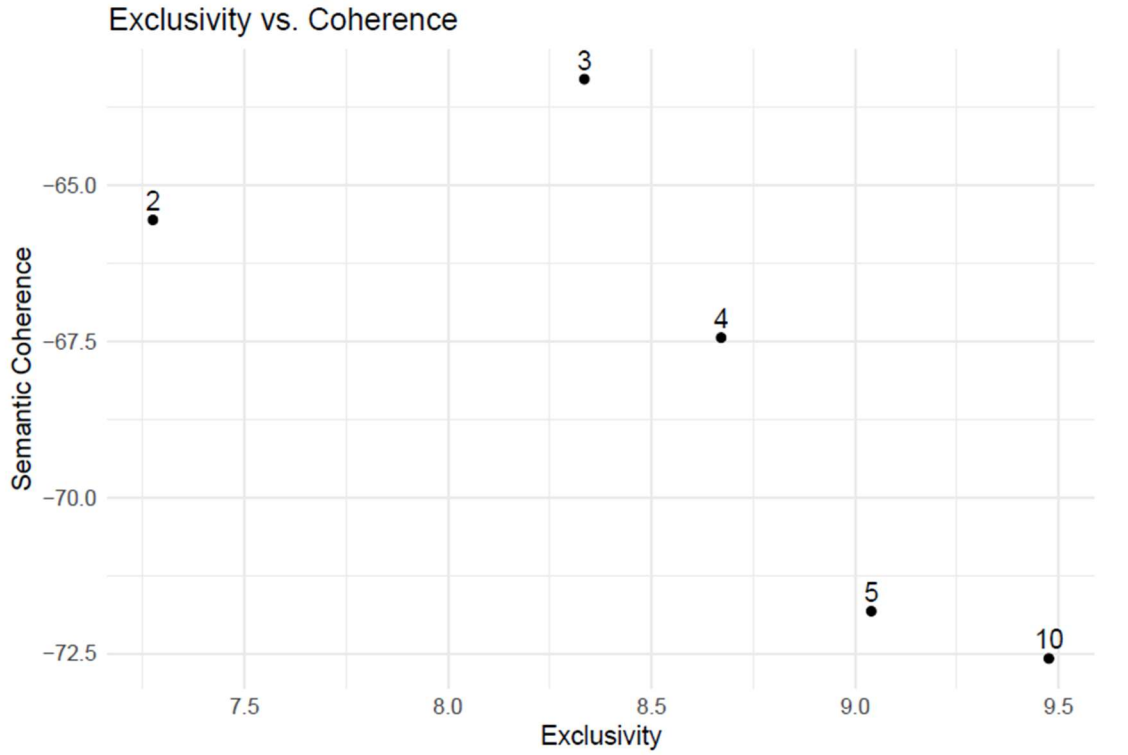
The code for analysing the data related to the discourse of different geographic regions on biodiversity is available in the R file named *FinalProject-DataAnalysisRegion*.

a) Results of selecting the optimal K

The optimal K was 3, based on obtaining the highest composite score after running searchK.

```
#Specify the K values
K_values <- c(2, 3, 4, 5, 10)

#Run searchK
k_result <- searchK(documents = docs,
                    vocab = vocab,
                    K = K_values,
                    prevalence = ~ Region,
                    data = meta,
                    verbose = FALSE,
                    seed = 42,
                    init.type = "Spectral")
```

```
#Plotting Coherence vs. Exclusivity
coherence <- sapply(k_result$results$semcoh, function(x) x[1])
exclusivity <- sapply(k_result$results$exclus, function(x) x[1])
plot_data <- data.frame(Topics = K_values, Coherence = coherence, Exclusivity = exclusivity)

ggplot(plot_data, aes(x = Exclusivity, y = Coherence, label = Topics)) +
  geom_point() +
  geom_text(vjust = -0.5) +
  labs(title = "Exclusivity vs. Coherence", x = "Exclusivity", y = "Semantic Coherence") +
  theme_minimal()
```




```
print(paste("The maximum composite score is",max(composite_scores)))
```

```
## [1] "The maximum composite score is -27.4832562808134"
```

```
print(paste("The optimal K value is", K))
```

```
## [1] "The optimal K value is 3"
```

b) Running the STM model (for optimal K=3)

The STM model was executed with the optimal number of topics, K=3, using init.type = "Spectral" and gamma.prior = "L1". The decision-making process and technical details of the parameter choices are explained in the “stm()” part of Section IV of this report.

Despite targeted efforts to enhance the model's semantic coherence (such as including lower and upper thresholds during the preprocessing, adjusting the gamma.prior parameter, and others described in Section IV), the semantic coherence scores remained suboptimal, ranging from -49.5 to -81.1. On the other hand, the Exclusivity Scores were relatively satisfactory, ranging from 7.3 to 8.8, indicating that the topics were reasonably distinct from one another.

```r
#Fit the STM model with K optimal

stm_model <- stm( documents = docs,
                  vocab = vocab,
                  K = K,
                  prevalence = ~ Region,
                  data = meta,
                  max.em.its = 500,
                  init.type = "Spectral",
                  verbose = FALSE,
                  gamma.prior = "L1")
```

```r
#Evaluate the fitted model
exclusivity(stm_model)
semanticCoherence(model = stm_model, documents = docs)
```

```
[1] 8.765075 7.291069 8.828281
[1] -49.50376 -81.07457 -52.86394
```

c) Manual selection of the name of topics (for optimal K=3)

Based on the results of the word clouds, top words, and top documents, the topics were manually named as follows:

-Topic 1: ***Cartagena Protocol on Biosafety, Nagoya Protocol on Access and Benefit-Sharing, and other formal and procedural aspects of biodiversity agreements and conventions***

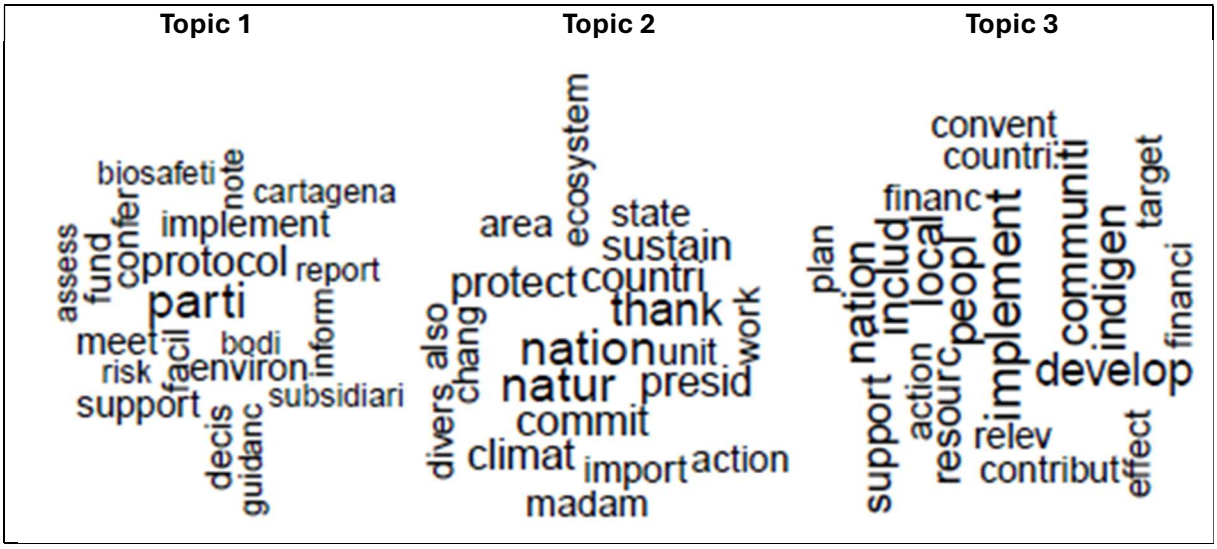
-Topic 2: ***Importance of protecting biodiversity, including restauration, ecosystem services and protected areas, as part of their national commitments***

