Documentation: PROJECT Apollo

Authors:
Harshvardhan Singh
University of Arizona

Dr. Charles J. Gomez
University of Arizona

Documentation Update 2.0:

New Direction:
We're emphasizing the changing dynamics of countries and not focus as much on the dynamics and content of the field (of astrophysics/astronomy). In other words, our primary focus won't be which topics dominate the field and what those topics are, what telescopes were used, etc. Our plan is to eventually do this analysis for hundreds of fields, meaning we'll need to create a pipeline in R and Python (not leveraging RStudio or Jupyter Notebook) on the HPCs, the codes have been written keeping this in mind for an easy integration with the pipeline.

As far as studying how the dominance/influence of a country changes when the field changes is concerned, Fields don't tend to radically change. They're fairly conservative. Our focus is on countries, but looking at their dynamics across fields. So, fields are in the background in terms of our analyses.

We're using a citation window of five years, so if a paper is published in the year 2000, we consider the number of citations received in the same year (2000) and the years 2001 through 2005. However, we also deflate that number using a citation deflation metric of the ratio of papers published in the year of the paper receiving citations (e.g., N_Cited_2000) and the total number of papers in the citing year (e.g., N_Citing_2000, N_Citing_2001, ..., N_Citing_2005). Now we only consider data published from **1990 to 2020**. This is because we're no longer taking a historical perspective of how fields develop. We split the corpus into five five-year periods: 1990-1994, 1995-1999, 2000-2004, 2005-2009, 2010-2015. We stop at 2015 to ensure a five-year time window for citations, so papers published in 2015 will have citations received from

2015 through 2020. We then run a separate STM for each of the five corpora: STM_1990-1994, STM_1995-1999, ..., STM_2010-2015.

For each STM, find the optimal number of K topics using the topic coherence and exclusivity measures. And use ChatGPT4 to label the topics. We compare the values of K across the STM models and then choose a suitable value for all of them to keep uniformity. The covariates for each STM model are top **15** countries and **6** continents. We chose the top N = 15 countries (FOR NOW) from all STM models 1990 to 2015 and kept it consistent across the 5 different STM models.

We decide this by checking semantic coherence and exclusivity to see which topics are redundant. Here's how to do this in STM (https://warin.ca/shiny/stm/#section-evaluate).

Moving forward, it will be interesting to decide how we choose the top 15 countries for the other fields.

On the one hand, the different countries should be consistent (i.e., the exact same) across fields. This is important because it may be the case that one country is less prominent in one field than another and we need a way to compare across fields. Although, our suspicion is that in the vast majority of cases, it'll be the same countries consistently in the top (e.g., US, UK, China, Japan, Germany, etc.). And this should be informed not by the data per se, but by prior work.

On the other hand, we could just pick the top 15 countries per field, by the number of papers in the field. This would make running the STMs easier. However, we should also consider more periphery countries that won't have a large presence in our data (e.g., countries in the Global South, etc.), like we do now. However, we aggregate across fields, how do we compare a country that may have it made it to the top 10 in astronomy but is classified as part of a wider region in another? For instance, India may be prevalent in astronomy but not so much in genetics and its part of a generic "ASIA" category?
When we prep our data and code to do hundreds of fields, we should look at the TOTAL number of papers pooled across all fields to determine which countries make up the top N = 15 by publication between 1990 and 2015. We should then see what percentage of papers these top N = 15 produce out of the entire (across fields) corpus. We want to choose an N where 80% of the papers produced by N countries between 1990 and 2015 across all fields in our data.

We calculate how prevalent the topic is across the corpus for each topic in each STM.
https://stackoverflow.com/questions/68984687/r-stm-topic-proportion-table
https://warin.ca/shiny/stm/#section-visualize
The topic prevalence is only calculated for that STM model (one of the 5 models), this is because even though two topics from different STM models (of different time intervals) might have very similar keywords, they are essentially two different topics.

Then, for each topic in each STM, we calculate the total number of citations received by each topic in the five-year time window but normalized by the citation deflation scores built from the number of papers published and the number of citing papers in the future year. Next, we capture the country covariate loadings on each topic for each STM and also capture the topic prevalence in the corpus from the STM and the citations received by that topic. We also get the indegree centrality of each country category for each of the five time periods (the number of citations [deflated] received in a five-year time period by each country) and the number of papers published by each country in each time period. (We suspect that countries that are prominent in very prevalent topics in the field (by proportion) are different (at least in terms of proportion) in topics that are highly cited. And this dynamic is changing over time.)

For the citation data, we only need the in-citations file. The file reads as the paper being cited ID, the year it was cited, and a long string of all the citations it received with the year the citation was published. To separate this long list of citing IDs out, we first parse by a comma delimiter using Python. We create a list of these citations, where each element is set up as: year_citingworkid. So, within each element of the list, we parse it by an underscore to filter cites within the five-year time frame.

In the case of international coauthors, like the US and India coauthoring a paper, we count it equally as 1 for the US and 1 for India.

We're going to focus on hundreds of separate citation networks, one for each field (assuming there are hundreds of fields). And, we're going to likely construct different citation networks for different time periods too. Whether papers are considered a part of multiple fields (and there are many) isn't really our concern here. The nature and dynamics of different fields isn't our focus here, either. Fields are a good way to peer into how countries are dominating science.

Data Cleaning and Preprocessing: Refer February 2024 documentation

After preprocessing the OpenAlex data, we merged columns for the citations count (5 year window) and deflated citations count into our main dataframe. There are only 15618 abstracts with non zero citations, this is because the work IDs of 188574 Abstracts (majority) were absent from the incitations dataset. But this is normal because it's pretty common for papers to never be cited, especially within our time window. We just keep them as zero citations raw and deflated.

We used the following formula for Deflated citation count -

```
def calculate_deflated_citations(row, growth_rate, baseline_year=2020):
    years_diff = baseline_year - row['year']
    deflation_index = (1 + growth_rate) ** years_diff
    return row['total_filtered_citations'] / deflation_index
```

This function should calculate each deflated citations for each row of the abstract dataframe

Here, the arguments for the function are -
row : represents the rows that we are iterating through each of the abstract dataframe using the df.apply() function
growth_rate

growth_rate: represents the annual growth rate of publication. I calculated it as following -

```
periods = 2015 - 1990
# Define the function to calculate CAGR
def calculate_cagr(end_value, start_value, periods):
    return (end_value / start_value) ** (1 / periods) - 1
```

baseline_year: The year to which citation counts are adjusted. I set it to year 2020

Also, I will briefly explain the body of the function -
years_diff = baseline_year - row['year'] : This calculates the difference between baseline year and the publication year
deflation_index = (1 + growth_rate) ** years_diff : calculates the deflation index
return row['total_filtered_citations'] / deflation_index: We return the value of deflated citation count with this formula

N.B. -
1. For the calculation of annual growth rate, we take the starting year as 1990 and the ending year as 2015
2. We take 2020 as the baseline year
3. We calculated the annual growth rate of publication as 10.79%, and used the decimal notation as 0.1079.

We created two CSV/dataframes:

CSV/Dataframe 1 | Summarized STM Topic Data

- [KEY] The name of the field (in this case, "astronomy/astrophysics").
- [KEY] STM time period (e.g., 1990-1994, 1995-2000, etc.)
- the topic number from the STM (e.g., 1, 2, 3, 4...etc.)
- the topic label from ChatGPT (e.g., "galactic nuclei," "experimental methods," etc.
- The raw number of citations the topic received in the five-year window as Sum and Percentage
- Citation deflated (sum of citation deflated) as Sum for one column and Percentages as another column.
- The prevalence of that topic in the corpus from the STM as Sum for one column and Percentages as another column.
- *Weighted deflated citation count for each topic proportion within every document (count and percentage)*

- *Total weighted citation proportion for the topics*
- *Total document proportion by topics*
- *Percentage of weighted citation proportion*
- *Percentage of total document document proportion*

CSV/Dataframe 2 | Summarized Country Covariates from the STM Model

- [KEY] The name of the field (in this case, "astronomy/astrophysics").
- [KEY] STM time period
- The country covariate column, where each row is a country covariate (e.g., "United States," "China," "South America," etc.)
- The number of citations received by the country covariate (as rows) in the five-year window as a Sum for one column and as a Percentage as another column.
- The citation deflated number of citations received by the country covariate in the five-year window as a Sum for one column and a Percentage as another column.
- The number of papers with at least one author from that country in that time period.
- Beta coefficient for that country from the STM
- The standard errors of that estimate of that country covariate from the STM.
- The t-value for the beta coefficient for that country from the STM.

Two additional dataframes -

1. Coauthorship dataframe - Calculates the topic load percentage for each country Here is the logic for creating the dataframe -

“First, we create a data frame of STM-model, document, and country authors, e.g.,

STM1, doc1, US, 3
STM1, doc1, China, 1
STM1, doc2, Canada, 4
STM1, doc3, UK, 2
....

Second, we calculate the percentages for each document as another column:

STM1, doc1, US, 3, 0.75
STM1, doc1, China, 1, 0.25
STM1, doc2, Canada, 4, 1.00
STM1, doc3, UK, 2, 1.00
.....

Drop the count column and keep the percentages.

Then, we take the document to topic percentages you already calculated:

STM1, doc1, Topic1, 0.0025
STM1, doc1, Topic2, 0.0003
STM1, doc1, Topic3, 0.00004
.....

And do an ALL/OUTER JOIN merge based on STM model and document (there should be A LOT of rows) :

STM1, doc1, US, Topic1, 0.75, 0.0025
STM1, doc1, US, Topic2, 0.75, 0.0003
STM1, doc1, US, Topic3, 0.75, 0.00004
.....

