NASA's Geospatial Metadata

Description

When considering datasets maintained by NASA, one might assume that most of this data involves measurements or observations from space. Surprisingly, most of NASA's datasets are actually geospatial measurements of Earth! For this analysis, I'm curious to learn more about the types of data the agency collects about our own planet. I aim to gain a more comprehensive understanding of how NASA contributes to science through the diverse range of data it gathers. This analysis employs simple exploration, word co-occurrences/correlations, TF-IDF, and topic modeling to understand relationships among NASA's data catalog.

Conclusions

Title and Description Word Pairs

Figures 1 and 2 illuminate the interconnected nature of words within NASA's data catalog, specifically within titles and descriptions. Figure 1, focusing on title word pairs with frequent co-occurrence, highlights "data" as a central hub, with strong connections to terms like "global," "release," and "comet."

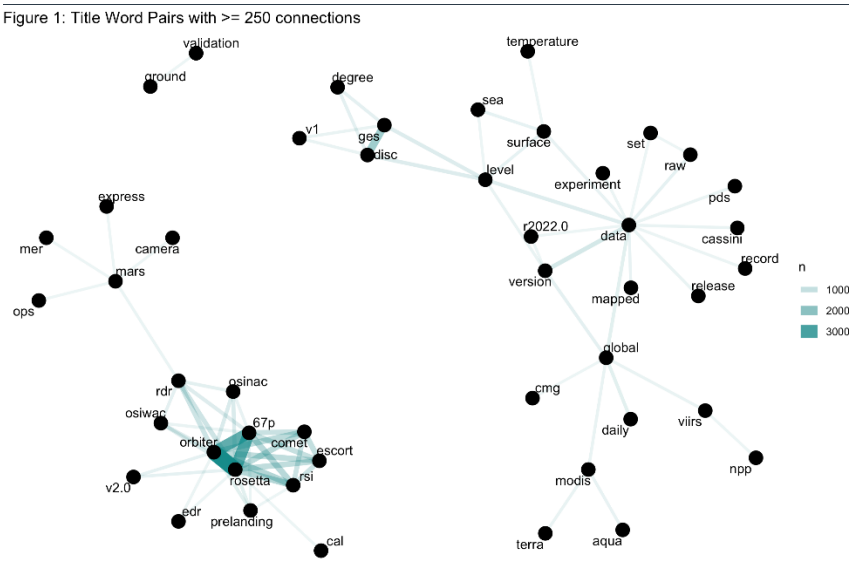

Figure 1: Title word pairs with >= 250 connections

Figure 2, which analyzes co-occurring word pairs within the more detailed descriptions, again places "data" at the forefront, revealing strong links to "spacecraft," "measurements," and "temperature." This reinforces the importance of observational data, particularly related to spacecraft measurements and temperature.