-Topic 3: ***Gender issues, inclusive development focusing on local communities, and indigenous rights in biodiversity***

The screenshots of the codes and the results of the word clouds, top words and top documents are provided below (c1, c2, and c3).

c.1 Word clouds

```{r}
# Visualize topics in a word cloud
for (k in 1:K) {
  cloud(stm_model,
        topic = k,
        max.words = 20,
        scale = c(0.9, 0.5),
        random.order = FALSE,
        rot.per = 0.4,
        main = paste("Topic", k))
}
```



c.2 Top Words

The top words per topic based on the results of the fitted model are:

```{r}
#Summarize the fitted model
summary(stm_model)
```

```
A topic model with 3 topics, 2792 documents and a 1665 word dictionary.
Topic 1 Top Words:
         Highest Prob: parti, protocol, meet, facil, environ, confer, implement
         FREX: biosafeti, cartagena, clearinghous, modifi, nagoya, replenish, subsidiari
         Lift: accuraci, andii, cbdcopadd, con, sectionc, verifi, vii
         Score: con, protocol, cartagena, biosafeti, facil, replenish, subsidiari
Topic 2 Top Words:
         Highest Prob: nation, natur, thank, commit, countri, protect, climat
         FREX: excel, peac, gentlemen, ladi, marin, republ, distinguish
         Lift: almost, basin, beauti, begun, biospher, bird, bold
         Score: univers, protect, restor, thank, excel, natur, presid
Topic 3 Top Words:
         Highest Prob: implement, develop, peopl, communiti, local, includ, indigen
         FREX: women, youth, local, privat, financ, indigen, knowledg
         Lift: taxonomi, blend, earthcentr, ineffici, metric, naturerel, rightsth
         Score: los, indigen, local, women, privat, right, communiti
```

c.3 Top Documents

The five top documents per topic, according to the results of the fitted model are:

```r
#Get the topic proportions for each document
topic_proportions <- as.data.frame(stm_model$theta)

#Get the original documents
reviews <- data$text

#List to store top documents
top_reviews <- list()

#For each topic (from 1 to K), identify the top documents
for (k in 1:K) {
  topic_scores <- topic_proportions[[k]]
  top_docs <- order(topic_scores, decreasing = TRUE)[1:5]
  top_reviews[[k]] <- reviews[top_docs]
}

#Print the top documents for each topic
for (k in 1:K) {
  cat(paste("\nTop document for topic", k, ":\n"))
  print(top_reviews[[k]])
}
```

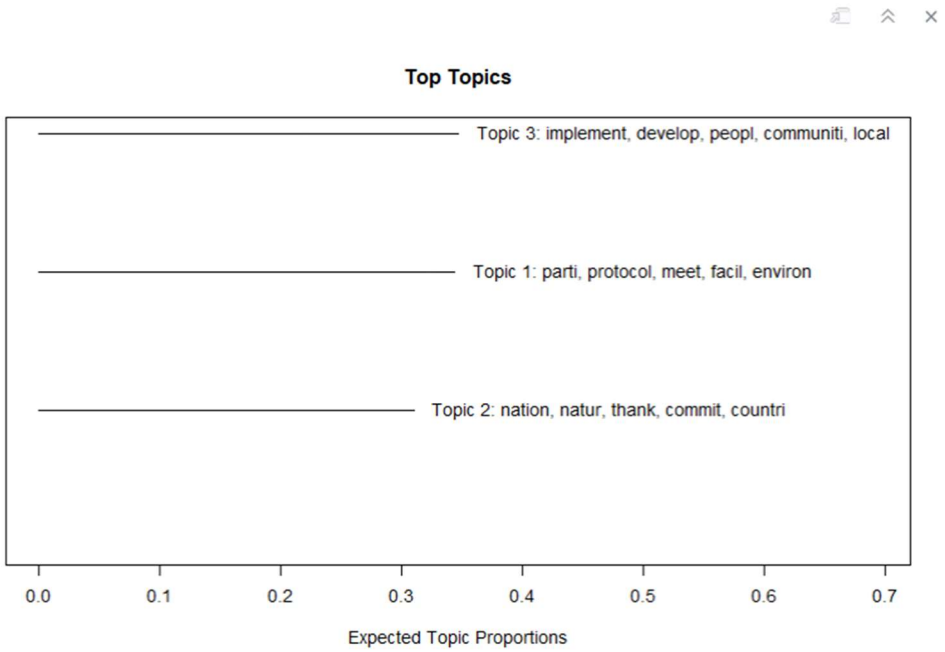
```
Top document for topic 1 :
[1] "parties protocol eleventh meeting comprise placeholder elements guidance annex additional guidance global environment facility placeholder
heading placeholder additional elements guidance adopted decisions items agenda sixteenth meeting conference parties convention biological
diversity decisions adopted conference parties serving meeting parties cartagena protocol biosafety eleventh meeting conference parties serving
meeting parties nagoya protocol"
[2] "engineered gene drives develop detailed outline additional guidance materials risk assessment living modified fish peer review preparation
work hoc technical expert group risk assessment convene online discussions openended online forum risk assessment risk management support work
hoc technical expert group risk assessment drafting detailed outline casebycase risk assessment living modified"
[3] "consideration conference parties serving meeting parties cartagena protocol fourteenth meeting item risk assessment risk management risk
assessment risk management conference parties serving meeting parties cartagena protocol biosafety recalling article cartagena protocol
biosafety recalling also decision december recalling decision november established process identification prioritization specific issues
regarding risk assessment living modified"
[4] "analyse information submitted parties paragraph decision basis prepare list prioritized topics guidance materials risk assessment may
needed according criteria set annex decision prepare report including draft voluntary guidance materials living modified fish list prioritized
topics subparagraph guidance materials risk assessment may needed consideration subsidiary body undertaking work hoc technical expert"
[5] "basis prepare list prioritized topics guidance materials risk assessment may needed according criteria set annex decision prepare report
including draft voluntary guidance materials living modified fish list prioritized topics subparagraph guidance materials risk assessment may
needed consideration subsidiary body undertaking work hoc technical expert group risk assessment shall consider synthesis"


Top document for topic 2 :
[1] "blessings upon environmental quality authority minister aziz abdukhakimov speech cbd cop high level segment national statements october
cali thank chair good afternoon ladies gentlemen extending deepest gratitude government hosting important event city cali leadership heshavkat
mirziyoyev president setting clear ambitious environmental agenda inline decade ecosystem restoration conducting massive afforestation
campaign"
[2] "ecosystem services ipbes unesco man biosphere mab stages protected area development first geobotanical reserve established creation nature
reserves sanctuaries protected areas grew first national park established hectares red book republic edition edition edition progress towards
global biodiversity targets state nature reserves territory national parks territory state sanctuaries territory total coverage"
[3] "combat impacts climate change desertification biodiversity loss recognizing addressing interrelated challenges requires integrated
approach thank translation statement english distinguished guests esteemed colleagues ladies gentlemen honor address cbd copi would like
express deep gratitude government hosting important event excellence would like make statement confirm stands resolute reaffirming full
commitment ambitious targets"
[4] "types also include various subtypes nature richness species diversity ecosystems forest ecosystems wetland ecosystems mountain ecosystems
plant species approximately recorded species endemic invertebrate species approximately species recorded vertebrate species home species bird
areas important bird areas ibas identified hosting globally threatened species semidesert ecosystems marine ecosystems coastal ecosystems
biodiversity conventions"
[5] "diversity protected lands west bank occupation confiscated natural year happening gaza strip genocide environmental genocide western
declared nature reserves turning healthy abomination considered man trees stone gaza valley destroyed important migratory birds africa europe
threatens wet area rest station protected migratory birds destruction habitats mammals reptiles ladies gentlemen platform call"
```

```
Top document for topic 3 :
[1] "cooperation synergies rio conventions global biodiversityrelated conventions multilateral agreements national subnational levels
integrated action plans programmes projects develop integrated finance solutions full involvement indigenous peoples local communities women
youth improving integration biodiversity objectives broader development cooperation frameworks national subnational accounting development
planning finance strategies using frameworks tools system environmentaleconomic accounting"
[2] "investments biodiversity continuing exploration scalingup appropriate finance solutions increasing cooperation synergies rio conventions
global biodiversityrelated conventions multilateral agreements national subnational levels integrated action plans programmes projects develop
integrated finance solutions full involvement indigenous peoples local communities women youth improving integration biodiversity objectives
broader development cooperation frameworks national subnational accounting development"
[3] "accessibility effectiveness efficiency transparency accountability resource provision use uptake accessibility effectiveness efficiency
transparency accountability resource provision use enhanced ensuring national subnational ownership aligning biodiversity policies national
development plans needs rights holders supporting policy coherence creating enhancing partnerships indigenous peoples local communities women
youth civil society enhancing capacitybuilding development technical assistance"
[4] "circumstances mainstreaming biodiversity private sector progressively aligning relevant private activities fiscal financial flows
objectives convention goals targets framework effort mainstream biodiversity finance mainstreaming biodiversity appropriate policies national
central banks regulatory authorities taking relevant distinct mandates account taking effective domestic action incentives including
inefficient subsidies harmful biodiversity accordance target framework reducing"
[5] "accessibility effectiveness efficiency transparency accountability resource provision use enhanced ensuring national subnational ownership
aligning biodiversity policies national development plans needs rights holders supporting policy coherence creating enhancing partnerships
indigenous peoples local communities women youth civil society enhancing capacitybuilding development technical assistance technological
cooperation financial planning effective resource use management improving"
```