Next, multiply the country coauthor percentages by the topic weights (we'll call this new column coauthor-topic-loads):
STM1, doc1, US, Topic1, 0.75 * 0.0025
STM1, doc1, US, Topic2, 0.75 * 0.0003
STM1, doc1, US, Topic3, 0.75 * 0.00004
.....

Then, do a group by STM model, topic, and Country and sum coauthor-topic-loads column:
STM1, Topic1, US, 25.25
STM1, Topic1, China, 15.34
STM1, Topic1, Canada, 7.233
.....

Finally, convert the coauthor-topic-loads to percentages by model and topic, so summing by model and topic equals 1; if you could send me the below CSV that'd be great :
STM1, Topic1, US, 0.35
STM1, Topic1, China, 0.24
STM1, Topic1, Canada, 0.05
.....
“
We merge the above data frame with our data frame on the number of citations received by topic to see which countries are "occupying" the most important topics. We hypothesize that the most prominent countries also occupy topics in the field that receive the most citations.

2. STM effects dataframe - to examine the effects of covariates on the prevalence of topics. Columns -
   Value: represents the point estimate of the effect that the covariate has on the prevalence of a particular topic. It indicates how much the presence or increase of the covariate value is expected to change the topic proportion, holding other covariates

constant. Positive values indicate an increase in topic prevalence with an increase in the covariate, while negative values suggest a decrease.

Proportion: indicates the proportion of the effect that is statistically significant, or it might represent the proportion of the bootstrap samples (if bootstrap methods are used for interval estimation) that supported the effect.

lower and upper: These columns represent the lower and upper bounds of a confidence interval (CI) or credible interval for the effect estimate. The lower value is the start of this interval, and the upper value is the end. If the interval does not include zero, it suggests that the effect is statistically significant at the chosen level of confidence.

Now in order to fit the STM model, we use the STM() function with the following set of arguments-

```
Research_topics <- stm(documents = out_text$documents,
                       vocab = out_text$vocab,
                       K = 35,
                       prevalence = prevalence_formula,
                       data = out_text$meta,
                       init.type = "Spectral",
                       max.em.its = 1000,
                       gamma.prior = 'L1')
```

Arguments -

1. documents:

documents = out_text$documents: This argument specifies the document-term matrix or a list where each entry corresponds to a document in the corpus. Each document is represented by a vector of word indices.

1. vocab:

vocab = out_text$vocab
This is the vocabulary of the corpus, a vector of words used across all documents. The vocabulary must align with the indices used in the documents list or matrix.

2. K:

K = 35
This parameter specifies the number of topics to be estimated by the model. Here, the model will estimate 35 distinct topics.

3. prevalence:

prevalence = prevalence_formula
This argument is a formula that defines how metadata influences the prevalence of topics across the documents. It's used to model topic prevalence as a function of document-level covariates included in the data argument.
We removed one regional variable (South America - for no particular reason) from the equation to avoid the dummy variable trap
The prevalence formula after which looks like the following -
prevalence_formula_str <- "~ US + FR + IT + CN + ES + AU + PL + RU + DE + IN + NL + GB + JP + CA + KR + NorthAmerica + Europe + Africa + Asia + Oceania"
prevalence_formula <- as.formula(prevalence_formula_str)

4. data:

data = out_text$meta
This parameter provides a data frame containing the metadata for each document. This metadata is referenced by the prevalence_formula to understand how different covariates may influence the presence and proportions of topics within the corpus.

init.type:
init.type = "Spectral"
This specifies the method for initializing the topic proportions used to fit the STM. “The "Spectral" method uses spectral decompositions for initialization, which can lead to faster convergence and often more interpretable topics compared to random initialization.”

max.em.its:
max.em.its = 1000
This sets the maximum number of iterations. Here, the algorithm is allowed to run up to 1000 iterations to reach convergence or until other stopping criteria are met, we choose such a high count so that the model will keep iterating till it converges (and not have a limitation), which typically takes fewer than 100 iterations in our model. It’s not a very high iterations count so I believe there isn’t much of overfitting despite setting such a high **max.em.its**

gamma.prior:
gamma.prior = 'L1'
This argument defines the type of prior to use for the topic proportions. The 'L1' prior encourages sparsity in the topic proportions, meaning that each document is likely to be represented by fewer topics, making the topics more distinct and interpretable.

The csv files are created modularly for each STM model, and then they are bound together. We have a column named “time_interval” to identify one of the 5 models each column belongs to.

We calculate load of the STM models from summary the function estimateEffect - https://www.rdocumentation.org/packages/stm/versions/1.3.7/topics/estimateEffect
Effects are plotted based on the results of estimateEffect which contains information on how the estimates are constructed. Note that in some circumstances the expected value of a topic proportion given a covariate level can be above 1 or below 0. This is because we use a Normal distribution rather than something constrained to the range between 0 and 1. If a continuous variable goes above 0 or 1 within the range of the data it may indicate that a more flexible non-linear specification is needed (such as using a spline or a spline with greater degrees of freedom).

Finally we use the summary(estimateEffect) function to obtain the beta coefficients and other statistical values. We observe that beta values are either positive or a negative value, the negative values could mean that in these instances perhaps the dependent variable (beta coefficient of the country covariate) decreases when comparing it with the dummy variable. So maybe we could say something like "these show a negative influence or effect for having a negative beta coefficient, relative to the South America (dummy variable)".

Documentation 1.0:

Content -

Summary -

The project is based on uncovering historical trends in research works in the field of Astronomy and astrophysics. Our goal is to study trends and patterns in which information and collaboration is shared in international collaboration. And for this, Astronomy and Astrophysics turns out to be the perfect field for our study because this in particular is one of those unique fields where international collaboration is crucial when it comes to sharing of satellite technology, sharing of cosmic and celestial data, etc. In the study we employ NLP algorithms and perform Network Analysis on a vast dataset. It's a pioneering approach of analyzing that is also novel in the context of Astronomy especially when it comes to topic modeling and citations network analysis. In order to better explain the project, I will talk about 3 overarching tasks containing their own set of sub-tasks. First task is to understand which countries contribute to different topics within the field of Astronomy and Astrophysics. This is done by implementing topic modeling on research works dating all the way back to 1824. The dataset for this task consists of 64,000 research abstracts along with their metadata which is important because I am using the very recently developed Structural Topic Modeling algorithm which is more advanced than LDA model because it also incorporates metadata into the model in order to produce a more nuanced set of topics. Second task is to identify the set of telescopes, satellites and arrays utilized by different countries to perform their research. This task is challenging because the data is unlabeled and requires a combination of manual labeling and classification algorithm with weak supervision. Third task is to perform Network Analysis to understand the patterns in which flow of Information and collaboration takes place among different countries and institutions. I am creating networks for different types of data including citation and coauthorship data (year 2000 onwards). And then also a network sharing satellite data, for which the second task was performed.

ChatGPT's generated abstract based on above summary -

Title: "Unveiling Global Trends in Astronomy and Astrophysics Research: A Novel Approach Using NLP and Network Analysis"

Abstract:

This pioneering research project aims to illuminate the historical and contemporary trends in the field of Astronomy and Astrophysics using advanced Natural Language Processing (NLP) algorithms and comprehensive Network Analysis techniques. The project is distinguished by its innovative methodology, involving the first-ever application of these techniques for topic modeling and citation network analysis in this field. Our approach encompasses three primary tasks, each with its own set of sub-tasks, to explore the global landscape of astronomical research.

The first task involves analyzing a dataset of 64,000 research abstracts dating back to 1824, along with their metadata. By employing the newly developed Structural Topic Modeling algorithm, which surpasses the conventional Latent Dirichlet Allocation (LDA) model by integrating metadata, we aim to discern how different countries contribute to various topics within Astronomy and Astrophysics. This nuanced topic modeling provides insights into the evolution and current state of research topics globally.

The second task addresses the challenge of identifying the telescopes, satellites, and arrays used by different countries for research. Due to the unlabeled nature of the data, this task requires a hybrid approach combining manual labeling and a weakly supervised classification algorithm. The outcome of this task is crucial for understanding the technological infrastructure supporting astronomical research worldwide.

The final task involves performing Network Analysis to uncover the patterns of information flow and collaboration among countries and institutions. This involves creating networks based on different types of data, including citation and coauthorship, as well as networks for sharing satellite data, which is informed by the findings of the second task.

This project is driven by the inherently international nature of Astronomy and Astrophysics, a field that necessitates global collaboration and data sharing. By mapping the intricate web of contributions, collaborations, and technological resources across countries, our study aims to provide a comprehensive understanding of the global dynamics in astronomical research. This novel approach not only sheds light on the historical and current trends in the field but also sets a foundation for future collaborative efforts and advancements in astronomical research.

ChatGPT’s Introduction -

Introduction:

The field of Astronomy and Astrophysics stands as a testament to human curiosity, driving us to explore the cosmos and seek answers to the universe's most profound mysteries. This quest for

knowledge is inherently global, transcending borders and necessitating international collaboration. As we delve deeper into the celestial realms, understanding the historical and contemporary landscape of research in Astronomy and Astrophysics becomes crucial. This study is poised at the forefront of such an understanding, employing pioneering methods to analyze and interpret trends, collaborations, and advancements in the field.

Astronomy and Astrophysics have evolved significantly since the first recorded telescopic observations. Over the centuries, this evolution has been marked by remarkable discoveries and technological advancements. However, the distribution and development of research in this field have not been uniformly explored or understood. Our project aims to fill this gap by offering a comprehensive analysis of the global trends in Astronomy and Astrophysics research, employing a novel combination of Natural Language Processing (NLP) algorithms and Network Analysis.