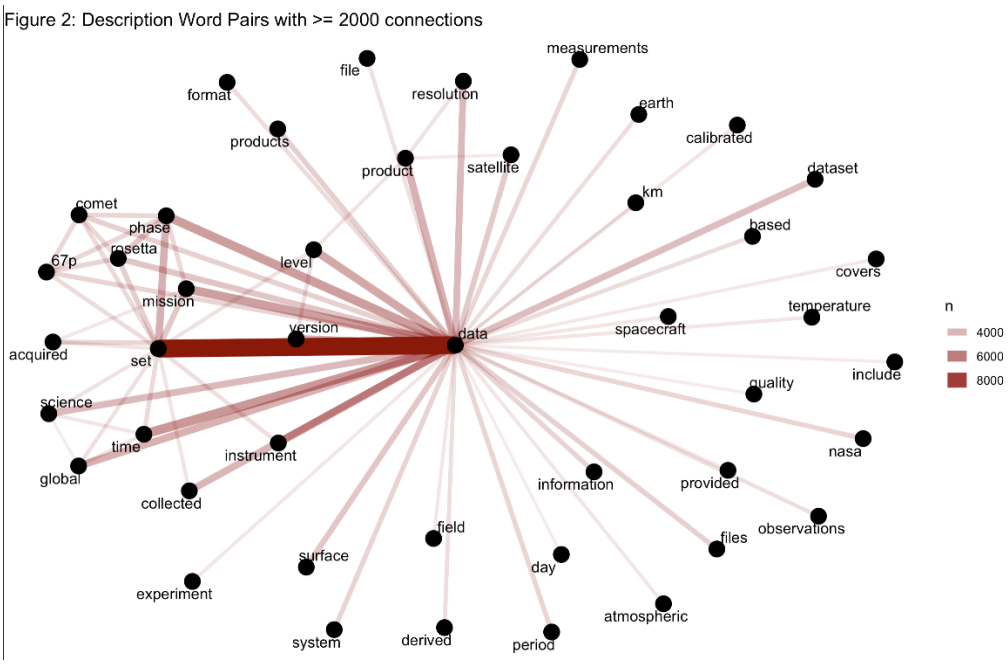

Figure 2: Description word pairs with >= 2000 connections

NASA Keyword Relationships

Moving beyond individual words, Figures 3 and 4 provide a deeper understanding of the relationships between keywords assigned to NASA datasets. Figure 3 visualizes frequently co-occurring keyword pairs, showcasing "Earth Science" as a central theme. Strong connections to "Atmosphere," "Oceans," and "Spectral Engineering" further delineate the agency's research focus. Interestingly, "International Rosetta Mission" emerges as a distinct cluster, highlighting its unique keyword associations.

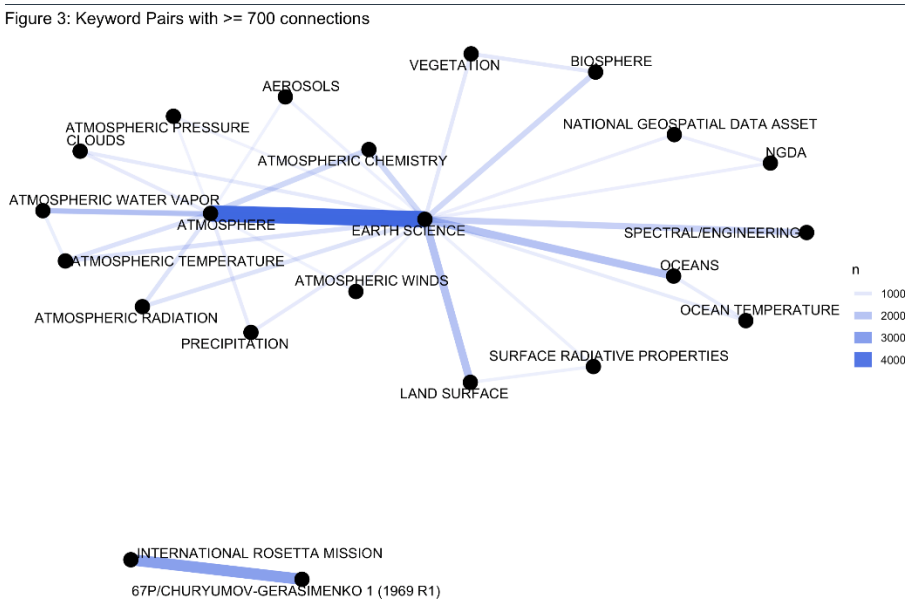

Figure 3: Keyword pairs with >= 700 connections

Figure 4 complements this analysis by visualizing strong correlations between keywords. The dense cluster centered around "Nucleic Acid Extraction" and "Library Construction" suggests a significant body of research related to biological and genetic analysis, potentially linked to astrobiology or human spaceflight. These network visualizations demonstrate the power of exploring keyword relationships to uncover hidden connections and clustered research areas within the NASA dataset.