d) Evaluation of topic imbalance

The three topics were in present in the documents in similar proportions, around one third each of them, as shown in the graph below:

```r
#Visualize topics
par(cex = 0.57)
plot (stm_model, type = "summary", n = 5)
```



e) Effect of geographical region on topic prevalence

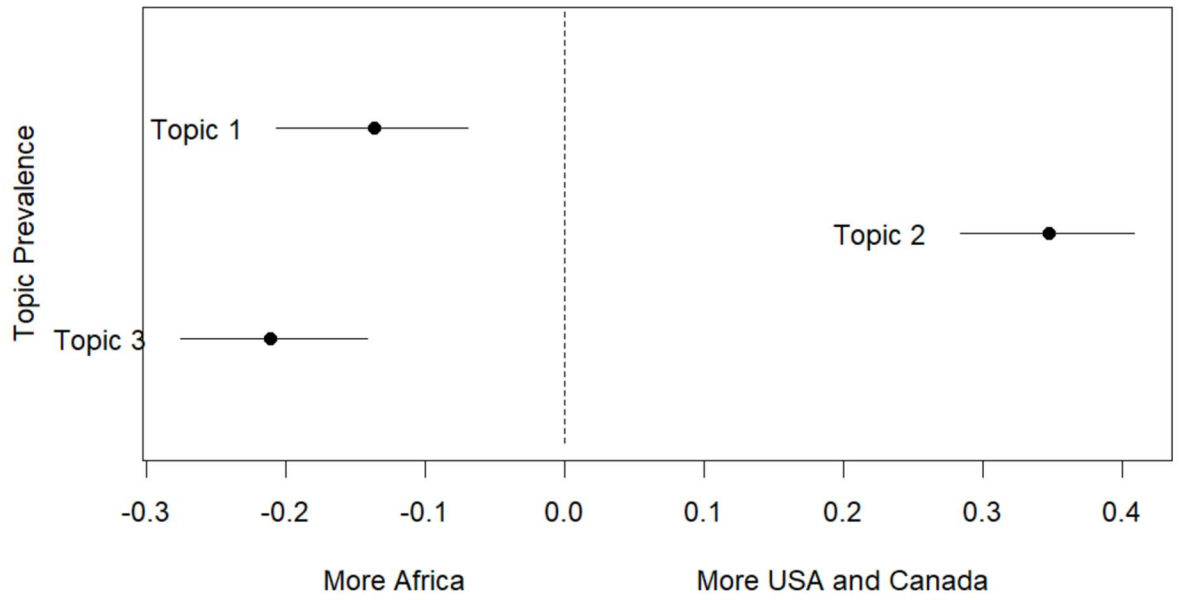
After visualizing the geographic regions against each other, I generated a total of 55 comparative graphs. The analysis revealed distinct patterns in topic prevalence across regions. For example, in Africa, **Topic 1** (focused on biosafety and benefit-sharing) and **Topic 3** (centered on inclusive development and indigenous rights) were notably more dominant compared to regions such as the USA and Canada (Graph 1) or Western Europe (Graph 2). In contrast, these latter regions

showed a stronger focus on **Topic 2** (emphasis on biodiversity protection as part of their national commitments).

This differentiation highlights how regional priorities and policies shape the discourse on biodiversity conservation. The comparison sheds light on global variations in thematic emphasis, illustrating the diversity of approaches toward biodiversity conservation.

