ANALYSING TRENDS ON NATURAL BIOTECHNOLOGY THROUGH TOPIC MODELING

Submitted by

HARISH G D

2327422

SUBJECT: Text and Social Media Analytics (TSMA)

CIA 2 Assignment

Guided by

Prof. ROSEWINE JOY

CHRIST SCHOOL OF BUSINESS AND MANAGEMENT

BENGALURU

October-2024

Overview of Nature.com:

Nature.com is a premier platform for accessing high-quality scientific research, part of the renowned Nature portfolio published by Springer Nature. It is especially valued for publishing peer-reviewed research articles, reviews, and opinion pieces across a diverse range of scientific disciplines, including life sciences, physical sciences, health sciences, and environmental sciences. Nature.com hosts flagship journals like Nature, which is known for groundbreaking studies, and specialized journals such as Nature Genetics, Nature Medicine, and Nature Sustainability, each focused on specific fields.

The website offers resources beyond journal articles, including news on scientific advancements, career advice for researchers, and insights into policy implications, making it a comprehensive hub for scientists, educators, and students. Additionally, Nature.com provides various levels of access, including open-access articles, subscription-based journals, and options for institutional access, ensuring that researchers and enthusiasts worldwide can access cutting-edge science.

Through these resources, Nature.com contributes significantly to advancing scientific knowledge, fostering cross-disciplinary collaboration, and supporting the global scientific community.

Nature.com further increases its value to the scientific community by offering unique features that enhance research accessibility, collaboration, and knowledge dissemination. One such feature is Nature Research Editing Services, which assists authors in refining their manuscripts, improving readability, and ensuring adherence to scientific standards. This service helps researchers effectively communicate complex findings, improving the chances of publication.

Additionally, Nature.com integrates interactive data visualizations and multimedia content, such as videos, infographics, and podcasts, which make scientific discoveries more engaging and accessible to a broader audience. Nature's Outlook series, which offers in-depth analysis on current topics like climate change, gene editing, and sustainable energy, provides a broader context that benefits both specialists and general readers.

The platform also promotes open science through initiatives like Nature Communications and Scientific Reports, which provide open-access options for authors and enable free access to research. Such initiatives reflect Nature.com's commitment to bridging the gap between academia and the public, advancing scientific understanding across disciplines and fostering more inclusive, collaborative research.

Types of Journals:

Nature.com hosts a range of specialized journals across various scientific fields, each focused on promoting advancements in distinct research areas. Here are a few notable ones:

1. Nature: As the flagship journal, Nature covers groundbreaking research across all scientific disciplines, from biology and chemistry to physics and Earth sciences. Known for its rigorous peer-review process, Nature publishes only the highest-impact research, making it one of the most prestigious scientific journals globally.

2. Nature Genetics: This journal focuses on studies in genetics and genomics, highlighting advances in understanding genetic mechanisms, gene therapy, population genetics, and molecular evolution. Researchers often publish landmark findings in genetics here, influencing research directions in medicine, biology, and evolutionary studies.

3. Nature Medicine: A leading journal for translational and clinical research, Nature Medicine publishes research that bridges the gap between laboratory discoveries and clinical applications. Its articles on disease mechanisms, therapeutic approaches, and clinical trials are instrumental for healthcare researchers and professionals working to improve patient care.

4. Nature Sustainability: This journal centers on research into sustainable development and environmental science. It includes studies on topics like climate change, renewable energy, and sustainable food systems, aiming to inform policy and practice for global sustainability challenges.

5. Nature Biotechnology: As a premier source for developments in biotechnology, this journal covers advancements in bioengineering, molecular diagnostics, synthetic biology, and pharmaceutical development. Nature Biotechnology serves as a critical resource for biotech researchers, industry professionals, and policy makers.

6. Nature Reviews Series: This series includes several review journals, such as Nature Reviews Molecular Cell Biology, Nature Reviews Cancer, and Nature Reviews Immunology. These journals publish high-quality reviews and perspectives on recent developments, synthesizing the latest research and providing a comprehensive understanding of complex scientific areas.

7. Nature Climate Change: A journal dedicated to research on climate science, policy, and societal impacts, Nature Climate Change publishes studies that explore global climate dynamics, adaptation strategies, and climate resilience. It is highly relevant for researchers, environmentalists, and policy makers tackling climate-related challenges.

8. Nature Astronomy: Catering to the field of astronomy and astrophysics, this journal publishes research on topics ranging from planetary science and cosmology to astrophysical phenomena. It brings together discoveries about the universe, benefiting astronomers and space scientists.

Each journal in the Nature portfolio combines stringent peer review, editorial expertise, and scientific innovation, making them vital resources across disciplines. They collectively enhance the availability of specialized knowledge, enabling scientists, educators, and decision-makers to stay informed on critical scientific advancements.