The use of NLP in this context is groundbreaking. While NLP has seen various applications in other domains, its utilization in analyzing research trends in Astronomy and Astrophysics, particularly through topic modeling and metadata integration, is unprecedented. Our approach involves analyzing a vast dataset of research abstracts dating back to 1824, which enables us to trace the evolution of topics and contributions over two centuries. The implementation of the Structural Topic Modeling algorithm, which incorporates metadata, allows for a more nuanced understanding of the data compared to traditional models like LDA.

Furthermore, the project extends beyond topic analysis to explore the infrastructure that underpins astronomical research. By identifying the array of telescopes, satellites, and arrays utilized globally, we gain insights into the technological and logistical aspects of research in different countries. This facet of the study is challenging due to the nature of the data, requiring innovative methods for data labeling and classification.

The cornerstone of this research lies in its Network Analysis. By creating networks based on citation, coauthorship, and satellite data sharing, we aim to unravel the complex patterns of collaboration and information flow across international borders. This analysis is crucial in understanding how knowledge and resources are shared and how the field progresses through collective efforts.

In essence, this study is not just about Astronomy and Astrophysics as isolated scientific disciplines but also about the global scientific community's collaborative efforts to advance our understanding of the universe. Through this research, we endeavor to provide a detailed map of the field's landscape, offering insights into past trends, current dynamics, and potential future directions. This work, therefore, serves as both a reflection of the field's rich history and a guidepost for its future trajectory.

1) Preprocessing and Preliminary analysis

The country column in our dataset contains more detailed information about the authors' affiliations, including first, middle, and last authors. However, this brings a level of complexity and potential confusion for machine learning models. Therefore, we propose a method to simplify the country column by retaining only the first and last author affiliations, ensuring a more straightforward representation of authorship.

N.B. We plan on revisiting this step and see how we could incorporate third authors in our study (to reflect country contributions more accurately). It's difficult to determine the right method of doing so, therefore we chose to exclude a third author for now.

We collapsed the countries (except the top 15 countries) into the following regions based on a low contribution of 8.5%-
North America
South America
Europe
Asia
Oceania
Africa

Aggregating less-represented countries into regions can help the model learn broader patterns without getting overfitted to noise from countries with sparse data. This step also reduces the dimensionality of the data, which can be beneficial for models that are sensitive to high-dimensional feature spaces.

List of Countries-

Top 15 - 91.47% of global contribution
1. USA (US) - 50.61%
2. Great Britain (GB) - 10.13%
3. Germany (DE) - 5.73%
4. Italy (IT) - 3.84%
5. Australia (AU) - 2.82%
6. Canada (CA) - 2.67%
7. Japan (JP) - 2.59%
8. China (CN) - 2.18%
9. Netherlands (NL) - 2.01%
10. France (FR) - 1.98%
11. Russia (RU) - 1.85%

12. Spain (ES) - 1.78%
13. India (IN) - 1.42%
14. Israel (IL) - 0.96%
15. Switzerland (CH) - 0.90%

The Americas (North/South America)
1. BR - Brazil
2. CL - Chile
3. VE - Venezuela
4. MX - Mexico
5. CO - Colombia
6. AR - Argentina
7. UY - Uruguay
8. JM - Jamaica
9. CR - Costa Rica
10. GT - Guatemala
11. HN - Honduras
12. PR - Puerto Rico
13. VI - U.S. Virgin Islands
14. BO - Bolivia
15. PE - Peru
16. EC - Ecuador
17. CU - Cuba

Europe
1. RS - Serbia
2. DK - Denmark
3. SE - Sweden
4. BE - Belgium
5. PL - Poland
6. IE - Ireland
7. FI - Finland
8. AT - Austria
9. HU - Hungary
10. LV - Latvia
11. UA - Ukraine
12. GE - Georgia
13. GR - Greece
14. HR - Croatia
15. TR - Turkey
16. AM - Armenia
17. CZ - Czech Republic

18. PT - Portugal
19. SK - Slovakia
20. IS - Iceland
21. NO - Norway
22. BG - Bulgaria
23. SI - Slovenia
24. EE - Estonia
25. CY - Cyprus
26. LU - Luxembourg
27. MT - Malta
28. MK - North Macedonia
29. RO - Romania
30. LT - Lithuania

Africa
1. NG - Nigeria
2. MZ - Mozambique
3. BW - Botswana
4. ZA - South Africa
5. GH - Ghana
6. RW - Rwanda
7. UG - Uganda
8. EG - Egypt
9. TZ - Tanzania
10. MU - Mauritius
11. ZW - Zimbabwe
12. NA - Namibia
13. SD - Sudan
14. TN - Tunisia
15. BD - Burundi

Asia/Oceania
1. KR - South Korea
2. TW - Taiwan
3. IR - Iran
4. KZ - Kazakhstan
5. KH - Cambodia
6. PK - Pakistan
7. VN - Vietnam
8. MY - Malaysia
9. SA - Saudi Arabia
10. UZ - Uzbekistan
11. TH - Thailand
12. NP - Nepal

13. LK - Sri Lanka
14. ID - Indonesia
15. TJ - Tajikistan
16. AE - United Arab Emirates
17. LB - Lebanon
18. SG - Singapore
19. AZ - Azerbaijan
20. IQ - Iraq
21. OM - Oman
22. KW - Kuwait
23. QA - Qatar
24. NZ - New Zealand
25. PS - Palestine
26. BD - Bangladesh

We also decided to bin countries into 11 intervals (not anymore) -

```
intervals <- list(
  c(2020, 2022),
  c(2015, 2019),
  c(2010, 2014),
  c(2005, 2009),
  c(2000, 2004),
  c(1995, 1999),
  c(1985, 1994),
  c(1975, 1984),
  c(1965, 1974),
  c(1900, 1964),
  c(1824, 1899)
)
```

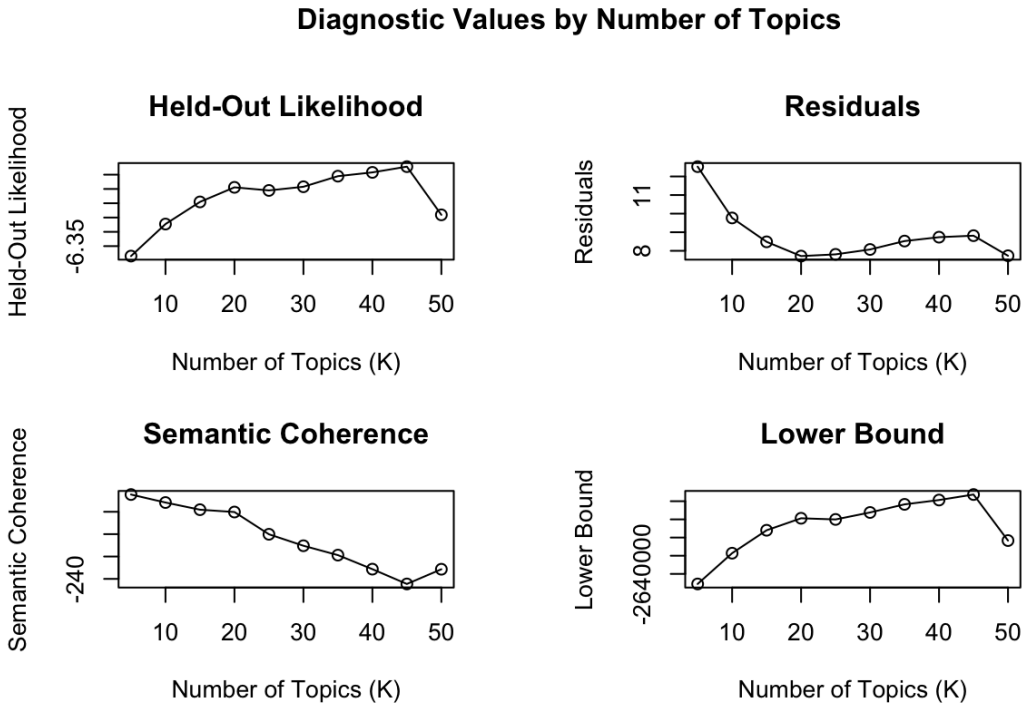
Binning publication years into intervals can help models capture trends over time without being overly sensitive to year-to-year fluctuations. This approach also helps in maintaining a manageable number of features.

It is important to use the right encoding for the models, we use weighted encoding for country contribution, where each country's contribution is represented by a float value on a scale of 0 to 100.

On the other hand, time intervals use one-hot-encoding where the time interval that the abstract belongs to is represented by 1, and all other time intervals are represented by 0.

2) STM Model

We utilized the searchK() function from the STM package to plot the following diagnostic values in order to figure out the optimal number of topics (K) we should have for our dataset -
a) **Held-out Likelihood** - This measures how well each model predicts unseen data. **Higher is better.**
b) **Residuals** - Measures how well the observed word counts are predicted by the model. **Lower is better**.
c) **Semantic Coherence** - Measure of topic quality, based on the idea that words in the same topic should co-occur in the documents. **Higher is better**.
d) **Lower Bound** - Lower limit for the log likelihood of the observed data given the model parameters. **Higher is better.**



It seemed that **45 topics** would be the optimal number **based on held-out likelihood and lower bound** and it did generate some highly interpretable

This, along with the preprocessing steps earlier helped us improve the quality of topics generated and reducing iterations required for model convergence from 250+ to 43 for the STM model with country contributions and time intervals as covariates covariates

We identified topics published by countries in collaboration with other countries, as well as topics independently published by the country. We also identified prevalent topics during different time periods

We first ran the STM model on the whole dataset, we then ran the model on samples of the dataset based on different countries and then different time intervals. We chose values of K for these sample data based on the output generated by the STM model run on the whole dataset.