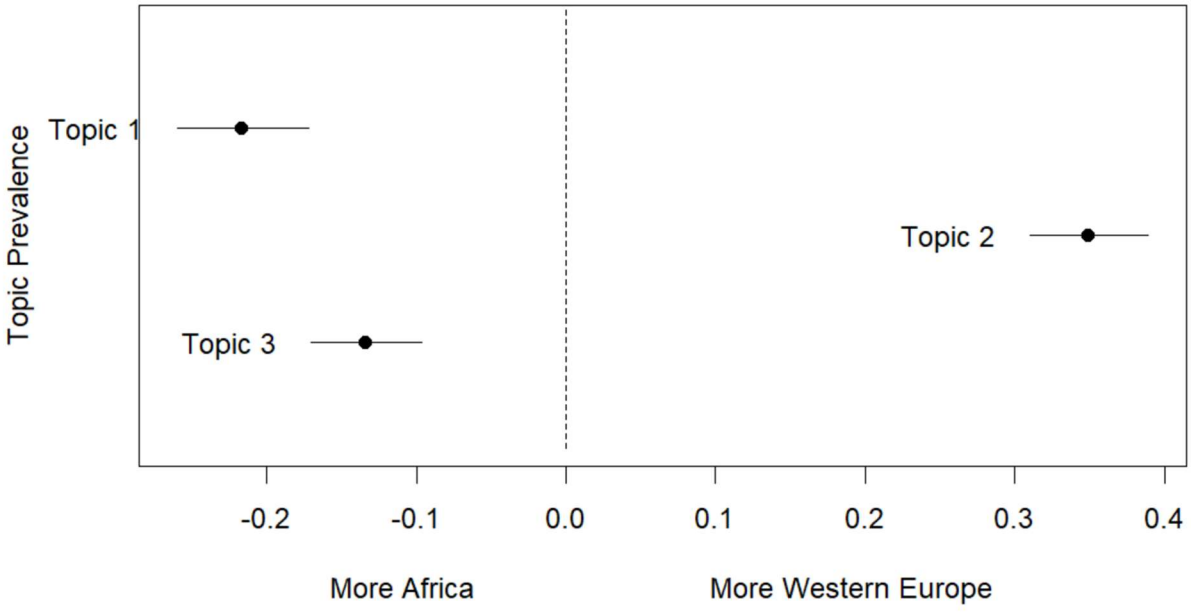
```r
#Plot of effects of region on topics
meta$Region <- as.factor(meta$Region)

regions <- levels(meta$Region)
for (i in 1:(length(regions) - 1)) {
  for (j in (i + 1):length(regions)) {
    plot(
      effects,
      covariate = "Region",
      topics = seq(1, K),
      model = stm_model,
      method = "difference",
      cov.value1 = regions[i],
      cov.value2 = regions[j],
      xlab = paste("More", regions[i], "                          ", "More", regions[j]),
      ylab = "Topic Prevalence",
      main = paste("Effect of", regions[i], "vs", regions[j], "on Topics"),
      labeltype = "custom",
      custom.labels = paste("Topic", 1:K)
    )
  }
}
```

Graph 1:
Effect of Africa vs USA and Canada on Topics



Topic 1 focuses on biosafety and benefit-sharing; **Topic 2** emphasizes biodiversity protection as part of national commitments; and **Topic 3** centers on inclusive development and indigenous rights.

Graph 2:
Effect of Africa vs Western Europe on Topics



Topic 1 focuses on biosafety and benefit-sharing; **Topic 2** emphasizes biodiversity protection as part of national commitments; and **Topic 3** centers on inclusive development and indigenous rights.

5.2. Global trends over time on biodiversity conservation (1994-2024)

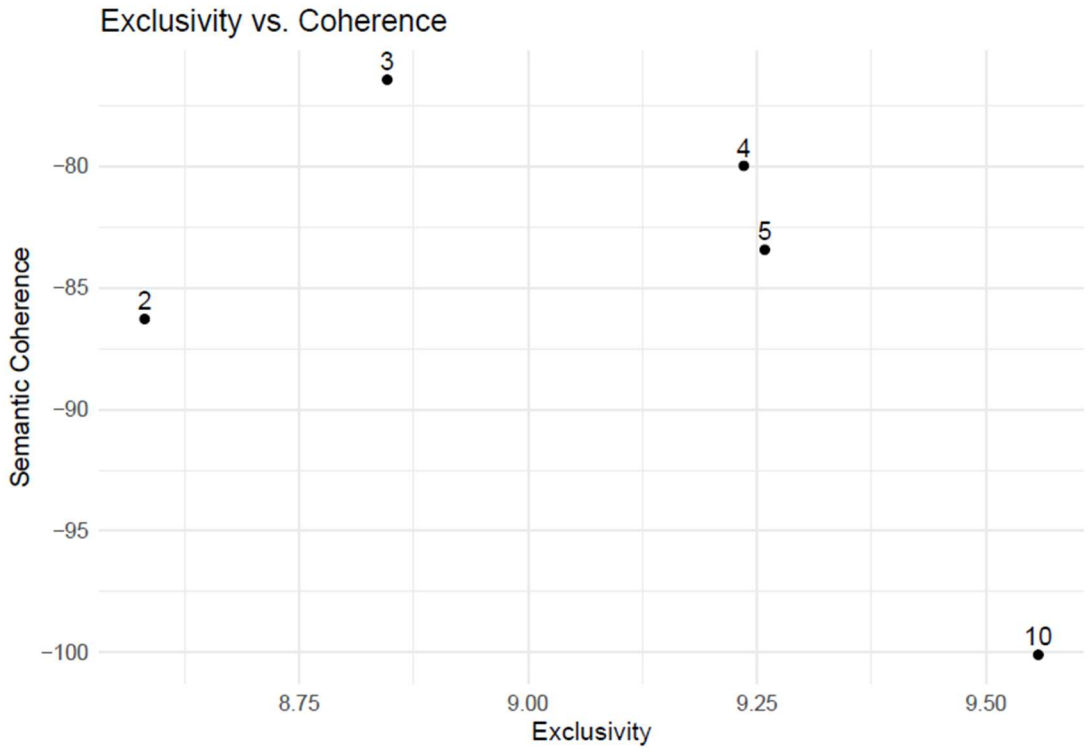
The code for analysing the data related to the global trends over time on biodiversity discourse is available in the R file named *FinalProject-DataAnalysisTrends*.

a) Results of selecting the optimal K

The optimal K was 3, based on obtaining the highest composite score after running searchK.

```
#Specify the K values
K_values <- c(2, 3, 4, 5, 10)

#Run searchK
k_result <- searchK(documents = docs,
                    vocab = vocab,
                    K = K_values,
                    prevalence = ~ Region,
                    data = meta,
                    verbose = FALSE,
                    seed = 42,
                    init.type = "Spectral")
```

```
#Plotting Coherence vs. Exclusivity
coherence <- sapply(k_result$results$semcoh, function(x) x[1])
exclusivity <- sapply(k_result$results$exclus, function(x) x[1])
plot_data <- data.frame(Topics = K_values, Coherence = coherence, Exclusivity = exclusivity)

ggplot(plot_data, aes(x = Exclusivity, y = Coherence, label = Topics)) +
  geom_point() +
  geom_text(vjust = -0.5) +
  labs(title = "Exclusivity vs. Coherence", x = "Exclusivity", y = "Semantic Coherence") +
  theme_minimal()
```



```
print(paste("The maximum composite score is",max(composite_scores)))
```

```
## [1] "The maximum composite score is -33.7982358809095"
```

```
print(paste("The optimal K value is", K))
```

b) Running the STM model (for optimal K=3)

The STM model was executed with the optimal number of topics, K=3, using init.type = "Spectral" and gamma.prior = "L1". The decision-making process and technical details of the parameter choices are explained in the "stm()" part of Section IV of this report.