Publication Policy:

Nature's publication policy is built around principles of scientific rigor, transparency, and integrity, ensuring that all published research meets high standards of quality. Their policy emphasizes several key areas:

1. **Rigorous Peer Review**: All manuscripts submitted to *Nature* journals undergo a thorough peer review process, typically involving multiple reviewers who are experts in the relevant field. This process ensures the validity, originality, and impact of the research, with an emphasis on novelty and scientific advancement.
2. **Data Availability and Transparency**: Nature mandates that authors make all relevant data publicly available and accessible. This includes sharing datasets, code, and detailed methodologies where possible, allowing other researchers to replicate findings and build upon previous studies.
3. **Ethical Standards**: Ethical considerations are core to Nature's publication policy. Authors must confirm that their research adheres to ethical standards in areas like human or animal research, conflicts of interest, and data privacy. For instance, research involving human participants requires informed consent, and studies with animal subjects must adhere to institutional or national ethical guidelines.
4. **Authorship Accountability**: Each author listed on a paper is required to contribute significantly to the research, and all authors must agree to the final manuscript before submission. This ensures accountability and prevents issues like ghost authorship or honorary authorship. The corresponding author takes responsibility for communicating with the journal and managing revisions.
5. **Disclosure of Competing Interests**: Nature requires authors to disclose any potential conflicts of interest that may influence the research or its interpretation. This could include financial support from companies or personal relationships that could bias results, ensuring transparency for readers.
6. **Plagiarism and Misconduct**: Nature strictly prohibits plagiarism, data fabrication, and falsification. Submitted manuscripts are screened for originality using plagiarism-detection software, and any instances of misconduct are addressed according to the Committee on Publication Ethics (COPE) guidelines. Authors found violating these guidelines may face retraction or banning from future submissions.
7. **Open Access Options**: Nature provides several publication models, including open access, which allows authors to make their work freely available to readers worldwide. Open access articles incur a publication fee but increase the reach of the research, fostering wider accessibility and potential impact.
8. **Reproducibility and Reporting Standards**: To strengthen reproducibility, Nature's policy requires detailed reporting of experimental procedures, statistical analyses, and controls. Clear and comprehensive reporting is emphasized to enable other researchers to verify or replicate studies, addressing reproducibility issues across scientific fields.

Nature's publication policy supports a transparent, ethical, and reproducible approach to scientific research. By prioritizing these standards, Nature aims to ensure that all publications

contribute positively to scientific knowledge and maintain trust within the academic community.

Data Source: Journal Of Nature Biotechnology:

As part of the analysis, a total of 51 research articles were downloaded from the journal of Biotechnology. This journal, which focuses on various aspects and trends of Biotechnology provides peer-reviewed, open-access content that is freely available online.

These articles cover a range of topics within the Nature Biotechnology domain, including:

1. **Genetic Engineering and Gene Editing**: Research on CRISPR-Cas9, TALENs, and other gene-editing tools that allow precise modifications in DNA. Studies often explore applications in therapeutic gene editing, crop improvement, and gene drive technology.
2. **Synthetic Biology**: Work in this area involves designing and engineering new biological parts, organisms, and systems. Topics include creating synthetic genomes, designing metabolic pathways, and developing synthetic cells with custom functions for medical or industrial use.
3. **Bioinformatics and Computational Biology**: The application of computational tools and machine learning in analyzing biological data, such as genome sequencing, protein structure prediction, and drug discovery. This also includes developing algorithms and databases for managing large-scale biological datasets.
4. **Biopharmaceuticals and Therapeutics Development**: Research focuses on the discovery, engineering, and production of biologics like monoclonal antibodies, vaccines, and cell and gene therapies for various diseases. This area also covers personalized medicine and targeted therapies based on genetic profiles.
5. **Regenerative Medicine and Stem Cells**: Studies on stem cell technology, including the development of induced pluripotent stem cells (iPSCs), tissue engineering, and regenerative therapies. Research often investigates applications in treating injuries, degenerative diseases, and organ repair.

The journal is well-regarded for its rigorous peer-review process and contributions to Nature Biotechnology related research. The articles downloaded will serve as a foundation for a comprehensive analysis of current trends, practices, and challenges in Bio Technology, with an emphasis on their practical implications for businesses.

Text Preprocessing and Analysis

Step 1: Word Count

To begin, we calculated the total word count across all 51 PDF documents. This was done by splitting the text of each document into individual words and summing the total count across all files.

- **Total word count across all PDFs**: 5,534,071

Step 2: Text Lowercasing

Following the word count, each document's text was converted to lowercase to maintain consistency, ensuring that words differing only in case would be treated as the same, thereby reducing redundancy.

```
ion combined with limited proteolysis-coupled mass spectrometry (flip-ms), a structural proteomics work
proval earlier this year of \na non-engineered t cell immunotherapy, \niovance's amtagvi (lifileucel), w
n-efficiency, low-toxicity and scalable in vivo genome editing if efficacious rnp-lnp complexes can be r
cture are not readily \nscalable to heterogeneous tissues. here we introduce droplet hi-c, which uses a
vendors help specifically \nwith recruitment for clinical trials, a task \ndifficult for big and small
```

Step 3: Tokenization Using Regular Expressions
To divide the text into meaningful units (words), each document was tokenized using a regular expression to extract alphabetic sequences. Only alphabetic characters (A-Z, a-z) were retained, excluding numbers, punctuation, and special characters.

