

**How does gender inequality manifest in the workplace, and what contributing factors can be identified through a case study analysis, aiming to understand the challenges and dynamics surrounding this issue?**

**ÖMER CEM TABAR - 2081169**

**ISMAIL DEHA KÖSE - 2072544**

**ANIL ÖZFIRAT - 2087154**

**ZAFER ÇINAR - 2041428**

** **

**Abstract:** Gender inequality in the workplace is a pressing issue that requires immediate attention. Despite progress, disparities persist in areas like pay, promotions, and leadership representation. Addressing this problem is crucial, not only ethically but also for fostering creativity, innovation, and organizational success. Gender-based barriers hinder both individual and collective potential. Rectifying this inequality is not just about fairness but is essential for creating dynamic and successful workplaces that leverage diverse talents, experiences, and perspectives.

**Motivation:** Our project aims to comprehensively explore gender inequality in the workplace, covering issues like unequal pay, glass ceilings, gender bias, and stereotypes. We'll delve into the challenges individuals, especially women, face in professional environments, examining occupational segregation and industry-specific gender gaps. Our focus extends beyond monetary disparities to encompass broader themes such as workplace discrimination, harassment, and societal expectations. Additionally, we will investigate the role of corporate culture and policies in perpetuating or mitigating gender-based challenges. Analyzing Reddit posts and comments will provide valuable insights into lived experiences, highlighting both barriers and opportunities for addressing workplace gender inequality.

**1.Data Collection & Processing**

Having a larger volume of posts and comments significantly enhances the effectiveness of topic detection algorithms by providing a more comprehensive dataset for analysis. A greater quantity of data enables algorithms to identify recurring patterns and trends with higher accuracy, leading to more robust topic modeling. However, collecting a substantial amount of data from Reddit poses challenges due to limitations on API requests and potential ethical considerations. Retrieving a large dataset requires careful adherence to Reddit's policies and often involves time-consuming efforts to avoid overloading the servers or violating terms of use, making the process both complex and resource-intensive.

When exploring posts and comments on Reddit related to women and gender inequality, you might observe subtopics such as:

* **Gender Pay Gap:** Discussions on salary disparities between men and women in various industries and professions.
* **Workplace Discrimination:** Experiences of discrimination based on gender, ranging from biased hiring practices to unequal treatment at work.
* **Glass Ceiling:** Conversations about barriers women face in reaching higher-level positions within organizations.
* **Sexual Harassment:** Posts discussing instances of harassment and inappropriate behavior in the workplace.
* **Diversity and Inclusion:** Conversations about the importance of creating diverse and inclusive workplaces, and the challenges faced in achieving this.
* **Career Advancement Challenges:** Discussions on obstacles hindering women's career growth and opportunities for professional development.
* **Stereotypes and Biases:** Exploration of gender-based stereotypes and biases affecting women in different aspects of their professional lives.
* **Maternity and Family Leave:** Conversations about challenges and policies related to maternity leave, family planning, and work-life balance.
* **Intersectionality:** Consideration of how factors like race, ethnicity, and sexual orientation intersect with gender, influencing the experiences of women in the workplace.
* **Empowerment Initiatives:** Discussions on programs, organizations, or initiatives designed to empower women professionally and address gender inequality.

These subtopics can provide a rich and diverse set of perspectives when analyzing posts and comments related to women and gender inequality on Reddit.

Under these subtopics, we have collected the amount of posts and comments as follows:

| **# of posts considered** | | | **18.019** | | |
| --- | --- | --- | --- | --- | --- |
| **# of comments considered** | | | **1.113.569** | | |

**Superficial Cleaning:** Basic cleaning operations are performed on the text data, such as removing HTML tags, converting text to lowercase, removing emojis, and punctuation. The output is saved in a new column called text sup clean.