8) Network Analysis

We have two CSV files -

1. “INPUT_SQL_Citation_Data_Astronomy_and_Astrophysics.csv”: Contains citation data within the field of Astronomy and Astrophysics.
2. “INPUT_ROR_Metadata.csv”: Contains metadata related to ROR codes, including institution names and countries.

We filter out irrelevant data and normalize the citation counts to ensure proportional representation across the network, and construct a network edge list capturing citations between institutions.

After constructing the edge list, we generated a corresponding graph to carry out a network analysis. Following are some of our observations -

1. The edge list data frame is really huge with 75113657 rows, which gives us a huge sample size of connections between the nodes and edges.
2. There are 102 nodes and 3947 edges with 69 self loops.

Centrality Measures -

3. I did some centrality measures including Degree Centrality, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality and PageRank. It can be viewed in the network analysis pdf file *("Citation Data Analysis with networkx and Node2Vec - Jupyter Notebook.pdf")*.
4. USA, Italy, Germany and the UK have high degree centrality scores (highly connected within the network) whereas Puerto Rico, Lebanon and Nepal have very low scores.
5. USA has the highest closeness centrality score (it can spread information most efficiently within the network) whereas Vietnam, Bangladesh and Ecuador have some of the lowest.
6. USA leads again in betweenness centrality (it acts as a bridge for the flow of information between other countries), interestingly there are many countries with 0 betweenness centrality.
7. Countries like the USA, Germany and the UK again have high Eigenvector Centrality scores (connected to other well connected nodes)
8. The USA has the highest PageRank (this reinforces its status as the most central and influential node in the network).
9. Interestingly, some countries like Cambodia have a relatively high PageRank even though they have lower degree centrality. Maybe because in the context of network structure they are important somehow.

Clustering-

10. I also calculated the clustering coefficient for each country in the graph. The average coefficient is 0.8614, which is quite high which means that the nodes tend to create tightly knit groups with a high density of ties.
11. Countries like Netherlands, USA, Australia and others have relatively lower clustering coefficients (still significant), I suppose it's because these countries have a broad network, and their connections are not restricted to just their neighbours. But I need to dig deeper into this hypothesis to make sure I am not thinking out of my own bias.
12. Countries like Indonesia, Nepal, Jordan, Tunisia and Luxembourg have clustering coefficients of 1, meaning their immediate neighbours are tightly connected to each other and perhaps even insulated with the rest of the network.
13. Other countries like Israel, Serbia and Switzerland have very high clustering coefficients, I suppose they have a strong interconnectivity among their neighbours (strong regional collaborations) but they might also be connected to other nodes in the network.

Connectivity -

Weakly Connected

14. The graph has 1 weakly connected component. So if we ignore the direction of the edges, there is a path between any two nodes in the graph.
15. Size of the largest weakly connected component is 102 (the only weak component). It is also the total number of nodes in the graph, interestingly
16. Nodes that are part of this weakly connected component are listed under 'largest_weakly_connected'

Strongly Connected

17. There are 11 strongly connected components in the graph (every node is reachable from every other node for that component).
18. Largest strongly connected component consists of 92 nodes
19. Nodes that are part of this weakly connected component are listed under 'largest_strongly_connected'

20. I was unable to measure average path length and graph diameter because I got the error: Graph is not strongly connected
22. **Degree Distribution:** The lowest degree is 2, so there are nodes with only two connections, whereas the highest degree is 188. Also, the degree values also vary alot, meaning that the graph is a heterogeneous network; some nodes have much more connections than the others.

25: **Graph Density**: The density of our graph is 0.3831. Which means that there are only 38% of the total number of possible edges.

Node2Vec

26. Finally I also generated Node2Vec embeddings, and saved it as 'node2vec_embeddings.csv'.

27. I used these embeddings to discover that the most similar nodes to the USA are Costa Rica, Palestine, French Polynesia, Bangladesh and Latvia. These results are rather surprising to me.
28. I used KMeans clustering to identify clusters of similar nodes based on these embeddings. I have added them in the 'Cluster' column of the 'node2vec_embeddings.csv' dataframe.
29. I generated a t-SNE plot by reducing the embedding dimensionality to 2. The proximity of the nodes to other nodes suggest how similar their vector representations are.

What are these centralities?

Degree Centrality: Measures the number of direct connections a node has. In citation networks, this can indicate how many different countries a particular country is directly connected with through citations.

Closeness Centrality: Indicates how close a node is to all other nodes in the network. A high closeness centrality means a node can quickly interact with all others, suggesting a central position in the network's communication flow.

Betweenness Centrality: Reflects how often a node appears on the shortest paths between other nodes. A high betweenness centrality suggests that a country plays a significant role as a bridge in the citation network, linking various research clusters.

Eigenvector Centrality: Takes into account not only the number of connections a node has but also the quality of these connections. A country with connections to other highly connected countries will have a higher eigenvector centrality.

PageRank: Developed by Google, it's a way of measuring the importance of nodes in a network. It considers the number and quality of links to a node to estimate how important the node is. In citation networks, this can indicate the influence of a country's research output.

Research Questions -

1. ***Are countries that are prevalent in topics that are widespread in the corpus different from countries that are prevalent in topics that are cited a lot?***
2. ***How does this vary over time?***

1 The US and other few elite, wealthy countries are dominating citations received, topics being researched over time?

Countries like the US and other wealthy nations do dominate in terms of citations received and topics researched (and possibly influence the research topics in astronomy and astrophysics). This is inferred from the centrality measures (degree, closeness, betweenness, eigenvector centrality, and PageRank) calculated for each country. Higher centrality scores for countries like the US indicate a more dominant position in the academic citation network.

This dominance could be attributed to more substantial research funding, better infrastructure, and higher numbers of research institutions in these countries.

The studies by Liu & Rousseau (2019) and Velden, Yan, & Lagoze (2017) highlight how citation networks can reflect power structures and the cognitive structure of disciplines, respectively, underscoring the role of economically powerful countries in shaping research agendas.

Power Structures (Liu & Rousseau, 2019): This study shows that the way research papers cite each other can actually tell us about power dynamics in the academic world. Essentially, it suggests that countries with more economic power and resources tend to have their research papers cited more often. This high rate of citations not only reflects their influence but also their ability to shape what topics or areas of study are considered important in various academic fields.

Cognitive Structure of Disciplines (Velden, Yan, & Lagoze, 2017): This research looks at how the subjects or themes of research papers are interconnected through citations. It suggests that the topics researched and cited in wealthier countries can influence the overall direction and areas of focus within entire academic disciplines. This means that the research interests and priorities of economically powerful countries can shape the overall structure and development of different fields of study.

https://jscires.org/article/282
https://link.springer.com/article/10.1007/s11192-017-2299-9

https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0068-2

2. How are we describing trends over time that support that story?

We have only analyzed the cumulative data so far. In order to discuss trends over time, we would have to examine how the centrality measures and the overall network structure have evolved.

We could do this by looking at changes in the network's density, connectivity, or the emergence of new influential nodes (countries) over time.

The work by Xie et al. (2015) and Ciotti et al. (2015) suggests that citation patterns and network structures evolve in a certain way that potentially reinforces the dominance of already influential countries. This could be due to a combination of historical precedence in the field and ongoing investments in research.

Hence the trend analysis in citation networks shows an evolving dominance of certain countries over time.

https://link.springer.com/article/10.1007/s11192-017-2299-9

3. Is it the case that the US, et al., dominate everything because they produce the most papers, so how are these figures on centrality, topic dominance, etc., actually revealing?

High centrality reveal not just quantity but the influence and integration of the nodes (countries) within the research community.

The dominance of countries like the US in citation networks isn't just due to the volume of papers produced. While higher output can contribute to higher centrality measures, these metrics reveal more nuanced aspects of influence and connectivity. For example, a country with high betweenness centrality might be crucial in connecting different research clusters, indicating a role as a knowledge bridge rather than just a volume leader. Similarly, eigenvector centrality reflects not just quantity but the quality of connections, highlighting influential relationships.

Studies by Hu et al. (2021) and Zhang et al. (2016) imply that centrality in citation networks is not only about the volume of research output but also about the impact and interconnectedness of the research. Wealthy countries tend to have more central positions in these networks, indicating their role in setting research trajectories.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120687

4. Why do countries like Cambodia have such a high PageRank?

The high PageRank of Cambodia in this network is intriguing. Papers from Cambodia, though perhaps fewer in number, are possibly cited by highly influential papers or are central in a smaller, highly interconnected sub-network.

This phenomenon could be analogous to the "small but mighty" concept in network dynamics, where even nodes with fewer connections can have significant impact if those connections are influential.

Hu et al. (2021) discuss the aging effect in evolving scientific citation networks, which could provide insights into how certain nodes gain prominence over time (Hu et al., 2021).

https://link.springer.com/article/10.1007/s11192-021-03929-8

Literature reviews -

1. Souma, W., Vodenska, I., & Chitkushev, L. T. (2020). Classification of Paper Values Based on Citation Rank and PageRank. Journal of Data and Information Science, 5(4), 57-70. https://jscires.org/article/282

Study Objective: Souma, Vodenska, and Chitkushev (2020) aimed to enhance the evaluation of scientific papers beyond traditional citation counts by integrating PageRank, a method originally used by Google for ranking web pages.

Citation Rank vs. PageRank: The study contrasts the conventional citation rank, which counts the number of times a paper is cited, with PageRank, which also considers the importance of the citing papers.

Data Analysis: The authors analyzed a large dataset from the Science Citation Index Expanded (SCIE), covering over 34 million papers and 591 million citations from 1981 to 2015.