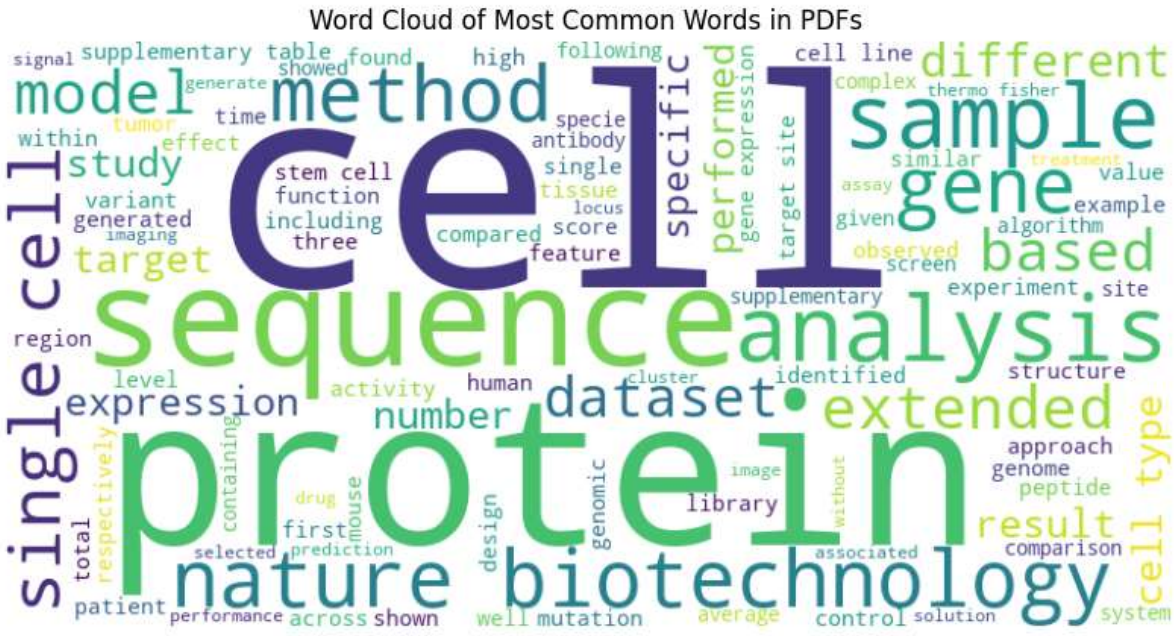
Step 4: Stop Word Removal
Common stop words (e.g., "and," "the," "is") were removed using the NLTK library's predefined list. This step helped focus the analysis on words relevant to the articles' content by excluding frequently occurring words that do not contribute significant meaning.

```
Total number of Stop words removed: 235882
```

Step 5: Cleaning Irrelevant Words
Beyond stop words, additional irrelevant words—such as common acronyms, publisher-related terms (e.g., "doi," "open access"), and miscellaneous terms (e.g., "https")—were identified and removed. A customized list of such words was created to further refine the token list and better focus on content related to CSR while filtering out irrelevant publishing jargon.

Step 6: Lemmatization
Finally, the WordNetLemmatizer from the NLTK library was applied to reduce words to their base or root form (e.g., "studies" to "study"), allowing words with the same meaning in different forms to be treated uniformly.