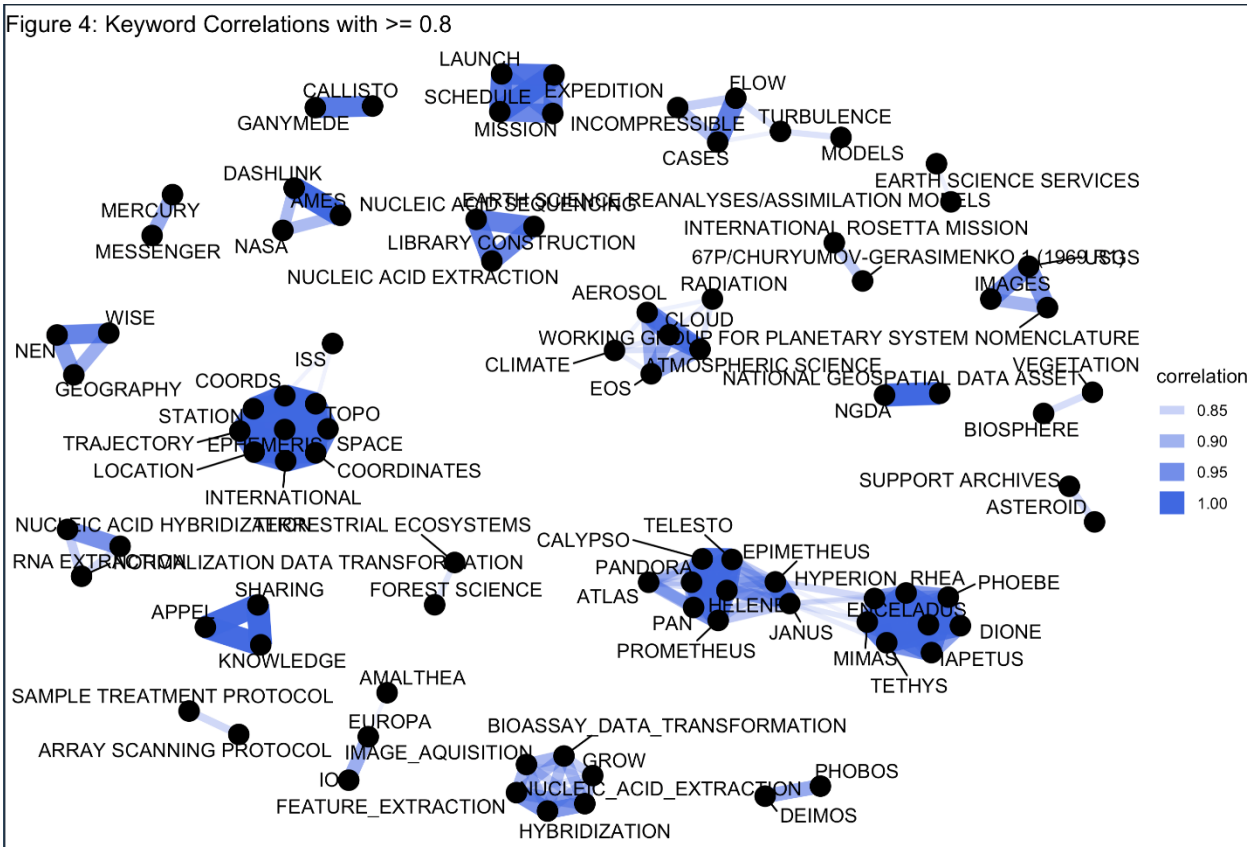

Figure 4: Keyword correlations >= 0.8

Keyword-Specific Insights

Figure 5 provides a granular perspective on NASA dataset descriptions by examining the highest scoring words within specific keywords using Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a numerical statistic that reflects how important a word is to a document within a collection of documents. A high TF-IDF score suggests that the word appears frequently in a specific document but is relatively rare across the entire corpus, indicating its importance to that document's theme.

In Figure 5, we focus on six prominent keywords: "Earth Science," "Atmosphere," "Precipitation," "Climate," "Clouds," and "Atmospheric Science." The plot reveals insightful patterns, such as the prominence of "brutsaert" within "Earth Science," "giovanni" within "Atmospheric Science," and "lysimeter" within "Precipitation," likely indicating specific instruments, datasets, or researchers associated with these fields. By visualizing high TF-IDF words within key thematic

areas, we gain a more nuanced understanding of the specific terminology and research areas characterizing some of NASA's Earth-centric datasets.