Identifying Paper Categories: The research categorized papers into four groups: high-quality, prestige, emerging, and popular, based on their Citation Rank and PageRank scores.

High-Quality Papers: The study found that high-quality papers were predominantly in biochemistry, molecular biology, and chemistry.

Prestige Papers: Prestige papers, identified by their influential citations, were mainly in computer science, engineering, and information science.

Emerging Papers: Emerging papers, which are newer and gaining attention, were noted in biochemistry, molecular biology, and in the journal “Cell.”

Popular Papers: Popular papers, frequently cited but not necessarily impactful in their field, were common in multidisciplinary sciences.

Practical Implications: The findings offer a nuanced approach to evaluating academic papers, useful for forecasting emerging research fields and aiding science policy decisions.

Originality and Value: This study's approach is unique in applying PageRank to a vast citation network, providing a more comprehensive evaluation of academic papers than citation counts alone.

2. Hu, F., Ma, L., Zhan, X., Zhou, Y., Liu, C., Zhao, H., & Zhang, Z.-K. (2021). The aging effect in evolving scientific citation networks. Scientometrics, 126(6), 4297-4309. https://doi.org/10.1007/s11192-021-03929-8

Study Focus: The paper by Hu et al. (2021) explores how scientific papers become less likely to be cited as they age, known as the "aging effect" in citation networks.

New Model: They propose a new model based on hypergraph theory, where one 'hyperedge' can link to several nodes (papers), to better understand how citations evolve over time.

Data Sources: Their analysis is based on large datasets from the American Physical Society (APS) and the Digital Bibliography & Library Project (DBLP), covering diverse disciplines.

Aging Effect Analysis: The study shows that older publications gradually lose their attractiveness, with their likelihood of being cited decreasing exponentially over time.

Simulation and Theoretical Solutions: They developed both theoretical solutions and simulation analyses to validate their model, ensuring it accurately represents real-world citation patterns.

Impact of Early Publications: The research found that early publications in a field tend to have a significant initial impact, but this influence diminishes rapidly as they age.

Collective Citation Patterns: By examining how the entire citation network behaves, the study sheds light on general trends in how scientific research is referenced over time.

Temporal Dynamics: The model captures the temporal dynamics of citations, showing how the relevance of scientific work changes as the field evolves.

Practical Implications: This research can help in understanding how scientific ideas spread and become obsolete, which is crucial for researchers in tracking the evolution of scientific knowledge.

Significance: The study provides a more nuanced view of citation patterns, moving beyond simple counts to understand the deeper dynamics at play in scientific literature

Zotero - citation with google docs plug in

February 2024 -

Updated dataset - The dataset has 540,515 research abstracts, but out of those we have got 247,117 duplicates (these duplicates might still have unique concept_id and work_id, but duplicate year, title, abstract and country).

Preprocessed Data -

Duplicates
After removing duplicate research abstracts and keeping just the first instance of the duplicates, we get a cleaner dataset of 416,953 research abstracts

540,515 - 416,953 = 123,562 dropped out abstracts
But 247,117/2 = 123,558 research abstracts meaning there might have been a few instances of more than just 2 duplicates

Null/NaN values
There were 41 rows with null/NaN values (each of those had a null value in either the title, abstract, or country columns)
We dropped all of these rows and are now down to **416,912** rows

This filtered dataset is saved as 'filter_data_feb_2024.csv'.

How did we bin countries -

We gave individual contributions to the top 12 countries because they have contributions of more than 2%.
We binned the rest of the countries into regions, such that in total we have got 18 countries and regions.

Countries Analysis -

1. Certain countries (like India and UK) are lower in the number of publication rankings, however they are higher in contribution ranking. This tells me that these countries might be collaborating with other countries to a lesser degree, and instead concentrate more of their contributions in fewer publications.

For DIM
Define how document influence will be measured within the dataset. Modify or extend existing topic modeling approaches to incorporate influence metrics, possibly involving network analysis or citation analysis if the influence is defined by inter-document references. Implement algorithms that can track and update influence scores over time or across topics.

STM (time covariates) generated topics -

N.B. - My initial thoughts - would it be challenging to decide on topic names that would be good enough to generalize and encapsulate all the abstracts that are labeled as that particular topic?

```
Topic 1 Top Words:
        Highest Prob: evalu, assess, inform, tool, part, manag, standard, practic,
systemat, refer, lower, updat, softwar, industri, famili
        FREX: evalu, assess, inform, tool, part, standard, systemat, refer, lower,
updat, industri, collect, databas, requir, reliabl
        Lift: collect, -site, accelerometer, assess, automobil, databas, daynight,
definit, diagon, ensur, ephemerid, fermilab, fisheri, geoloc, hrg
        Score: evalu, assess, inform, practic, manag, tool, part, standard, systemat,
softwar, refer, industri, updat, onlin, servic
Topic 2 Top Words:
        Highest Prob: new, test, first, experi, calibr, earli, initi, year, recent,
absolut, overview, obtain, descript, occult, methodolog
        FREX: test, first, calibr, earli, descript, occult, -orbit, campaign,
radiometr, recent, parallax, -flight, seven, astrometri, spectrophotometr
        Lift: -orbit, earli, first, parallax, sim, test, vicari, -flight, aqua,
glass-jwst, matiss, microarcsecond, midi, multi-sit, occult
        Score: new, first, test, calibr, experi, earli, initi, year, recent, overview,
-orbit, -ski, occult, absolut, radiometr
Topic 3 Top Words:
        Highest Prob: survey, spectroscop, quasar, select, redshift, deep, agn, sky,
photometr, digit, large-scal, hubbl, catalog, plane, izi
        FREX: survey, quasar, redshift, sky, photometr, digit, catalog, plane,
continuum, sdss, catalogu, releas, faint, sloan, count
        Lift: cmass, mathsf, point-spread, single-epoch, spt-sz, arecibo, continuum,
emission-lin, for, isophot, lclcyclcα, mock, qsos, radio-quiet, sdss
        Score: survey, quasar, redshift, spectroscop, agn, photometr, sky, izi,
catalog, deep, select, digit, sloan, sdss, hubbl
Topic 4 Top Words:
        Highest Prob: even, kepler, end, far, barrier, similar, morn, illumin,
african, distinct, iri, quadrupol, ini, ici, tradit
        FREX: kepler, end, far, barrier, similar, morn, illumin, african, distinct,
iri, ini, ici, tradit, occup, ancient
        Lift: canaria, custom, egyptian, end, far, freeform, ihi, inhabit, isi,
poetri, processor, recept, renewcommandencodingdefaultot, violenc, accept
        Score: even, kepler, morn, far, end, similar, barrier, african, illumin, iri,
distinct, newcommandcyr, renewcommandrmdefaultwncyr, usepackageamsbsi,
usepackageamsfont
Topic 5 Top Words:
        Highest Prob: properti, applic, sphere, character, synthesi, via, effici,
enhanc, materi, glass, hybrid, nanoparticl, prepar, hierarch, assembl
        FREX: properti, applic, sphere, character, synthesi, effici, enhanc, materi,
nanoparticl, hollow, colloid, phosphor, uniform, silica, modif
        Lift: alumina, batteri, close-pack, core-shel, dielectrophoresi, double-shel,
electrod, encapsul, enzymat, ersupsup, nanocomposit, photodegrad, pore, powder, ser
        Score: properti, sphere, applic, synthesi, character, enhanc, effici, materi,
via, glass, hollow, nanoparticl, phosphor, prepar, colloid
Topic 6 Top Words:
```

Highest Prob: model, simul, radiat, interact, wind, particl, flow, turbul, transfer, flux, acceler, numer, heat, charg, instabl
FREX: simul, interact, particl, turbul, numer, instabl, spheric, driven, magnetohydrodynam, compress, equilibrium, dynamo, drift, tube, mhd
Lift: magnetohydrodynam, alfvén, astron, buoyant, collisionless, compression, corioli, hard-spher, kelvin-helmholtz, large-eddi, long-tim, mhd, microswimm, multifluid, non-id
Score: model, simul, particl, radiat, wind, interact, turbul, instabl, flux, numer, acceler, charg, transfer, flow, hydrodynam
Topic 7 Top Words:
Highest Prob: star, format, rotat, neutron, giant, abund, chemic, red, rate, element, cool, rapid, pulsat, histori, low-mass
FREX: star, format, neutron, abund, red, cool, rapid, branch, evolv, tauri, spot, agb, subdwarf, asteroseismolog, quark
Lift: cno, evolv, kic, neutron-captur, wolf-rayet, abund, binar, bursti, byikepleri, c-rich, carbon-enhanc, deconfin, double-mod, gravity-mod, hamburgeso
Score: star, format, rotat, neutron, giant, abund, red, pulsat, cool, chemic, low-mass, element, rapid, branch, tauri
Topic 8 Top Words:
Highest Prob: sight, magic, mathemat, higher, western, atlas, translat, hidden, centuri, clock, phenomenon, art, hand, margin, plain
FREX: sight, magic, mathemat, higher, atlas, translat, hidden, centuri, clock, phenomenon, art, hand, margin, plain, vega
Lift: alchemi, barcelona, decod, dinosaur, fem, garcilaso, ino, lope, pau, plankton, renaiss, spectrometr, theatr, vertebr, watt
Score: sight, mathemat, magic, western, centuri, higher, atlas, hidden, clock, translat, art, del, vega, plain, hand
Topic 9 Top Words:
Highest Prob: implic, magnitud, earthquak, area, sequenc, ground, site, seismic, record, locat, deform, import, zone, northern, attenu
FREX: implic, magnitud, earthquak, area, sequenc, ground, site, seismic, record, locat, deform, northern, attenu, damag, exampl
Lift: accretionari, alpin, amplif, andes, bedrock, colorado, cosmogen, craton, damag, drop, exampl, fidel, flank, gulf, magma
Score: earthquak, implic, magnitud, seismic, area, ground, site, sequenc, record, deform, locat, northern, basin, import, attenu
Topic 10 Top Words:
Highest Prob: dark, wave, matter, gravit, cosmolog, probe, constraint, power, theori, graviti, lens, problem, strong, constrain, equat
FREX: dark, wave, matter, gravit, cosmolog, probe, constraint, theori, graviti, lens, constrain, weak, modifi, reioniz, len
Lift: acoustic-grav, bardeen, dark-energi, infin, -cm, bang, bianchi, bose-einstein, boson, brans-dick, cdm, chameleon, chaplygin, cusp, dilaton
Score: dark, matter, gravit, wave, cosmolog, lens, constraint, graviti, theori, probe, power, strong, constrain, halo, weak
Topic 11 Top Words:
Highest Prob: system, optic, telescop, design, perform, integr, instrument, array, sensor, spectrograph, oper, camera, astronom, larg, adapt
FREX: telescop, instrument, array, spectrograph, camera, astronom, mirror, interferomet, coronagraph, guid, apertur, prototyp, keck, cryogen, capabl
Lift: aberr, anisoplanat, apertur, arraytitl, keck, linc-nirvana, opticstitl, slump, space-bas, titleadapt, titlean, titleintegr, wht, x-ifu, applicationstitl
Score: system, optic, telescop, design, spectrograph, instrument, array, perform, sensor, integr, camera, oper, mirror, interferomet, astronom
Topic 12 Top Words:

Highest Prob: analysi, process, statist, featur, revisit, interpret, boundari, code, separ, one, main, probabl, analyz, term, signific
FREX: analysi, process, statist, featur, revisit, interpret, boundari, code, separ, one, probabl, analyz, term, various, comment
Lift: analyz, full, separ, addendum, featur, histogram, introduct, low-level, multifract, probabl, show, term, triton, various, analysi
Score: analysi, process, statist, featur, revisit, code, interpret, boundari, separ, one, issu, main, analyz, probabl, special
Topic 13 Top Words:
Highest Prob: primari, case, respons, human, cell, patient, factor, secondari, vaccin, treatment, booster, report, clinic, diseas, covid-
FREX: primari, human, cell, patient, vaccin, treatment, booster, report, clinic, diseas, dose, cancer, express, gene, sever
Lift: cancer, cardiovascular, chemotherapi, cortic, diagnosi, dose-adjust, inhal, intestin, kidney, malign, needl, rat, spinal, therapeut, thrombosi
Score: primari, patient, vaccin, human, case, cell, respons, booster, diseas, treatment, covid-, clinic, factor, cancer, dose
Topic 14 Top Words:
Highest Prob: galact, sourc, line, variabl, cloud, larg, sampl, small, extrem, discoveri, iii, nuclei, bright, extend, magellan
FREX: galact, line, small, extrem, nuclei, bright, magellan, center, extinct, cepheid, broad, excess, centr, proper, sourc
Lift: period-luminos, absorption-lin, aeb, alpha, balmer, broad, center, centimet, cepheid, excess, exomol, extrem, high-latitud, low-luminos, mbm
Score: galact, cloud, line, sourc, variabl, magellan, nuclei, discoveri, sampl, larg, bright, iii, extrem, small, lumin
Topic 15 Top Words:
Highest Prob: natur, point, view, versus, challeng, depth, late, stage, discrimin, saturn, histor, independ, final, spread, launch
FREX: natur, point, view, versus, challeng, depth, late, stage, discrimin, saturn, histor, independ, final, spread, launch
Lift: audibl, brazil, choic, discrimin, equatorial-depth, janeiro, late, minima, necess, offer, reusabl, saturn, tongu, yes, attempt
Score: natur, view, point, versus, challeng, depth, stage, late, discrimin, saturn, final, histor, independ, launch, spread
Topic 16 Top Words:
Highest Prob: physic, result, product, astrophys, collis, program, laboratori, decay, lunar, preliminari, mont, carlo, heavi, demonstr, scintil
FREX: physic, result, product, astrophys, collis, laboratori, decay, lunar, preliminari, mont, carlo, heavi, demonstr, scintil, relev
Lift: apollo, authent, budget, calorimet, carlo, cern, constanc, contamin, counter, cuor, free-fal, heavy-ion, hertz, high-luminos, hyperveloc
Score: result, physic, product, astrophys, collis, laboratori, mont, carlo, program, lunar, preliminari, decay, heavi, demonstr, scintil
Topic 17 Top Words:
Highest Prob: determin, paramet, band, posit, precis, angl, loss, accuraci, navig, gap, accur, fit, edg, gps, vision
FREX: determin, paramet, band, precis, angl, loss, accuraci, navig, gap, accur, fit, edg, gps, vision, bridg
Lift: circuit, josephson, programm, seamless, simpli, u-shap, afm, ambigu, bds, bds-, cut, decorrel, determin, fbg, fill
Score: determin, paramet, band, posit, precis, accuraci, navig, angl, gap, loss, accur, edg, fit, vision, displac
Topic 18 Top Words:

Highest Prob: space, satellit, environ, earth, global, explor, mission, intern, communic, channel, life, spacecraft, mar, link, constel

FREX: space, satellit, environ, earth, global, explor, mission, communic, channel, spacecraft, mar, link, constel, station, weather

Lift: station, aerothermodynam, aviat, crew, cubesat, downlink, dual-spin, enceladus, geoscienc, hail, inter-satellit, intersatellit, iss, juic, langley

Score: space, satellit, mission, earth, environ, explor, intern, communic, global, constel, mar, life, spacecraft, station, link

Topic 19 Top Words:

Highest Prob: region, molecular, object, map, toward, planetari, core, nebula, compact, outflow, reveal, classif, central, dens, alma

FREX: object, map, toward, core, nebula, compact, outflow, reveal, dens, alma, maser, orion, ira, submillimet, clump

Lift: afgl, filamentari, fountain, iji, ophiuchus, orion, protostar, starless, star-form, yso, map, alma, alma-pil, australi, barnard

Score: region, nebula, object, molecular, planetari, core, toward, outflow, map, compact, alma, maser, star-form, orion, dens

Topic 20 Top Words:

Highest Prob: emiss, radio, burst, gamma-ray, jet, origin, fast, possibl, associ, relativist, afterglow, long, grb, connect, short

FREX: emiss, radio, burst, gamma-ray, jet, origin, possibl, afterglow, long, grb, short, blazar, high-energi, multiwavelength, synchrotron

Lift: compton, intraday, markarian, parsec-scal, pks, radio, baselin, bats, bepposax, blueshift, chime, chimefrb, cocoon, comptel, curtain

Score: emiss, radio, burst, gamma-ray, jet, origin, fast, grb, afterglow, relativist, blazar, possibl, gamma-ray, long, associ

Topic 21 Top Words:

Highest Prob: light, supernova, type, neutrino, shock, curv, distanc, remnant, signatur, shell, explos, expans, peculiar, progenitor, asymmetr

FREX: light, supernova, type, neutrino, shock, curv, distanc, remnant, signatur, shell, explos, expans, peculiar, progenitor, bubbl

Lift: explod, peculiar, blast, cassiopeia, core-collaps, core-collaps, ctb, deflagr, echo, electron-captur, expans, fail, fast-mov, flavor, flavour

Score: supernova, light, type, neutrino, remnant, shock, curv, distanc, signatur, progenitor, explos, shell, core-collaps, peculiar, nucleosynthesi

Topic 22 Top Words:

Highest Prob: crystal, depend, singl, film, novel, growth, layer, quantum, thin, luminesc, interfac, silicon, photoluminesc, thick, well

FREX: singl, film, growth, layer, quantum, thin, interfac, silicon, photoluminesc, thick, well, subsub, bulk, strain, grown

Lift: cubic, disloc, microcav, multilay, phonon, photoreflect, piezoelectr, sic, subsub, -layer, -optic, acceptor, aggregation-induc, algaa, algan

Score: crystal, film, singl, quantum, thin, layer, depend, photoluminesc, luminesc, novel, growth, dope, silicon, grown, epitaxi

Topic 23 Top Words:

Highest Prob: ratio, contribut, analys, supsup, negat, aspect, quantif, step, der, synthes, newli, und, axi, von, lung

FREX: ratio, contribut, analys, supsup, negat, aspect, quantif, step, der, synthes, newli, und, axi, von, van

Lift: den, der, diagnos, dosimet, endogen, lineshap, ortho--para, quantif, seoul, suitabl, telemetri, van, von, xxix, allen

Score: ratio, contribut, supsup, analys, der, negat, und, aspect, quantif, von, step, synthes, van, waal, lung

Topic 24 Top Words:

Highest Prob: studi, investig, experiment, comput, compar, combin, theoret, defect, alloy, detail, comprehens, positron, feasibl, epr, photoemiss
FREX: studi, investig, experiment, comput, combin, theoret, comprehens, positron, feasibl, mössbauer, steel, iab, initioi, nqr, exaf
Lift: endor, exaf, feasibl, gelatin, monoenerget, studi, xaf, xmcd, cao, comprehens, comput, cosi, electron-irradi, fes, fmri
Score: studi, investig, experiment, theoret, compar, comput, combin, defect, alloy, positron, epr, photoemiss, esr, feasibl, mössbauer
Topic 25 Top Words:
Highest Prob: two, candid, three, tidal, photon, indic, second, identifi, disrupt, arteri, dual, confirm, screen, planar, follow-
FREX: two, candid, three, tidal, photon, indic, second, identifi, disrupt, dual, confirm, screen, planar, follow-, dimension
Lift: adaptive-opt, biophys, chiron, complementar, disrupt, dual, fals, fetus, piv, radio-emit, three-bodi, weigh, abnorm, abrupt, alarm
Score: two, candid, tidal, three, photon, indic, arteri, identifi, disrupt, second, dual, confirm, velocimetri, screen, planar
Topic 26 Top Words:
Highest Prob: leagu, perform, footbal, nation, sport, support, valu, player, major, movement, profession, train, evid, european, match
FREX: leagu, footbal, nation, sport, valu, player, major, profession, train, european, match, soccer, examin, team, injuri
Lift: bargain, cyclist, diplomaci, ncaa, passion, trainer, ventur, academi, australian, basketbal, brand, career, club, commod, decis
Score: leagu, footbal, player, sport, profession, nation, soccer, team, basebal, perform, injuri, major, game, valu, train
Topic 27 Top Words:
Highest Prob: scienc, review, social, perspect, unit, open, state, media, side, critic, law, cultur, world, empir, histori
FREX: scienc, review, perspect, unit, media, critic, law, cultur, world, empir, look, access, scientif, modern, polici
Lift: ahead, anniversari, argentin, bigger, citizen, citizenship, crimin, critic, crowd, darwin, disciplin, encourag, existenti, fact, good
Score: scienc, review, social, perspect, unit, cultur, world, open, polit, side, public, media, polici, communiti, american
Topic 28 Top Words:
Highest Prob: impact, event, level, meter, qualiti, air, china, horizon, build, delta, forest, urban, soil, oil, citi
FREX: impact, level, meter, horizon, build, forest, urban, soil, citi, river, recoveri, smart, sustain, pollut, mega
Lift: estuari, soil, agricultur, arid, beef, beij, bit, build, canopi, cattl, citi, clay, destin, egypt, farm
Score: impact, event, level, meter, qualiti, air, china, urban, delta, horizon, soil, citi, build, river, forest
Topic 29 Top Words:
Highest Prob: effect, influenc, differ, thermal, characterist, pressur, stabil, bear, condit, composit, size, behavior, bodi, due, speed
FREX: effect, thermal, stabil, bear, condit, load, static, friction, lay, lubric, hen, rotor, tilt, axial, thrust
Lift: bed, broiler, egg, elastohydrodynam, fluidiz, herringbon, hydrostat, micropolar, orthotrop, rotor-bear, self-act, sine, spindl, squeeze-film, textur
Score: effect, bear, influenc, pressur, thermal, characterist, differ, stabil, behavior, composit, condit, size, journal, bodi, load
Topic 30 Top Words:

Highest Prob: dwarf, disk, orbit, search, planet, around, disc, transit, white, hot, limit, exoplanet, inner, companion, protoplanetari
FREX: disk, planet, around, white, limit, exoplanet, inner, companion, protoplanetari, jupit, brown, debri, terrestri, radius, extrasolar
Lift: circumbinari, circumplanetari, jovian, kepler-, tess, ultracool, unseen, brown, cancri, cha, co-orbit, engulf, exocomet, exoplanetari, exozodiac
Score: dwarf, disk, planet, orbit, around, search, disc, white, hot, exoplanet, transit, protoplanetari, companion, jupit, limit
Topic 31 Top Words:
Highest Prob: use, base, method, control, trajectori, estim, predict, approach, optim, network, improv, track, motion, algorithm, generat
FREX: base, control, trajectori, predict, approach, optim, network, track, robot, vehicl, plan, neural, mobil, machin, manipul
Lift: ballist, crane, cruis, exoskeleton, gait, glide, humanoid, iter, kalman, network-bas, pedestrian, recurs, reentri, time-vari, wireless
Score: trajectori, control, base, method, robot, optim, track, vehicl, use, network, estim, algorithm, predict, learn, plan
Topic 32 Top Words:
Highest Prob: x-ray, binari, hole, black, accret, pulsar, period, transient, eclips, x-ray, merger, hard, soft, nova, close
FREX: x-ray, binari, hole, black, accret, pulsar, period, transient, eclips, soft, nova, close, outburst, supermass, millisecond
Lift: algol, aql, coalesc, cygni, double-lin, herculi, nicer, psr, sagitta, vulpecula, accret, pulsar, accretor, advect, advection-domin
Score: x-ray, binari, hole, black, accret, pulsar, x-ray, merger, eclips, period, transient, supermass, nova, hard, outburst
Topic 33 Top Words:
Highest Prob: data, develop, monitor, research, detector, project, compon, current, futur, technolog, nois, status, valid, advanc, construct
FREX: data, develop, monitor, detector, project, current, futur, nois, status, valid, advanc, construct, progress, anomali, sar
Lift: cross-calibr, intercalibr, low-background, sentinel-, advanc, amazon, ambient, anomali, bug, c-band, confer, current, detector, develop, eo-
Score: data, develop, research, monitor, detector, project, current, futur, status, technolog, nois, advanc, compon, valid, construct
Topic 34 Top Words:
Highest Prob: univers, shape, morpholog, book, south, speci, press, geometr, along, paper, island, new, cloth, india, york
FREX: univers, shape, morpholog, south, speci, press, geometr, along, paper, cloth, york, indian, africa, edit, fli
Lift: anura, call, chicago, cloth, colo, contour, cornel, cranial, craniofaci, dart, descriptor, diptera, geometr, guinea, harvard
Score: univers, book, shape, press, morpholog, south, cloth, speci, paper, geometr, york, along, island, cambridg, india
Topic 35 Top Words:
Highest Prob: spectra, electron, absorpt, state, ion, vibrat, excit, raman, atom, carbon, transit, hydrogen, reaction, molecul, calcul
FREX: spectra, electron, absorpt, ion, excit, raman, atom, hydrogen, reaction, oxid, solid, oxygen, amin, radic, bond
Lift: magnesium, aceton, acetyl, acetylen, adduct, adenin, aldehyd, amid, anharmon, anhydr, annul, antimoni, benzen, bimolecular, bismuth
Score: spectra, electron, absorpt, vibrat, ion, excit, raman, atom, hydrogen, oxid, molecul, fluoresc, carbon, reaction, state
Topic 36 Top Words:

Highest Prob: observ, solar, activ, flare, coron, sun, cycl, comet, asteroid, corona, infer, loop, eject, ionospher, fragment

FREX: observ, solar, flare, coron, sun, cycl, comet, asteroid, corona, loop, eject, ionospher, nucleus, chromospher, sunspot

Lift: cme, cmes, eit, extreme-ultraviolet, field-align, ground-bas, heliosheath, nanoflar, nucleus, pchuryumov-gerasimenko, sdo, sep, spectrophotometri, sumer, trail

Score: observ, solar, activ, flare, coron, comet, sun, asteroid, corona, cycl, eject, loop, chromospher, nucleus, erupt

Topic 37 Top Words:

Highest Prob: variat, chang, spatial, pattern, rate, within, tempor, across, total, individu, long-term, environment, behaviour, genet, climat

FREX: variat, chang, pattern, within, tempor, across, environment, behaviour, genet, climat, caus, biolog, common, trend, explain

Lift: charact, common, drosophila, intraspecif, ancestri, ape, bark, bird, captiv, chromosom, climat, consequ, cryptic, deer, ecolog

Score: variat, chang, spatial, pattern, genet, tempor, across, individu, within, season, climat, environment, total, behaviour, rate

Topic 38 Top Words:

Highest Prob: fluctuat, primordi, set, presenc, warm, without, cmb, inflat, scenario, maximum, asymmetri, consist, topolog, remov, minim

FREX: fluctuat, primordi, set, presenc, warm, without, cmb, inflat, scenario, maximum, asymmetri, consist, topolog, remov, minim

Lift: scenario, b-mode, curvaton, d-brane, fine-tun, five-dimension, ghost, gibb, halo-independ, impos, isotropi, likelihood, litebird, maximum, minim

Score: primordi, fluctuat, inflat, cmb, set, warm, scenario, without, presenc, maximum, non-gaussian, topolog, asymmetri, consist, planck

Topic 39 Top Words:

Highest Prob: galaxi, distribut, function, gas, relat, local, dust, densiti, scale, group, luminos, nearbi, spiral, host, angular

FREX: distribut, gas, local, luminos, spiral, fraction, seyfert, starburst, bar, absorb, irregular, intergalact, narrow-lin, densiti, ism

Lift: atizilt, floccul, sfr, illustristng, intragroup, irregular, ism, narrow-lin, vec, distribut, local, arp, atlasd, bulgeless, cartwheel

Score: galaxi, gas, distribut, dust, function, relat, local, densiti, luminos, spiral, nearbi, group, scale, seyfert, host

Topic 40 Top Words:

Highest Prob: spectroscopi, imag, high, spectral, infrar, resolut, ultraviolet, near, reflect, spectromet, high-resolut, near-infrar, transform, transmiss, visibl

FREX: spectral, resolut, ultraviolet, spectromet, high-resolut, transform, transmiss, visibl, wavelength, fourier, synthet, broadband, microscop, breakdown, deconvolut

Lift: synthet, acousto-opt, aotf, breakdown, cavity-enhanc, deconvolut, deconvolv, diffract, dual-comb, excitation-emiss, fabry-perot, fourier, high-spectr, high-throughput, imagestitl

Score: spectroscopi, imag, infrar, high, spectral, resolut, spectromet, ultraviolet, high-resolut, reflect, near-infrar, near, transform, fourier, visibl

Topic 41 Top Words:

Highest Prob: measur, use, doppler, laser, plasma, veloc, techniqu, signal, puls, flow, intens, profil, beam, modul, rang

FREX: measur, laser, techniqu, puls, modul, rang, radar, diagnost, discharg, coher, devic, ultrasound, tomographi, volum, ultrason

Lift: ablat, coupler, laser-doppl, titlelas, -fiber, -nm, -vivo, arterioven, atmospheretitl, atmospheric-pressur, bandwidth, bistat, c-mod, cathet, coher

Score: measur, doppler, laser, plasma, use, veloc, puls, techniqu, signal, beam, radar, ultrasound, flow, discharg, rang
Topic 42 Top Words:
Highest Prob: general, deriv, differenti, ring, form, number, simpl, geometri, normal, coeffici, free, beyond, random, approxim, reduc
FREX: general, deriv, ring, form, number, simpl, geometri, normal, coeffici, free, beyond, approxim, partial, stabl, altern
Lift: altern, arbitrari, axe, coeffici, compat, converg, exterior, fisher, formul, formula, general, log, normal, number, parabol
Score: general, deriv, ring, form, number, differenti, coeffici, simpl, geometri, approxim, normal, free, random, beyond, reduc
Topic 43 Top Words:
Highest Prob: cluster, mass, evolut, stellar, ngc, massiv, evid, popul, young, globular, kinemat, halo, photometri, metal, way
FREX: cluster, mass, stellar, ngc, globular, photometri, way, milki, stream, bulg, abel, gaia, merg, virgo, igaiai
Lift: edr, praesep, stream, apoge, chemo-dynam, collind, color-magnitud, color–magnitud, color–magnitud, colour–magnitud, dsph, evolut, galactocentr, galaxy', geminigmo
Score: cluster, stellar, ngc, mass, evolut, globular, massiv, milki, young, popul, halo, way, kinemat, photometri, open
Topic 44 Top Words:
Highest Prob: surfac, temperatur, atmospher, water, interstellar, diffus, medium, organ, correct, sens, ice, remot, concentr, irradi, ocean
FREX: surfac, water, interstellar, sens, ice, remot, concentr, irradi, ocean, sea, retriev, aerosol, vertic, grain, stratospher
Lift: ecmwf, ssmi, submesoscal, -situ, aeronet, airfoil, altimetri, analogu, bathymetri, cdom, chlorophyll-, clear-ski, cmip, cyclon, downward
Score: surfac, temperatur, atmospher, water, interstellar, medium, sens, ice, diffus, remot, organ, retriev, sea, ocean, aerosol
Topic 45 Top Words:
Highest Prob: time, can, mean, index, delay, seri, north, day, green, note, refract, east, middl, travel, veget
FREX: time, can, mean, index, delay, seri, north, day, green, note, refract, east, travel, veget, burn
Lift: acuiti, airlin, alga, bath, broken, cautionari, closer, conting, day, deeper, delay, dictat, discern, dive, doa
Score: time, can, mean, index, delay, seri, day, north, green, note, travel, east, middl, knight, refract
Topic 46 Top Words:
Highest Prob: magnet, field, reson, frequenc, low, scatter, acoust, coupl, nuclear, mode, induc, oscil, electr, doubl, circular
FREX: magnet, low, scatter, doubl, circular, cross, low-frequ, dichroism, section, reson, frequenc, tunnel, dipol, harmon, high-frequ
Lift: elf, ellips, harmon, kdp, lossi, low-frequ, magnetometri, optical–opt, schumann, textor, low, apparatus, caviti, circular, cross
Score: magnet, field, reson, scatter, frequenc, low, acoust, nuclear, coupl, circular, electr, oscil, mode, dichroism, proton
Topic 47 Top Words:
Highest Prob: energi, cosmic, spectrum, ray, background, observatori, microwav, propag, anisotropi, gamma, trace, anomal, cosmic-ray, virtual, extens
FREX: energi, cosmic, ray, background, microwav, propag, anisotropi, gamma, trace, anomal, cosmic-ray, virtual, extens, mev, web
Lift: meteorolog, aci, altitud, anisotropi, anomal, antiproton, cabl, censorship, cerenkov, cosmic, cosmic-ray, cta, harden, highest, iwilkinson

```
        Score: cosmic, energi, ray, background, spectrum, microwav, observatori,
anisotropi, gamma, propag, cosmic-ray, anomal, trace, auger, virtual
Topic 48 Top Words:
        Highest Prob: structur, dynam, polar, phase, complex, chain, solut, liquid,
nmr, molecular, polym, linear, mix, order, protein
        FREX: phase, chain, liquid, polym, order, mixtur, conform, solvent, contain,
chiral, ionic, aggreg, ligand, block, solid-st
        Lift: coarse-grain, contain, electrolyt, fold, hydrophob, melt, mesomorph,
micell, short-chain, unfold, -atom, achir, acryl, amphiphil, andc
        Score: structur, dynam, polar, chain, nmr, complex, liquid, phase, solut,
polym, molecular, protein, mixtur, dielectr, relax
Topic 49 Top Words:
        Highest Prob: role, relationship, among, cours, visual, associ, educ,
understand, school, children, student, languag, health, work, exposur
        FREX: relationship, cours, visual, understand, school, student, languag, work,
sleep, teach, speech, person, memori, daili, represent
        Lift: addict, adolesc, allophon, arithmet, attent, awar, bing, caffein,
cardiometabol, chronobiolog, coarticul, cochlear, cognit, conson, corpus-bas
        Score: cours, children, school, student, among, languag, educ, sleep, care,
health, role, adult, relationship, visual, work
Topic 50 Top Words:
        Highest Prob: detect, comparison, correl, direct, mechan, potenti, multipl,
sensit, insight, use, evid, identif, generat, motion, infer
        FREX: direct, detect, mechan, potenti, sensit, correl, comparison, insight,
multipl, dragon, hoc, pick-, thermochem, prismat, peek
        Lift: sensit, direct, potenti, mechan, detect, insight, correl, comparison,
multipl, use, evid, identif, generat, infer, motion
        Score: detect, direct, comparison, mechan, correl, potenti, sensit, multipl,
insight, use, evid, identif, generat, motion, infer
```

ChatGPT generated topic names -

Research Evaluation and Tools
Astronomical Instrumentation and Calibration
Cosmological Surveys and Redshift Studies
Kepler Mission and Exoplanet Discoveries
Material Properties and Nanotechnology
Astrophysical Simulations and Models
Stellar Evolution and Neutron Stars
Historical Astronomy and Celestial Phenomena
Seismology and Earthquake Studies
Dark Matter and Cosmological Theories
Telescope and Astronomical Instrument Design
Data Analysis and Statistical Methods
Immunology and Vaccine Research
Galactic Nuclei and Active Galactic Nuclei (AGN)
Philosophical Perspectives on Nature and Science
Particle Physics and Astrophysics
Precision Measurement and Navigation Technologies

Space Missions and Satellite Technology
Interstellar Medium and Molecular Clouds
High-Energy Astrophysics and Gamma-Ray Bursts
Supernovae and Neutrino Astronomy
Crystal Growth and Semiconductor Physics
Chemical Analysis and Synthesis
Experimental and Theoretical Studies in Physics
Binary Star Systems and Tidal Effects
Sports Performance and Analytics
Science Communication and Social Perspectives
Environmental Impact and Urban Development
Mechanical Engineering and Material Science
Exoplanet Searches and Disk Structures
Control Systems and Robotics
X-Ray Astronomy and Black Hole Studies
Data Acquisition and Sensor Technology
Cultural and Geometric Studies
Spectroscopy and Molecular Physics
Solar Physics and Space Weather
Genetic Variation and Environmental Adaptation
Cosmological Inflation and Primordial Fluctuations
Galaxy Evolution and Gas Dynamics
High-Resolution Spectroscopy and Imaging
Measurement Techniques and Laser Technology
Mathematical Formulations and General Theories
Stellar Clusters and Galactic Structure
Atmospheric Studies and Earth Sciences
Time Series Analysis and Geographical Studies
Magnetic Resonance and Electromagnetic Theory
High-Energy Cosmic Rays and Gamma-Ray Astronomy
Molecular Dynamics and Phase Transitions
Education and Cognitive Development
Detection Techniques and Sensory Systems

STM (CxT covariates) generated topics -

present proof of work permission
business manager in department
not much paper work

The two CSV/dataframes we need:

CSV/Dataframe 1 | Summarized STM Topic Data

- [KEY] The name of the field (in this case, "astronomy/astrophysics").
- [KEY] STM time period (e.g., 1990-1994, 1995-2000, etc.)
- the topic number from the STM (e.g., 1, 2, 3, 4...etc.)
- the topic label from ChatGPT (e.g., "galactic nuclei," "experimental methods," etc.
- The raw number of citations the topic received in the five-year window as Sum and Percentage
- Citation deflated (sum of citation deflated) as Sum for one column and Percentages as another column.
- The prevalence of that topic in the corpus from the STM as Sum for one column and Percentages as another column.

CSV/Dataframe 2 | Summarized Country Covariates from the STM Model

- [KEY] The name of the field (in this case, "astronomy/astrophysics").
- [KEY] STM time period
- The country covariate column, where each row is a country covariate (e.g., "United States," "China," "South America," etc.)
- The number of citations received by the country covariate (as rows) in the five-year window as a Sum for one column and as a Percentage as another column.
- The citation deflated number of citations received by the country covariate in the five-year window as a Sum for one column and a Percentage as another column.
- The number of papers with at least one author from that country in that time period.
- Beta coefficient for that country from the STM
- The standard errors of that estimate of that country covariate from the STM.
- The t-value for the beta coefficient for that country from the STM.