Top Words in Journal of Nature Biotechnology

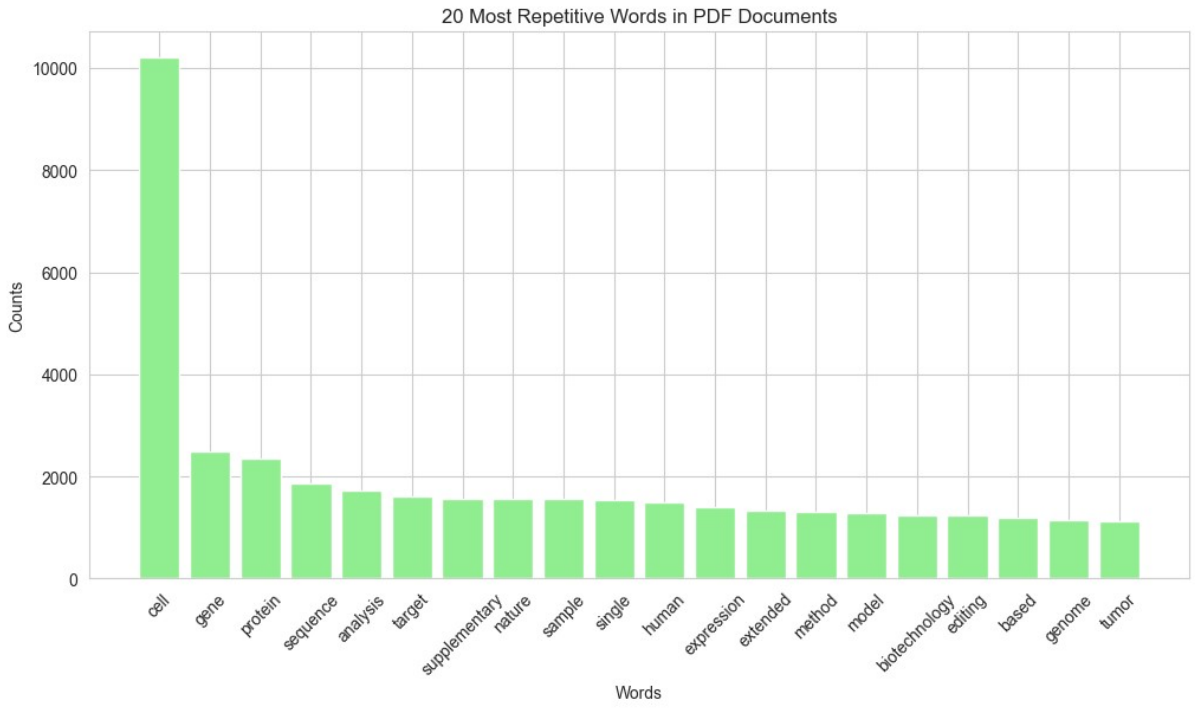


The word cloud suggests that the PDFs analyzed were likely related to biological research, with a focus on genetics, proteomics, and cell biology. Key themes include:

- Cellular research: Studies on cell differentiation, gene expression, and signaling pathways.
- DNA sequence analysis: Analysis of DNA sequences for gene identification, mutation detection, and genome-wide association studies.
- Protein research: Investigation of protein structure, function, and interactions.
- Drug discovery and development: Screening compound libraries to identify potential drug candidates.
- Computational methods: Use of machine learning and other computational techniques to analyze biological data.

Overall, the word cloud indicates a broad range of biological research topics, with a particular emphasis on genomics, proteomics, and cell biology

Most Frequent Words in the Articles:

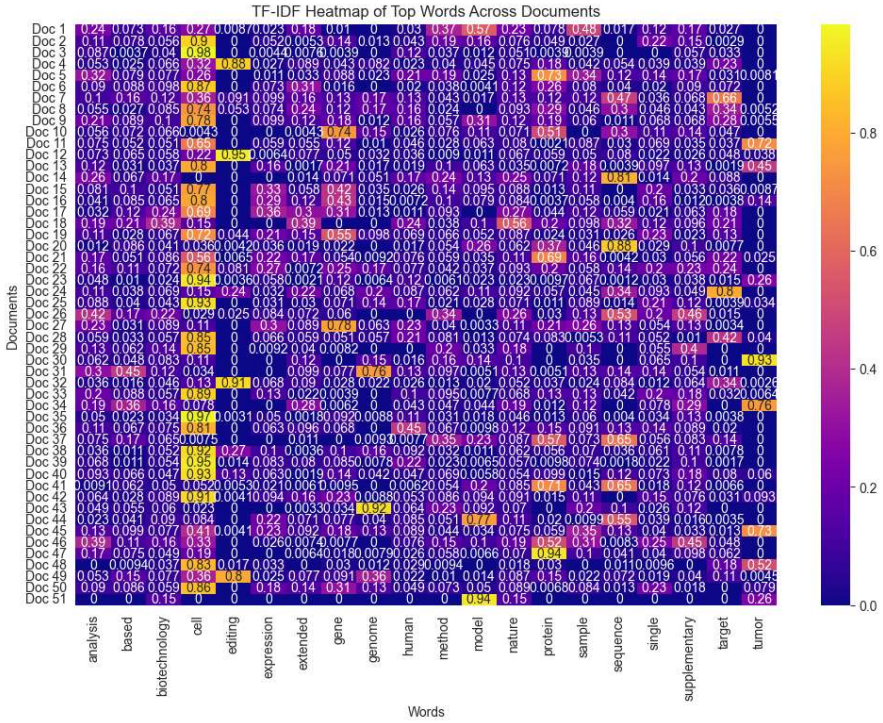


Key Insights

- **"Cell" is the most frequent word:** This suggests that the PDFs likely contain a significant amount of content related to cellular biology, genetics, or related fields.
- **Other biological terms are prominent:** Words like "gene," "protein," "sequence," and "analysis" further support the biological theme.
- **"Supplementary" and "nature" are also frequent:** These terms might indicate that the PDFs include supplementary materials or discuss natural phenomena.