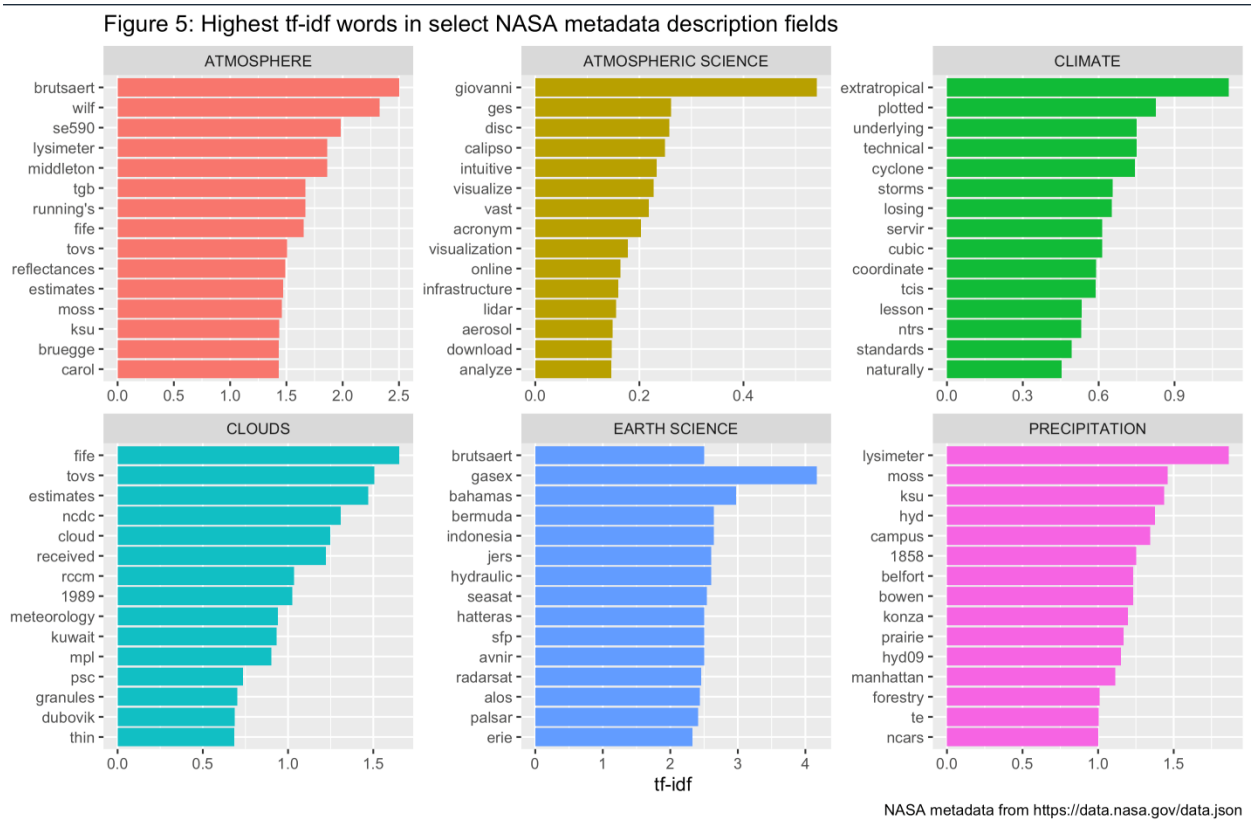

Figure 5: Highest TF-IDF words in select NASA metadata description fields

Latent Derichlet Allocation (LDA)

Figure 6 unveils latent themes within NASA dataset descriptions using a topic model generated through Latent Dirichlet Allocation (LDA). LDA is a statistical method that identifies underlying topics within a corpus by analyzing the distribution of words across documents. In this case, we trained an LDA model to discover 12 distinct topics within the NASA dataset descriptions. It is certainly possible that a larger k value is more appropriate, but chose 12 for timely code execution and data processing.