Despite targeted efforts to enhance the model's semantic coherence (such as including lower and upper thresholds during the preprocessing, adjusting the gamma.prior parameter, and others described in Section IV), the semantic coherence scores remained suboptimal, ranging from -66.8 to -88.9. On the other hand, the Exclusivity Scores were relatively satisfactory, ranging from 8.0 to 9.4, indicating that the topics were reasonably distinct from one another.

```r
#Fit the STM model with K optimal

stm_model <- stm( documents = docs,
                  vocab = vocab,
                  K = K,
                  prevalence = ~ Year,
                  data = meta,
                  max.em.its = 500,
                  verbose = FALSE,
                  init.type = "Spectral",
                  gamma.prior = "Pooled")
```

```r
#Evaluate the fitted model
exclusivity(stm_model)
semanticCoherence(model = stm_model, documents = docs)
```

```
[1] 9.371155 7.965566 8.879276
[1] -70.30640 -66.81333 -88.86965
```

c) Manual selection of the name of topics (for optimal K = 3)

Based on the results of the word clouds, top words, and top documents, the topics were manually named as follows:

-Topic 1: ***Capacity building, monitoring, and implementation support for biodiversity conservation***

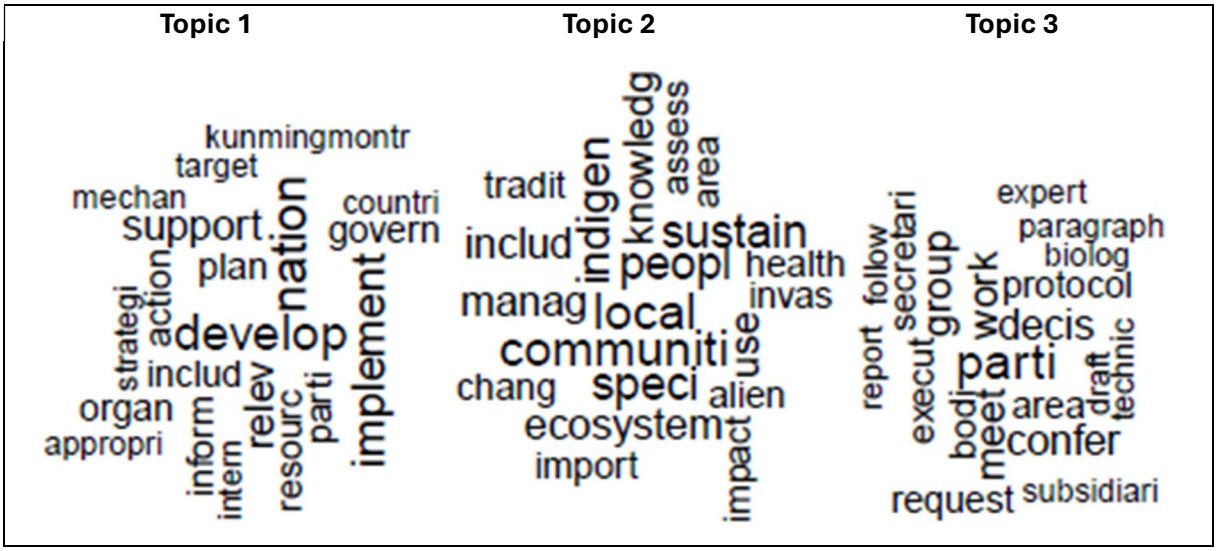
-Topic 2: ***Local communities, indigenous peoples and their importance for biodiversity conservation efforts***

-Topic 3: *Governance and decision-making processes oriented to biodiversity conservation*

The screenshots of the codes and the results of the word clouds, top words and top documents are provided below (c1, c2, and c3).

c.1 Word clouds

```
```{r}
# Visualize topics in a word cloud
for (k in 1:K) {
  cloud(stm_model,
        topic = k,
        max.words = 20,
        scale = c(0.9, 0.5),
        random.order = FALSE,
        rot.per = 0.4,
        main = paste("Topic", k))
}
```
```



c.2 Top Words

The top words per topic based on the results of the fitted model are:

```{r}
#Summarize the fitted model
summary (stm_model)

```

```
A topic model with 3 topics, 6739 documents and a 1808 word dictionary.
Topic 1 Top Words:
         Highest Prob: develop, nation, implement, relev, support, action, includ
         FREX: capacitybuild, cooper, nation, mobil, financ, resourc, multilater
         Lift: arid, assum, blend, commensur, fiscal, headlin, idea
         Score: develop, nation, plan, action, capacitybuild, strategi, implement
Topic 2 Top Words:
         Highest Prob: local, communiti, peopl, indigen, speci, sustain, use
         FREX: communiti, indigen, invas, chang, alien, tradit, impact
         Lift: carbon, classif, flora, livestock, pattern, pressur, taxa
         Score: speci, invas, alien, communiti, peopl, indigen, risk
Topic 3 Top Words:
         Highest Prob: parti, meet, decis, work, group, confer, execut
         FREX: meet, group, confer, bodi, subsidiari, draft, advic
         Lift: abyss, acipens, adhoc, ahead, azor, basin, benthic
         Score: meet, confer, draft, subsidiari, parti, session, group
```

c.3 Top Documents

The five top documents per topic, according to the results of the fitted model are:

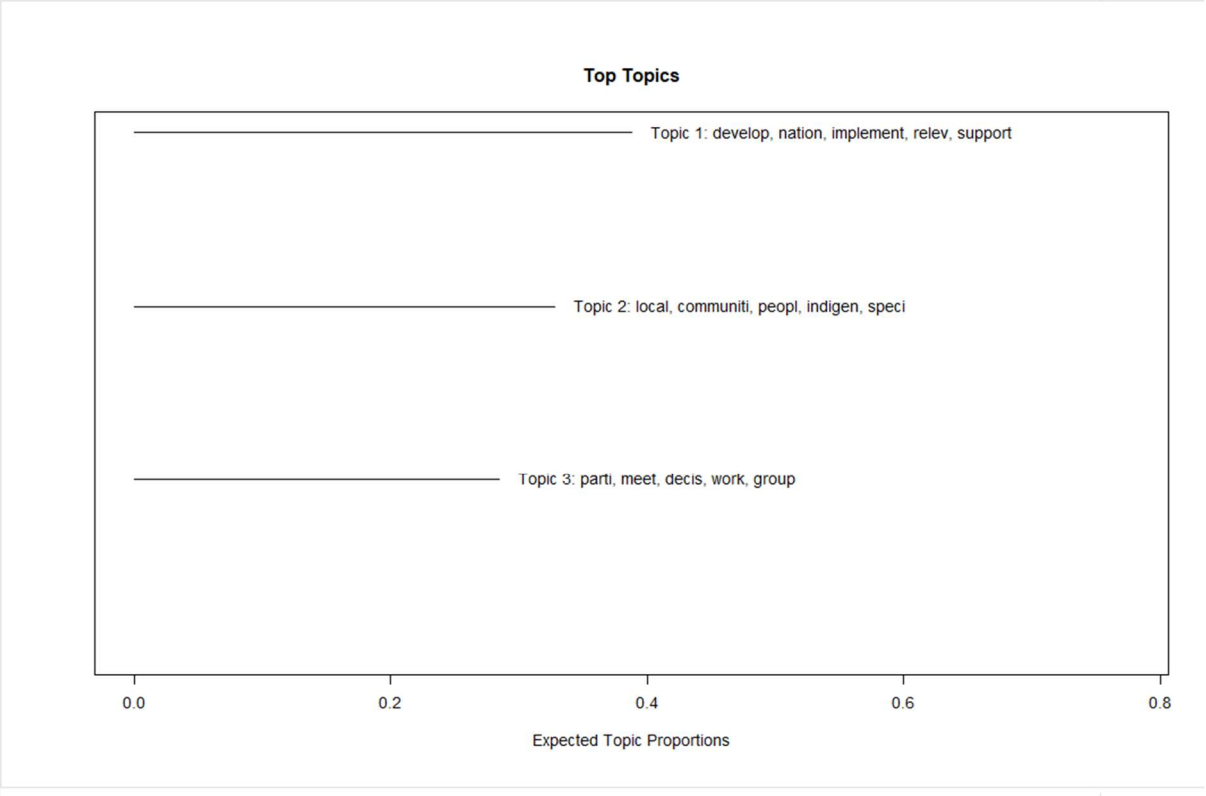
```{r}
#Get the topic proportions for each document
topic_proportions <- as.data.frame(stm_model$theta)

#Get the original documents
reviews <- data$text

#List to store top documents
top_reviews <- list()

#For each topic (from 1 to K), identify the top documents
for (k in 1:K) {
  topic_scores <- topic_proportions[[k]]
  top_docs <- order(topic_scores, decreasing = TRUE)[1:5]
  top_reviews[[k]] <- reviews[top_docs]
}

#Print the top documents for each topic
for (k in 1:K) {
  cat(paste("\nTop document for topic", k, ":\n"))
  print(top_reviews[[k]])
}
```

```
Top document for topic 1 :
[1] "to facilitate the compilation and use of these headline component and complementary indicators at the national level enabled by effective
national biodiversity monitoring systems and other information systems capacitybuilding and development activities technology and other support
will be required the secretariat together with organizations identified in the indicator metadata sheets as data providers are invited to
provide guidelines and information for the design or improvement and implementation of national monitoring systems to support the collection of
data and the calculation of headline indicators in this way parties will be able to effectively use the headline indicators as well as
component and complementary indicators supported by adequate means of implementation including capacitybuilding and development and technical
and scientific cooperation to fill monitoring gaps especially for developing countries"
[2] "national biodiversity strategies and action plans should promote and support increased efforts and actions and improved implementation and
consistency over time in a cooperative and flexible manner ensuring responsibility and transparency of information on national targets
reflecting as applicable all the goals and targets of the kunmingmontreal global biodiversity framework and including information regarding
means of implementation for developing countries in particular the least developed countries and small island developing states in line with
the relevant articles of the convention"
[3] "the mechanisms recognize the specific challenges faced by developing countries and the need for international cooperation to support them
accordingly means of implementation including capacitybuilding and development and technical and financial support will be provided to parties
especially to developing country parties to enable the implementation of these mechanisms for responsibility and transparency including
information on transparency of the support provided and received and provide a full overview of aggregate support provided"
[4] "urges all parties to implement their national biodiversity strategies and action plans in accordance with article of the convention on
biological diversityaccording to national circumstances priorities and capabilities recognizing the need for enhanced support especially for
developing countries in particular the least developed countries and small island developing states as well as countries with economies in
transition through adequate timely predictable and easily accessible means of implementation from all sources on a needs basis and taking the
considerations for the implementation as set out in section c of the framework into account"
[5] "stressing that the capacities to generate collect organize and share biodiversityrelevant data are uneven and that developing country
parties in particular and especially the least developed countries and small island developing states among them require enhanced international
support through capacitybuilding and development technical and scientific cooperation and access to and transfer of technology to strengthen
national information and monitoring systems"


Top document for topic 2 :
[1] "improved management of soil and its biodiversity offers solutions for all sectors that rely on soils including forestry and farming while
it can simultaneously increase carbon storage improve water and nutrient cycling resilience to climate change while preventing and avoiding
potential impacts arising from the implementation of soil mitigation approaches and practices on indigenous peoples and local communities
including through naturebased solutions andor ecosystembased approaches and mitigate pollution soil biodiversity depends on the type of climate
mineral soil and type of vegetation and in turn this biodiversity has an effect on soil in order to maintain or restore the biodiversity of
soils it is necessary to maintain or restore their biophysical biochemical and biological properties soil biodiversity and its biotic
interactions are important levers to improve soil quality and function highlighting the importance of research monitoring and management that
is geared directly at soil biodiversity as an integrative part and key element of soil quality soil biodiversity is also crucial to improve not
only soil health but also plant animal and human health"
[2] "climate change is associated with more frequent extreme weather events like cyclones and flooding extreme weather events cannot only
transport invasive alien species to new areas but also cause disturbances in habitats which enable invasive alien species to establish
themselves and spread climateinduced extreme weather events can also lead to sudden human population movements and displaced people can
inadvertently transport invasive alien species"
[3] "wild bee nests in nature are in danger of depletion as a result of logging practices in and it has been shown that logging reduces the
number of wild bee nests and as a consequence pollinators which has implications for forest recovery or restoration logging also reduces the
forest habitat that contains suitable unoccupied nesting sites the loss of pollinators occurs even if the current rules for certified wood
management are taken into account"
[4] "identify map and prioritize areas important for essential ecosystem functions and services including ecosystems that are important for
food eg mangroves for fisheries for climate mitigation eg carbondense ecosystems such as forests peatlands mangroves for water security eg
mountains forests wetlands and grasses that provide both surface and groundwater for poverty alleviation eg ecosystems that provide subsistence
livelihoods and employment and for disaster risk reduction eg ecosystems that buffer impacts from coastal storms such as reefs seagrass beds
floodplains"
[5] "climate change is associated with more frequent extreme weather events such as drought cyclones and flooding as well as slowonset events
extreme events can contribute to the movement of invasive and potentially invasive alien species to new areas and cause disturbances in
habitats that enable invasive alien species to establish themselves and spread they can also lead to sudden human population movements and
displaced people can inadvertently transport invasive alien species with them"


Top document for topic 3 :
[1] "if comments are received and the proponent decides to address them and if necessary provide a revised version of the submission to the
secretariat the secretariat shall transmit the submission to the subsidiary body on scientific technical and technological advice and the
conference of the parties for their consideration the proponent may also request that the submission be discussed at a workshop on ecologically
or biologically significant marine areas held pursuant to paragraph of the present decision before it is submitted to the subsidiary body"
[2] "the rd plenary session of the meeting served in part as a second stocktake session held jointly with the conference of the parties serving
as the meeting of the parties to the cartagena protocol and the conference of the parties serving as the meeting of the parties to the nagoya
protocol during the stocktake session representatives heard a report from the president on the outcomes of the highlevel segment as well as
reports by the chairs of working group i working group ii and the contact group on budgetary matters on the progress made to date"
[3] "reminds all parties to the convention that contributions to the core programme budgets general trust fund for the convention general trust
fund for the core programme budget for the cartagena protocol and general trust fund for the core programme budget for the nagoya protocol are
due on january of the year for which those contributions have been budgeted and urges all parties to pay them promptly and requests that
parties be notified of the amount of their contributions as early as possible in the year preceding the year in which the contributions are
due"
[4] "having reviewed the experience in holding concurrently meetings of the conference of the parties the conference of the parties serving as
the meeting of the parties to the cartagena protocol and the conference of the parties serving as the meeting of the parties to the nagoya
protocol using the criteria determined in decision xiii and taking into account the views of parties observers and participants at the
thirteenth meeting of the conference of the parties to the convention the eighth meeting of the conference of the parties serving as the
meeting of the parties to the cartagena protocol and the second meeting of the conference of the parties serving as the meeting of the parties
to the nagoya protocol and through the surveys conducted after the meetings"
[5] "requests that the executive secretary expedite consultations with parties on the date and venue of the sixteenth meeting of the conference
of the parties the eleventh meeting of the conference of the parties serving as the meeting of the parties to the cartagena protocol and the
fifth meeting of the conference of the parties serving as the meeting of the parties to the nagoya protocol and in the absence of any viable
offer from a party by the end of december explore in consultation with the bureau arrangements to hold the meetings at the seat of the
secretariat"
```

d) Evaluation of topic imbalance

The three topics were in present in the documents in more or less similar proportions, around one third each of them, although the proportion of topic 1 was slightly higher than for the other topics, as shown in the graph below:

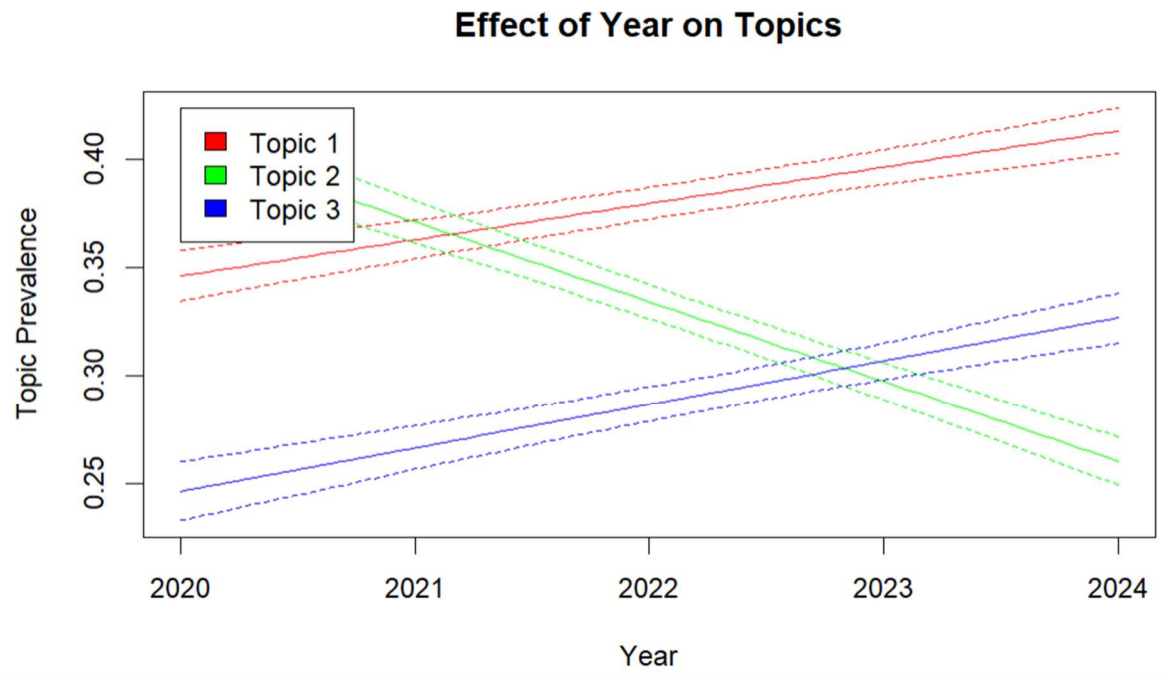
```{r}
#Visualize topics
par(cex = 0.57)
plot (stm_model, type = "summary", n = 5)

```



e) Effect of year on topic prevalence

After visualizing the effect of year on topic prevalence, **Topic 1** (focused on capacity building, monitoring, and implementation support) and **Topic 3** (centered on governance and decision-making processes) demonstrated increased prevalence in the most recent years. In contrast, **Topic 2** (emphasizing local communities and indigenous peoples) showed a declining trend over the years. Despite these patterns, the correlation between topic prevalence and year was not particularly strong. This was evident from the graph of correlation values, where Pearson correlation coefficients ranged between -0.22 and 0.12.

```r
#Plot of effects of year on topics
plot(effects, covariate = "Year",
     topics = seq(1, K),
     model = stm_model,
     method = "continuous",
     xlab = "Year",
     ylab = "Topic Prevalence",
     main = "Effect of Year on Topics",
     labeltype = "custom",
     custom.labels = paste("Topic", 1:K))
```



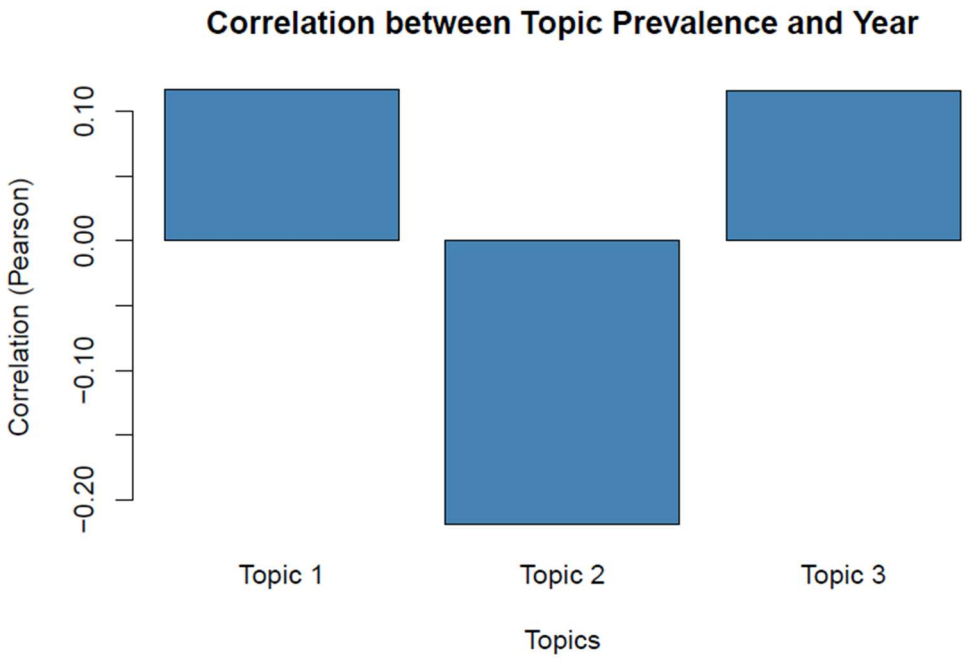

```r
#Topic prevalence
topic_prevalence <- as.data.frame(stm_model$theta)
ratings <- meta$Year

#Compute correlation between Rating and topic prevalence for each topic
correlations <- sapply(1:ncol(topic_prevalence), function(k) {
  cor(ratings, topic_prevalence[[k]], method = "pearson")
})

#Print the correlations for each topic
print("The correlations per topic are:")
```

```
## [1] "The correlations per topic are:"
```

```r
print(correlations)
```

```
## [1]  0.1166484 -0.2190608  0.1162207
```

```r
#Bar plot of correlations
barplot(correlations,
        main = "Correlation between Topic Prevalence and Year",
        xlab = "Topics",
        ylab = "Correlation (Pearson)",
        col = "steelblue",
        names.arg = paste("Topic", 1:length(correlations)))
```



VI. Conclusions and Lesson Learned

Natural Language Processing (NLP) and Structural Topic Modeling (STM) offer powerful tools for analyzing large volumes of text data. These methods enable the identification of key topics, trends over time, and geographic influences on discourse. These tasks would otherwise be prohibitively time-consuming to perform manually.