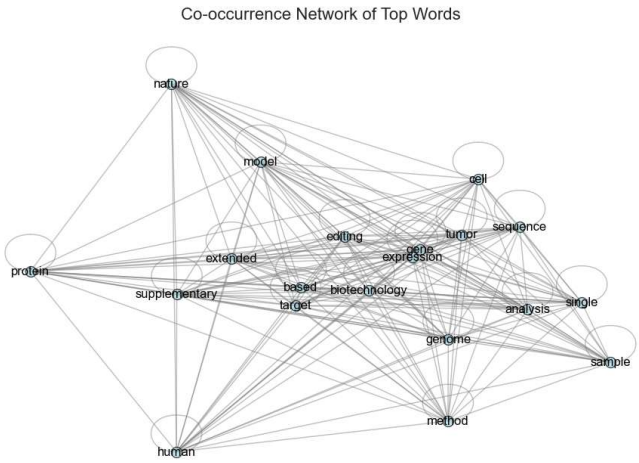
- **A few words suggest computational or technical aspects:** Words like "model" and "editing" hint at the use of computational tools or techniques in the analysis of biological data.

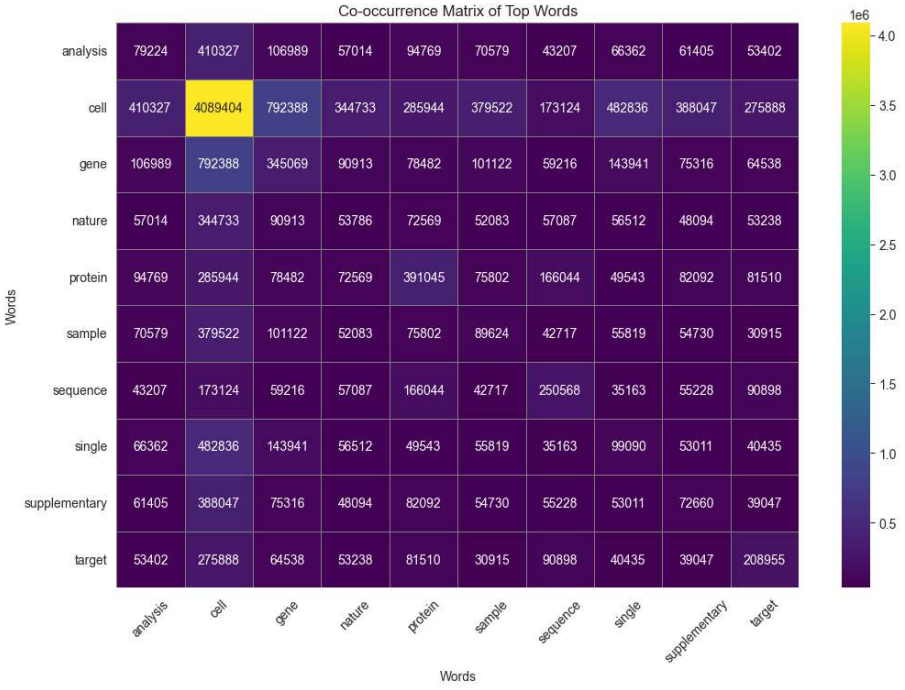
Term Frequency and Inverse Document Frequency:



- **Word Distribution:** The heatmap shows that some words, like "cell," "gene," and "protein," are relatively common across most documents, indicating a general biological theme. However, other words, such as "tumor," "extended," and "biotechnology," are more specific to certain documents.
- **Document Similarities:** Documents with similar color patterns likely share similar topics or themes. For example, documents with strong red or yellow colors in the "tumor" column might be focused on cancer research.
- **Document Differences:** Documents with distinct color patterns likely have different focuses. For instance, a document with a strong green color in the "editing" column might be about gene editing techniques, while a document with a strong blue color in the "nature" column might be about natural biology.

Co-Occurrence Matrix and Network Diagram:





Key Insights:

1. **Strong Co-occurrences:**
   - Cell and Gene: This pair has the highest co-occurrence count, indicating they frequently appear together, which makes sense given their strong biological relationship.
   - Gene and Protein: This pair also has a high co-occurrence count, reflecting the central dogma of molecular biology.
   - Nature and Protein: This pair suggests discussions about proteins in their natural context, such as within cells or organisms.
2. **Moderate Co-occurrences:**

- Analysis and Cell: This pair suggests that the documents involve analyzing cellular data or processes.
- Sequence and Gene: This pair indicates discussions about DNA or RNA sequences and their relation to genes.
- Sample and Target: This pair might suggest experimental setups where samples are analyzed to identify specific targets.

3. **Lower Co-occurrences:**
   - Single and Supplementary: These pairs have lower co-occurrence counts, suggesting they might appear in more specific contexts or less frequently.

I. Document-Term Matrix (DTM) Creation

Process:
After text preprocessing, the cleaned and lemmatized tokens from each article were combined back into a single string, enabling each document to be treated as a cohesive unit for analysis. The sklearn library's CountVectorizer was then used to generate a Document-Term Matrix (DTM).