Figure 6 presents the top 10 most relevant terms for each topic, ranked by their beta value, which represents their importance within the topic. For example, Topic 4 prominently features terms like "data," "analysis," "space," and "samples," suggesting a theme related to data analysis and sample collection in space exploration. Similarly, Topic 8 focuses on "Earth," "atmosphere," "temperature," and "precipitation," indicating a theme related to Earth's climate and atmospheric studies.

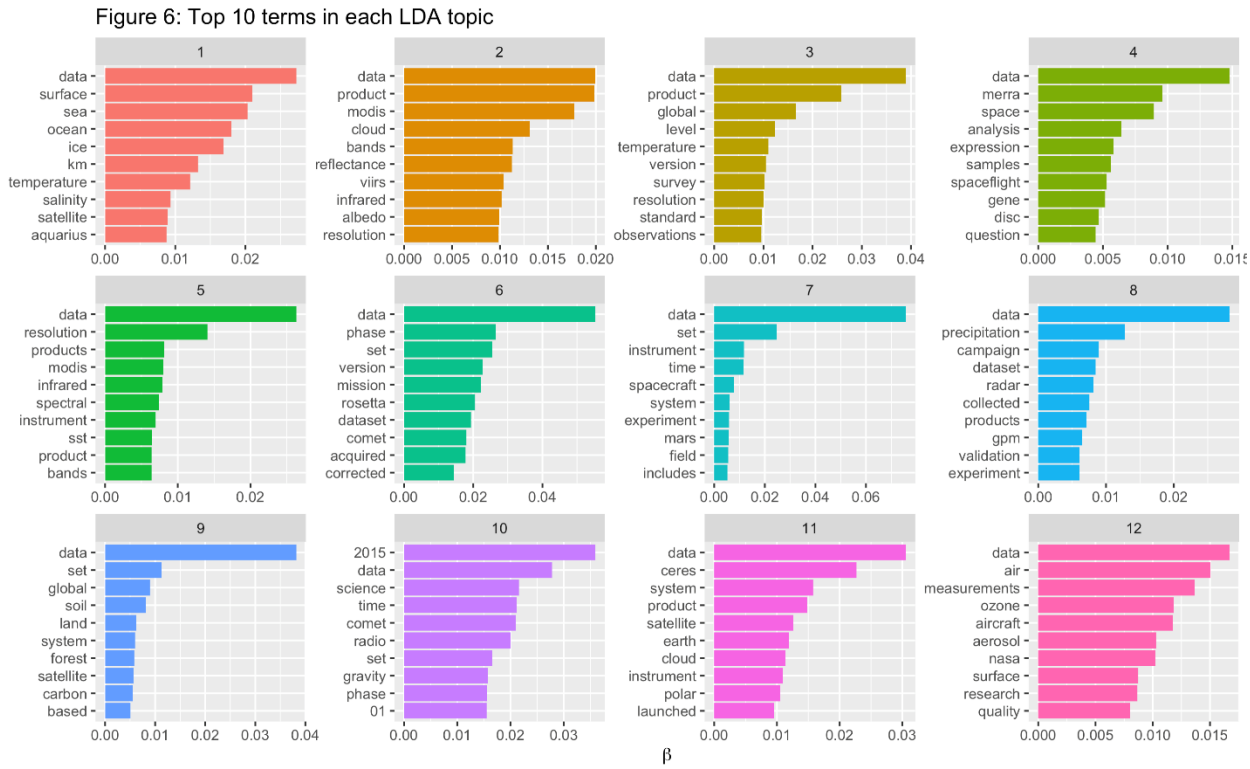

Figure 6: Top 10 terms in each LDA topic

Topic Analysis: Document Distribution and Probabilities

Figures 7 and 8 provide a deeper dive into the distribution of topics within the NASA dataset, utilizing the concept of gamma probability from our LDA model. The gamma probability represents the likelihood that a given document belongs to a particular topic. Figure 7 offers a bird's-eye view, illustrating the distribution of gamma probabilities across all 12 topics. The log scale on the y-axis highlights the wide range of probabilities, with peaks at both ends of the spectrum.