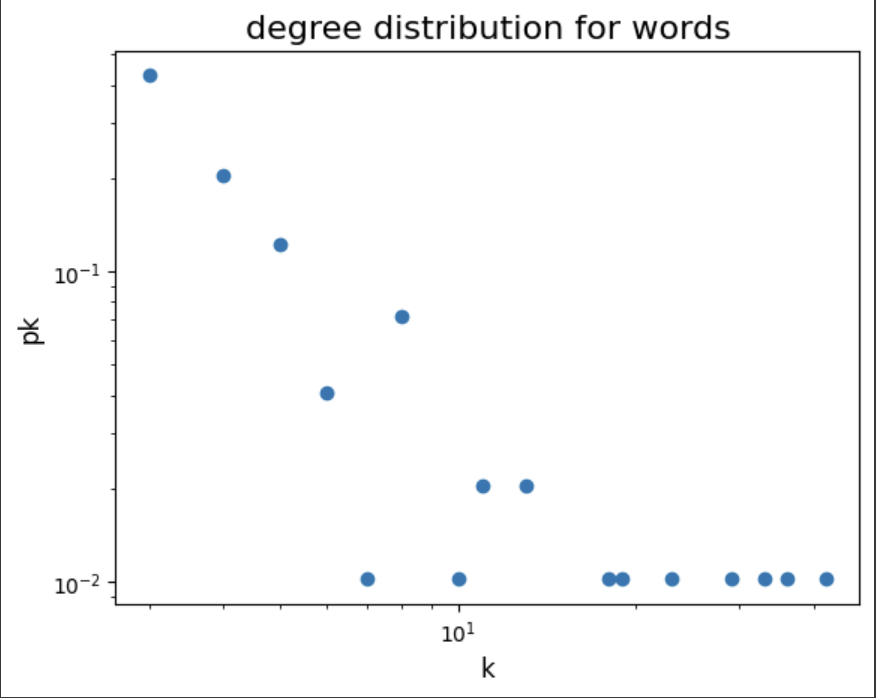
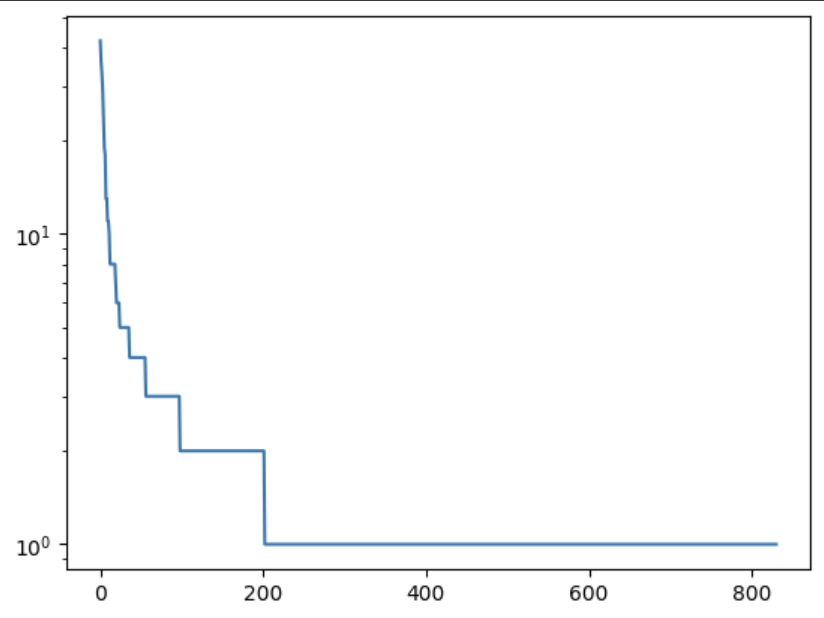
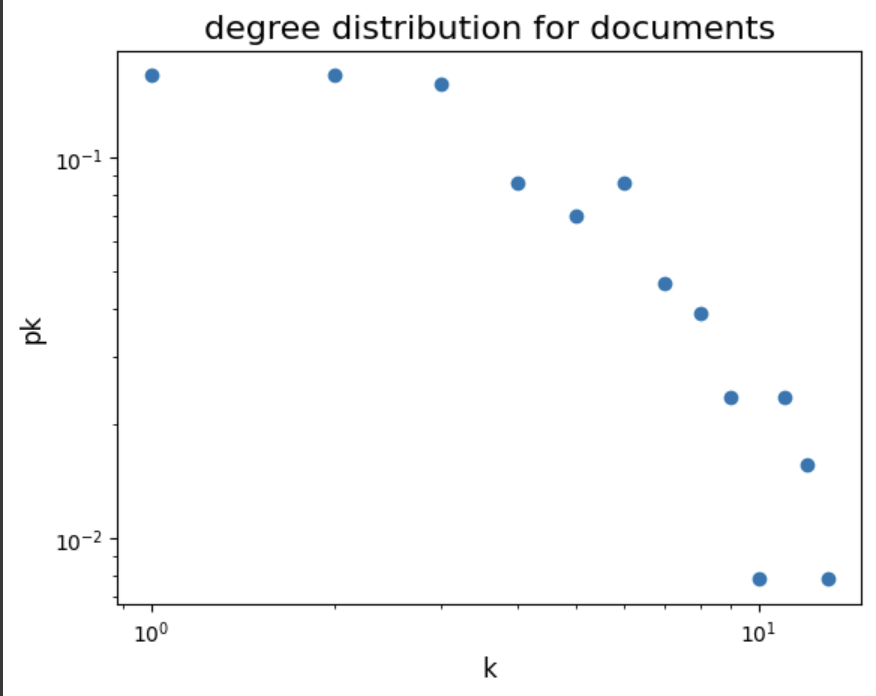
**Deep Cleaning:** Advanced cleaning operations are performed on the text data, including tok- enization, removal of stop words, and part-of-speech (POS) tagging. Two new columns, deep cleaning and deep cleaning pos, are created.