- **Document-Term Matrix (DTM):** This matrix has rows representing individual documents (i.e., articles) and columns representing unique terms. Each cell in the matrix indicates the frequency of a term within the corresponding document.
- The DTM is crucial for transforming text data into a numerical format that can be processed by machine learning models. It is a sparse matrix that captures term-document relationships, allowing for analysis of word frequencies and patterns across the corpus.

Output:
The resulting DTM had the following shape:

- Shape: (51, 27,883) — representing 51 rows (one for each article) and 27,883 unique terms (columns).

```
Document-Term Matrix shape: (51, 27883)
   aaaaa  aaaaaa  aaaaaaaga  aaaaaagcac  aaaaaagcaccgactcggtgcc  aaac  \
0      0       0         0           0                       0     0
1      0       0         0           0                       0     0
2      0       0         0           0                       0     0
3      0       0         0           0                       0     0
4      0       0         0           0                       0     0

   aaacagatcacccgctgagcgggttatctgttcnumber  aaai  aaalac  aaat  ...  zygosity  \
0                                        0     0       0     0  ...         0
1                                        0     0       0     0  ...         0
2                                        0     0       0     0  ...         0
3                                        0     0       0     0  ...         0
4                                        0     0       0     0  ...         0

   zygote  zygotic  zygous  zyla  zymo  zymoclean  zzzm  zzzzz  zzzzzz
0       0        0       0     0     0          0     0      0       0
1       0        0       0     0     0          0     6      0       0
2       0        0       0     0     0          0     0      0       0
3       0        0       0     0     0          0     0      0       0
4       0        0       0     0     0          0     0      0       0

[5 rows x 27883 columns]
```

II. LATENT DIRICHLET ALLOCATION (LDA) FOR TOPIC MODELING

After constructing the DTM, we applied Latent Dirichlet Allocation (LDA) to uncover underlying topics within the *Nature Biotechnology* articles. LDA is a popular topic modeling technique that assumes each document consists of several topics, with each topic represented by a unique distribution of words.

- **Number of Topics:** Initially, the LDA model was set to identify 7 topics, aiming to capture meaningful distinctions based on the articles' content in *Nature Biotechnology*.
- **Model Fitting:** The LDA model was then fitted to the DTM, allowing it to learn the topic distributions across documents and highlight the most relevant words associated with each topic.

LDA enables the discovery of hidden thematic structures within large document collections. For this biotechnology-focused analysis, topic modeling could reveal recurring themes such as genetic engineering, bioethics, drug development, sustainability, regulatory affairs, public health, and scientific innovation. These topics are essential for understanding the broader landscape of research within *Nature Biotechnology*

III. ANALYSIS OF KEY TOPICS

After fitting the LDA model, the top 15 words associated with each of the 7 topics were extracted. These words represent the most important terms that contribute to each topic, providing insight into the key themes present in the articles.

```
Topic 1:
protein sequence cell model analysis proteome supplementary gene complex method based peptide change sample structure

Topic 2:
cell editing target site sequence antibody protein extended variant human guide analysis mouse nature read

Topic 3:
cell human gene tumor mouse sample analysis study supplementary nature model single feature patient line

Topic 4:
cell gene single expression type extended perturbation spatial method score analysis raman nature drug state

Topic 5:
sequence protein genome supplementary cell gene read nature sample method design model high tumor structure
```

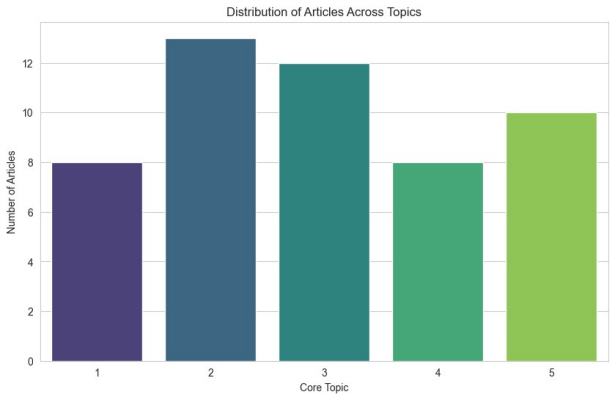
Findings:

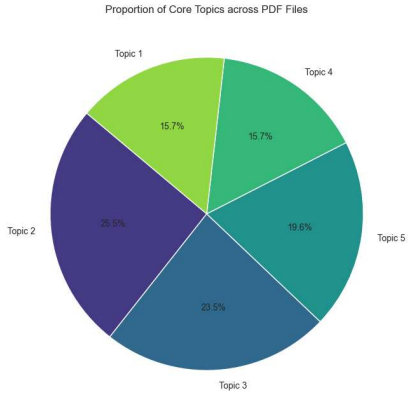
1. Topic 1: Protein and Cell Analysis
    - Key terms include "protein sequence," "cell model," "analysis," and "proteome."
    - This topic suggests a focus on the structural and functional analysis of proteins and their roles in cellular processes. The mention of "supplementary gene complex" and "peptide" implies an exploration of complex interactions within biological systems.