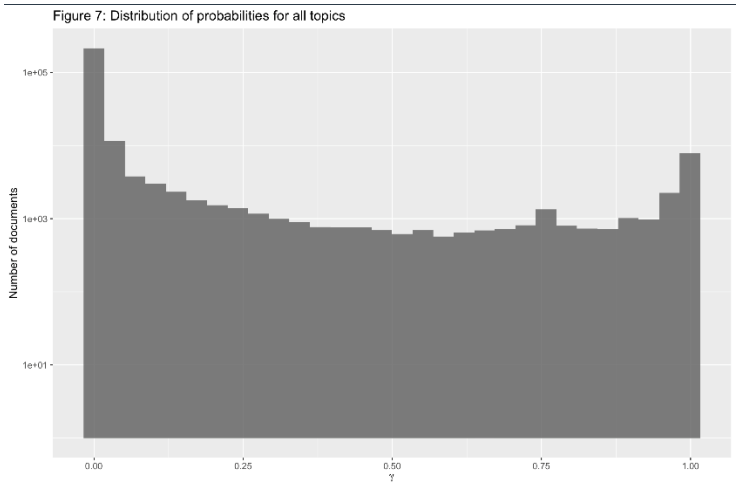

Figure 7: Distribution of probabilities for all topics

Figure 8 provides a more granular perspective by showcasing the gamma probability distribution for each topic individually. This allows us to see variations in topic prevalence.