Using STM, I analyzed 306 country statements from the 2024 UN Biodiversity Conference, each averaging 1.5 pages, resulting in approximately 459 pages. Similarly, I examined 498 decision reports from UN Biodiversity Conferences spanning 1994 to 2024, averaging 10 pages each, amounting to nearly 5,000 pages. Analyzing this volume of data manually would have been unfeasible within the project timeline.

6.1. Data Cleaning Challenges

A significant portion of time was spent cleaning and preparing the data. The documents were in diverse formats (e.g., Word, PDF, HTML) and languages (the six UN official languages: Arabic, Chinese, English, French, Russian, and Spanish). The challenges included:

- Numerical markers: Varied formats like “1.”, “i.”, or “(1)” required testing different removal methods.

- Country names: UN decisions use the formal name of the countries (e.g., “Bolivarian Republic of Venezuela” instead of “Venezuela”) which necessitated additional cleaning.

- Non-ASCII characters: Artifacts like “andii” or “vii” persisted in the text despite efforts to clean the data.

Moreover, issues arose from converting PDFs to editable formats or translating documents which introduced errors, such as line breaks within paragraphs, which disrupted semantic coherence. To address this, I artificially divided the text into 50-word paragraphs. However, this method helped, sometimes it created nonsensical paragraphs, which complicated the sematic coherence of the STM models.

6.2. Model Optimization

To improve semantic coherence, I employed several strategies:

- Custom stop words: Added domain-specific terms like “biodiversity” and “convention” to enhance the textProcessor function.

- Word thresholds: Applied upper and lower frequency thresholds to exclude overly common or rare words.

- Gamma priors: Used lasso regulation and pooled distributions to balance topic proportions.

Despite these efforts, challenges in data preparation still affected model quality, as the sematic coherence only yielded low two-digit negative scores, emphasizing the need for further refinement in future iterations.

6.3. Key Findings:

(a) Related to the Geographic Region Discourse (2024)

The identified topics were:

-Topic 1 (Biosafety and Benefit-Sharing)

-Topic 2 (Biodiversity as National Commitment)

-Topic 3 (Inclusive Development and Indigenous Rights)

The prevalence of the topics differed according to the geographic region. For instance, Topic 1 (biosafety and benefit sharing) and Topic 3 (inclusive development and indigenous rights) were more prevalent in regions such as Africa in contrast to USA, Canada and Western Europe, while Topic 2 (biodiversity as national commitment) was emphasized by the USA, Canada, and Western Europe in comparison to Africa

(b) Related to Global Trends (1994-2024):

- Topic 1 (Capacity Building, Monitoring, and Implementation): Increasing prevalence in recent years.

- Topic 2 (Local Communities and Indigenous Peoples): Declining trend over time.

- Topic 3 (Governance and Decision-Making Processes): Rising prominence.

These trends suggest a shift towards institutional and governance priorities in global biodiversity discourse, possibly at the expense of local and indigenous issues. Monitoring these shifts is crucial to ensure balanced policy focus.

6.4. Next Steps

In the future, I expect to enhance model accuracy and usability through:

-Improved data cleaning and further refined automated language-sensitive workflows.

-Language-specific STM models: analyze documents in their original languages through language specific STMs, and incorporate language as metadata in the current models, mainly in the model related to the geographic region discourse.

-Scalable framework designed to streamline the analysis of the data of future UN Biodiversity Conferences for automated insights.

Despite challenges, the results provide a robust foundation for examining regional and temporal dynamics in biodiversity discourse and can inform future analyses of evolving priorities in international biodiversity policies.

References

- CRAN R Project. 2023. *Package STM*.
  Available at: https://cran.r-project.org/web/packages/stm/stm.pdf

- FAO. 2018. *The State of the World's Biodiversity for Food and Agriculture*.
  Available at: https://openknowledge.fao.org/server/api/core/bitstreams/50b79369-9249-4486-ac07-9098d07df60a/content

- FAO. 2020. *The State of World Fisheries and Aquaculture*.
  Available at: https://www.fao.org/3/ca9229en/ca9229en.pdf

- IPBES. 2019. *Global Assessment Report on Biodiversity and Ecosystem Services*.
  Available at: https://ipbes.net/global-assessment

- R Documentation for STM's estimate effect. No Year. Available at:
  https://www.rdocumentation.org/packages/stm/versions/1.3.7/topics/estimateEffect

- R Documentation for STM's exclusivity. No Year. Available at:
  https://www.rdocumentation.org/packages/stm/versions/1.3.7/topics/exclusivity

- R Documentation for STM's preDocuments. No Year. Available at:
  https://www.rdocumentation.org/packages/stm/versions/1.3.7/topics/prepDocuments

- R Documentation for STM's searchK. No Year. Available at:
  https://www.rdocumentation.org/packages/stm/versions/1.3.7/topics/searchK

- R Documentation for STM's summary. No Year. Available at:
  https://www.rdocumentation.org/packages/stm/versions/1.3.7/topics/summary.STM

- R Documentation for STM's textProcessor. No Year. Available at:
  https://www.rdocumentation.org/packages/stm/versions/1.3.7/topics/textProcessor

- R Documentation for STM's word cloud. No Year. Available at:
  https://www.rdocumentation.org/packages/stm/versions/1.3.7/topics/cloud

- UNEP. 2020. *The State of the Environment Report*.
  Available at: https://www.unep.org/resources/report/state-environment-report

- WHO. 2019. *Traditional Medicine Strategy 2014-2023*.
  Available at: https://www.who.int/medicines/areas/traditional/en/

Appendices

A. Hardware

My laptop has the following specifications:

- Processor: AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz
- RAM: 16 GB
- GPU: NVIDIA GeForce RTX 3050 Laptop GPU
- System type: 64-bit operating system, x64-based processor
- Operating System: Windows 11 Home

B. Jupyter and R Files

The list of the Jupyter and R files created for the final project is provided below:

b.1 Jupyter files (for cleaning the data)

-FinalProject-DataCleaningRegion (for loading and cleaning the data associated with the geographic region differences on biodiversity discourse)

-FinalProject-DataCleaningTrends (for loading and cleaning the data associated with the trends on biodiversity discourse over time)

b.2 R files (for analysing the data)

-FinalProject_DataAnalysisRegion (related to the geographic region differences on biodiversity discourse)

-FinalProject_DataAnalysisTrends (related to the trends on biodiversity discourse over time)

I uploaded on Canvas these jupyter and R files, along with their pdf versions.

In addition, I provided the links to all data sources in this report (see Section II). As I cannot upload all the data sources (315 for country statements and 498 for the decision documents) on Canvas, I am only uploading the data files obtained after data cleaning (named outcomes_COP16_Countries.csv and outcomes_COP_Biodiversity.csv), which were the same that were uploaded to the R files for the data analysis (after data cleaning in Jupyter).