2. Topic 2: Gene Editing and Protein Targeting
    - The presence of terms like "cell editing," "target site," "antibody," and "protein" indicates a focus on gene editing technologies, such as CRISPR.
    - This topic highlights the applications of targeted gene modifications in both human and mouse models, emphasizing the relevance of these techniques in biomedical research.
3. Topic 3: Cancer Research and Genomics
    - The keywords "gene," "tumor," "human," and "patient" suggest a focus on cancer genomics and the analysis of genetic variations associated with tumors.
    - The use of terms like "sample analysis" and "single feature" implies that this research may involve detailed genomic studies on individual patients or tumor samples, potentially for personalized medicine.
4. Topic 4: Gene Expression and Perturbation Analysis
    - This topic features "cell," "gene," "expression," and "perturbation," indicating research into how genes are expressed in different conditions or treatments.
    - The inclusion of "spatial method" and "raman" suggests advanced techniques being used to analyze gene expression in specific locations within tissues, which is crucial for understanding spatial biology.
5. Topic 5: Genomic Sequencing and Structural Analysis
    - Key terms such as "sequence," "genome," "high tumor structure," and "design model" indicate a focus on sequencing technologies and their applications in studying complex genomic structures.
    - This topic points to the significance of high-throughput sequencing methods in understanding genetic architecture and its implications for diseases, particularly tumors.

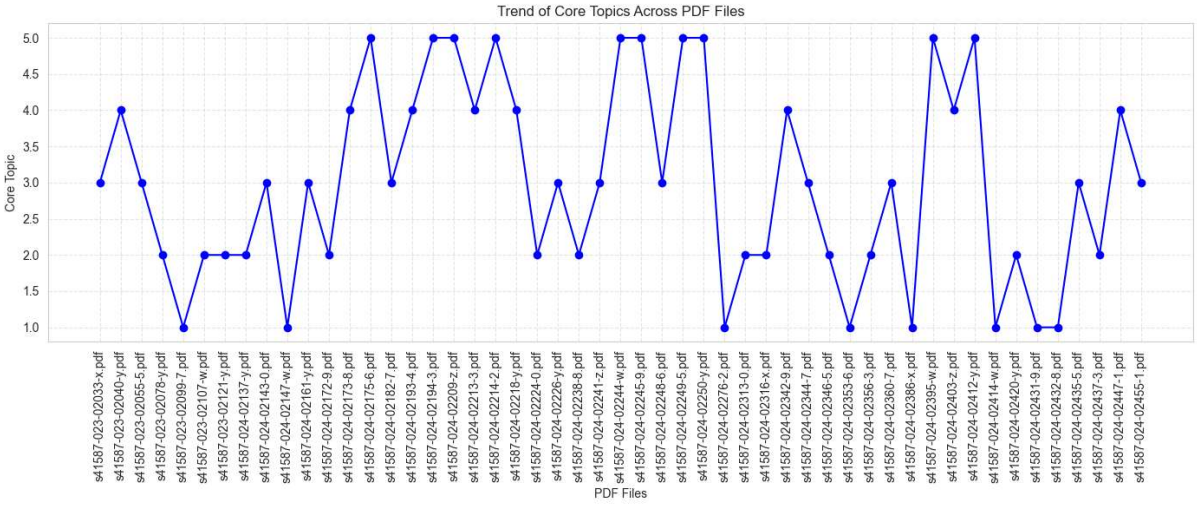
IV. TOPIC DISTRIBUTION AND DOCUMENT MAPPING

```
Count of articles per core topic:
Core Topic
1     8
2    13
3    12
4     8
5    10
Name: count, dtype: int64
```





- Topic Dominance: Topic 2 has the largest share, accounting for 25.5% of the content. This suggests that it is a central theme or focus area within the documents.
- Topic Distribution: The other topics (1, 3, 4, and 5) have relatively similar proportions, ranging from 15.7% to 23.5%. This indicates a balanced distribution of content across these topics.



Key Insights:

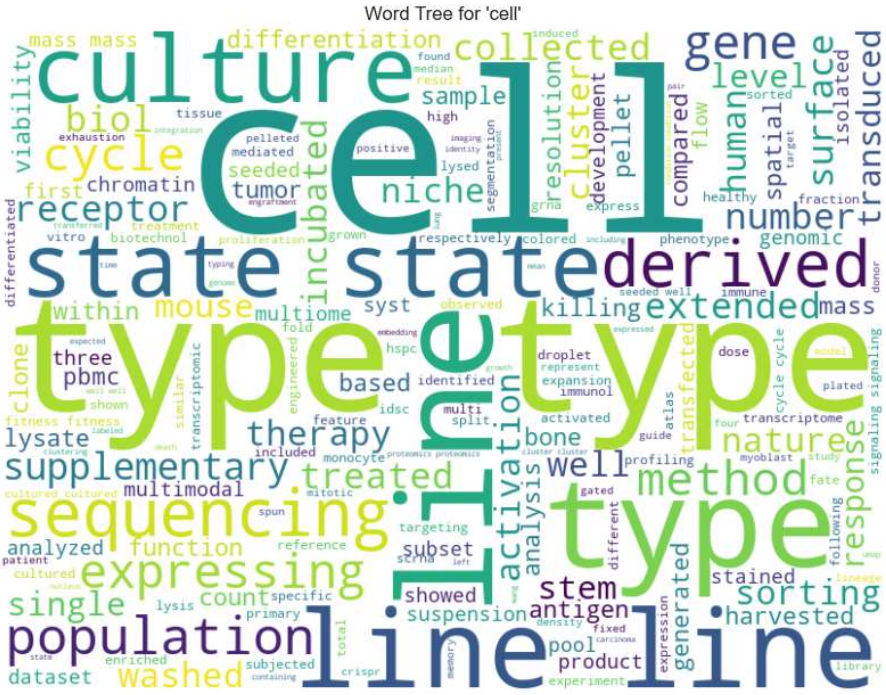
1. **Topic Fluctuation:** The plot shows that the prominence of the core topic fluctuates across different PDF files. It doesn't follow a consistent upward or downward trend.
2. **Peak and Trough:** There are instances where the topic's prominence peaks, indicating that it's a major focus in those specific files. Conversely, there are also troughs where the topic's relevance diminishes.
3. **Variability:** The plot highlights the variability in the importance of the core topic across the PDF collection. Some files might delve deeply into this topic, while others might only touch upon it briefly.

Word Tree:

Key Insights:

1. **Core Concepts:** The central theme is the gene itself, represented by the large font size. Surrounding it are words like "expression," "protein," "DNA," and "function," which are fundamental concepts related to genes.
2. **Gene Expression and Regulation:** Terms like "expression," "regulation," "transcription," and "translation" are prominent, highlighting the importance of gene expression in cellular processes.
3. **Gene Function and Variation:** Words like "function," "mutation," "variant," and "polymorphism" suggest an emphasis on the role of genes in determining phenotypes and the genetic basis of variation.
4. **Genetic Analysis and Technology:** Terms like "analysis," "sequencing," "editing," and "CRISPR" indicate the use of various techniques to study and manipulate genes.
5. **Biological Context:** Words like "cell," "organism," and "disease" emphasize the biological context in which genes operate.

Overall, the word tree reveals that the concept of "gene" is multifaceted and central to various fields of biological research, including genetics, molecular biology, and genomics. It highlights the interconnectedness of genes to other biological processes and their role in shaping the diversity of life.



Key Insights:

1. **Core Concepts:** The central theme is the cell itself, represented by the large font size. Surrounding it are words like "type," "function," "gene," and "protein," which are fundamental concepts related to cells.
2. **Cell Types and Differentiation:** Terms like "type," "stem," "differentiation," and "phenotype" highlight the diversity of cell types and their potential to specialize.
3. **Cellular Processes and Analysis:** Words like "function," "cycle," "metabolism," and "analysis" suggest an emphasis on understanding cellular processes and analyzing cellular data.
4. **Experimental Techniques:** Terms like "culture," "sample," "experiment," and "assay" indicate the use of various techniques to study cells.
5. **Biological Context:** Words like "tissue," "organism," and "disease" emphasize the cellular context within larger biological systems.

Overall, the word tree reveals that the concept of "cell" is multifaceted and central to various fields of biological research, including cell biology, genetics, and medicine. It highlights the interconnectedness of cells to other biological processes and their role in shaping the structure and function of organisms.

Findings and Conclusion:

1. Interdisciplinary Research: The identified topics reflect an interdisciplinary approach, integrating concepts from molecular biology, genomics, cancer research, and bioinformatics. This indicates that contemporary research is increasingly collaborative, utilizing diverse methodologies and perspectives.
2. Emerging Technologies: The emphasis on gene editing (Topic 2) and high-throughput sequencing (Topic 5) suggests that cutting-edge technologies are at the forefront of current research. These methods are driving significant advancements in understanding genetic and cellular mechanisms.
3. Personalized Medicine Focus: Topics 3 and 4 highlight the growing importance of personalized medicine, particularly in cancer treatment. Research is increasingly directed toward understanding individual genetic profiles to tailor therapies and improve patient outcomes.
4. Structural Biology: The recurring mention of protein structure and gene expression across multiple topics underscores the critical role of structural biology in understanding biological functions and disease mechanisms.
5. Future Directions: The findings suggest potential areas for future research, such as:
    - Investigating the interplay between proteins and genetic variations in disease.
    - Exploring novel gene editing techniques and their therapeutic applications.
    - Advancing spatial biology methodologies to enhance our understanding of tissue-specific gene expression.