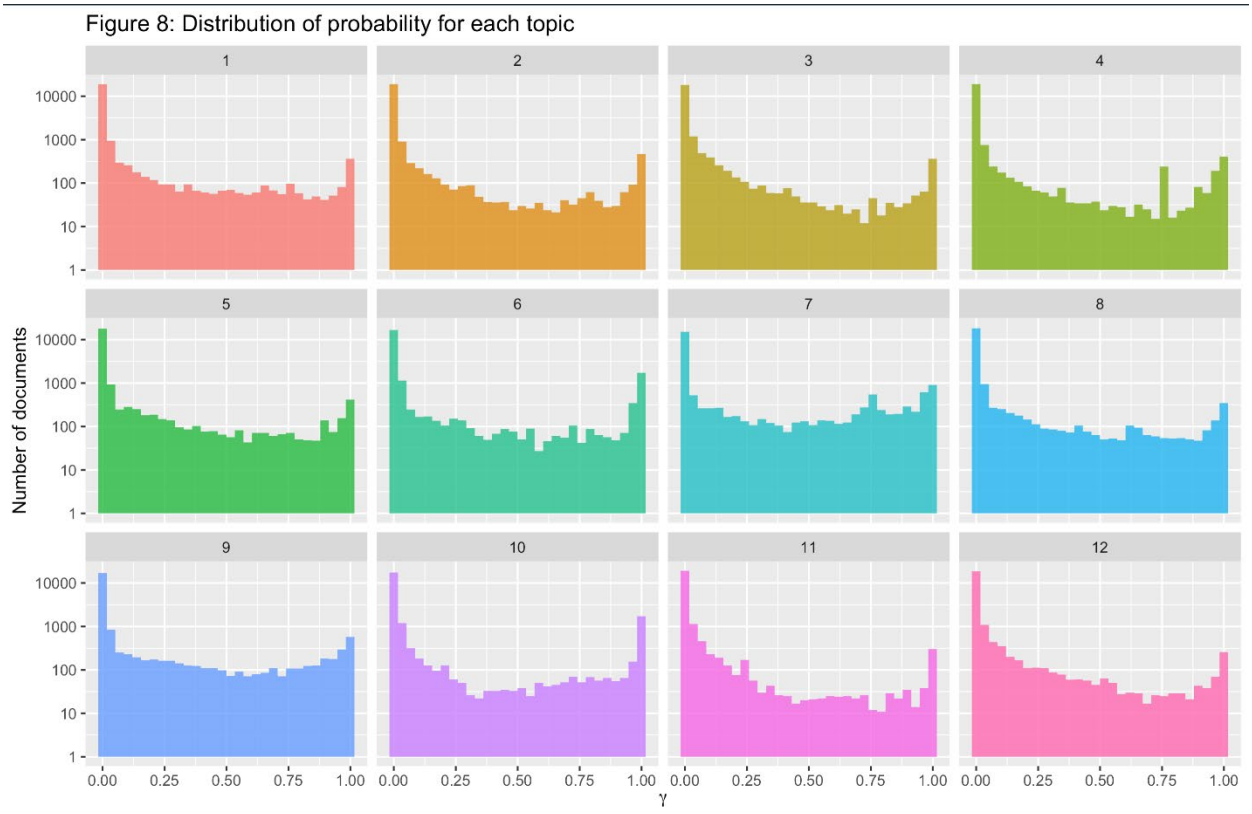

Figure 8: Distribution of probability for each topic

Connecting the Dots: Aligning Topics with Keywords

Figure 9 bridges the gap between the statistically derived topics and the human-assigned keywords within the NASA dataset. This connection allows us to attach more concrete labels and interpretations to the abstract themes identified through LDA. The plot focuses on documents with a high gamma probability (greater than 0.9) for a given topic, ensuring a strong thematic alignment. By counting the occurrences of keywords within these highly probable documents, we can identify the most prevalent keywords associated with each topic.

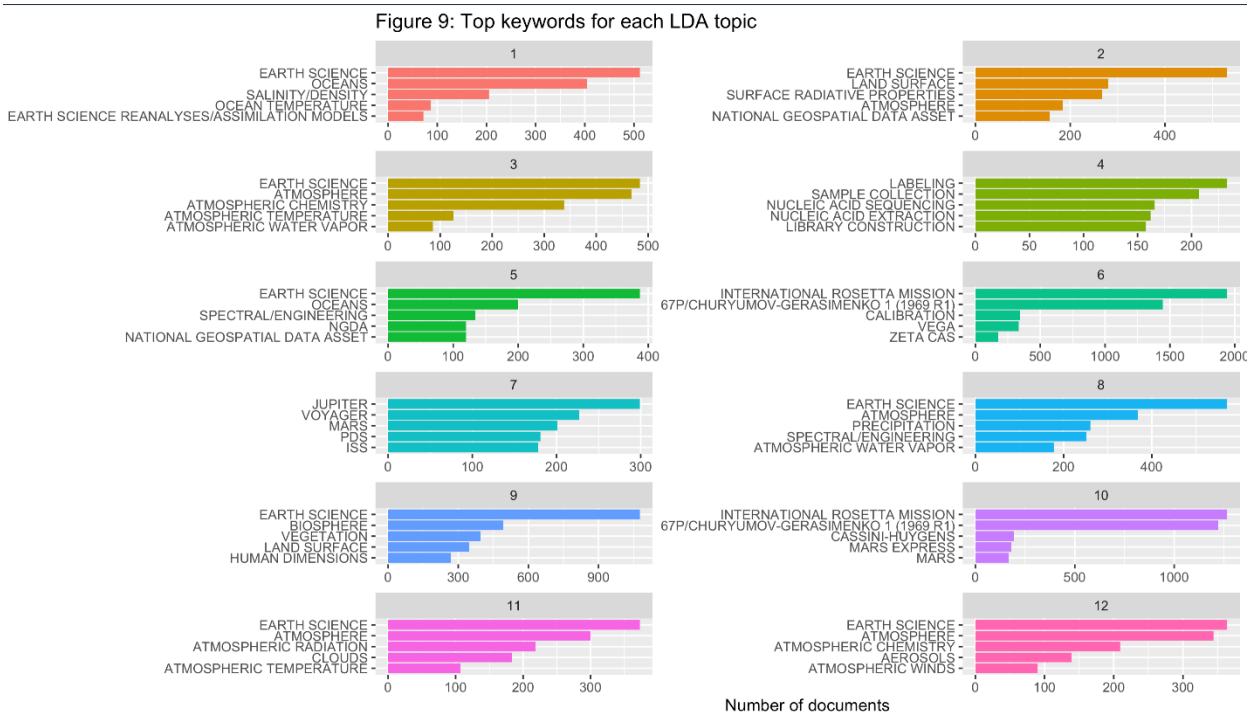

Figure 9: Top keywords for each LDA topic (k = 12)