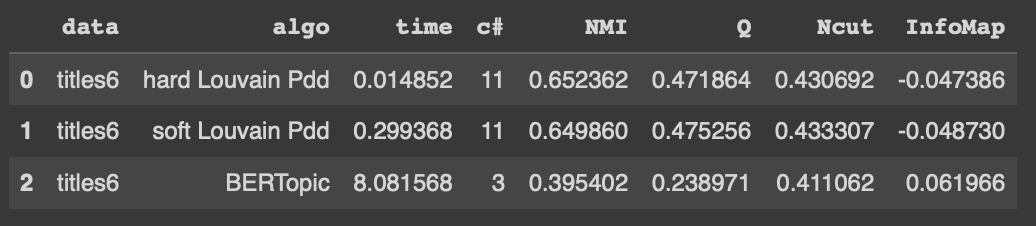
**Hashtag Extraction:** The code extracts hashtags from the translated text, if present, to analyze the relationship between words and hashtags.

**2.Building Semantic Network**

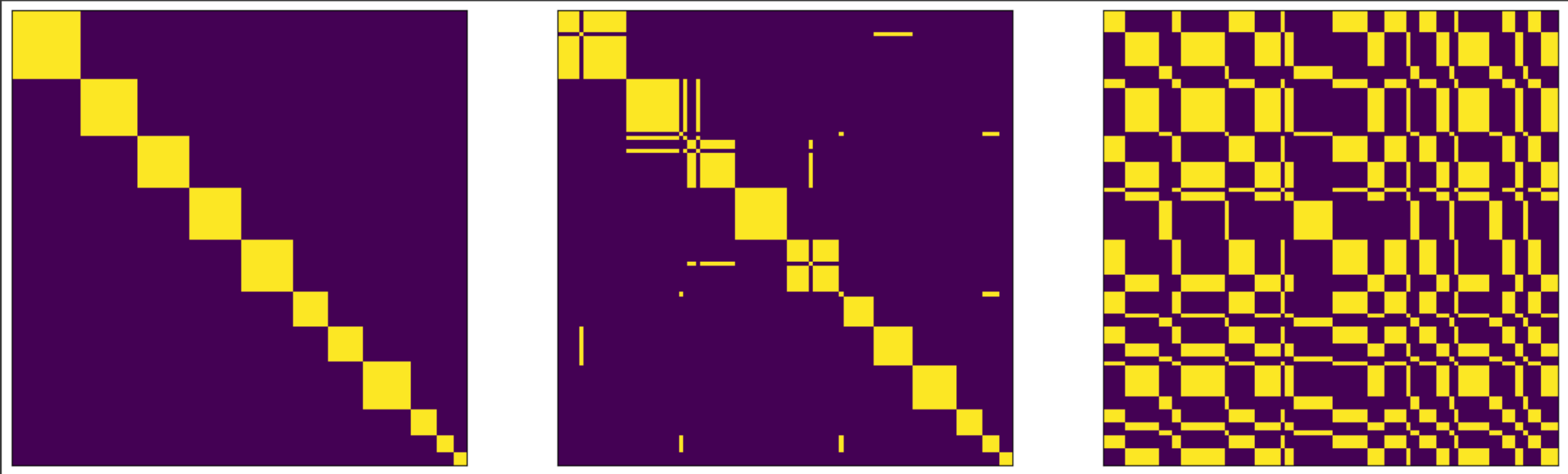
Creating a semantic network from Reddit posts involves extracting key themes from titles and comments to establish meaningful connections. Utilizing natural language processing, identified topics become nodes in the network, with edges representing relationships. This visual representation allows for a deeper exploration of discussion themes, enabling researchers to analyze patterns and relationships within the Reddit community.

* **Soft Louvain Community Detection:** Soft Louvain is a community detection algorithm that allows nodes to belong to multiple communities with varying degrees of membership, providing a nuanced understanding of overlapping relationships within a network.
* **Hard Louvain Community Detection:** Hard Louvain is a community detection algorithm that assigns each node to a single, non-overlapping community, simplifying the representation of network structures by distinctly categorizing nodes into mutually exclusive groups.
* **BERTopic Community Detection:** BERTopic is a community detection algorithm based on BERT embeddings, leveraging natural language processing to identify coherent topics within a text corpus and revealing underlying community structures in an unsupervised manner.

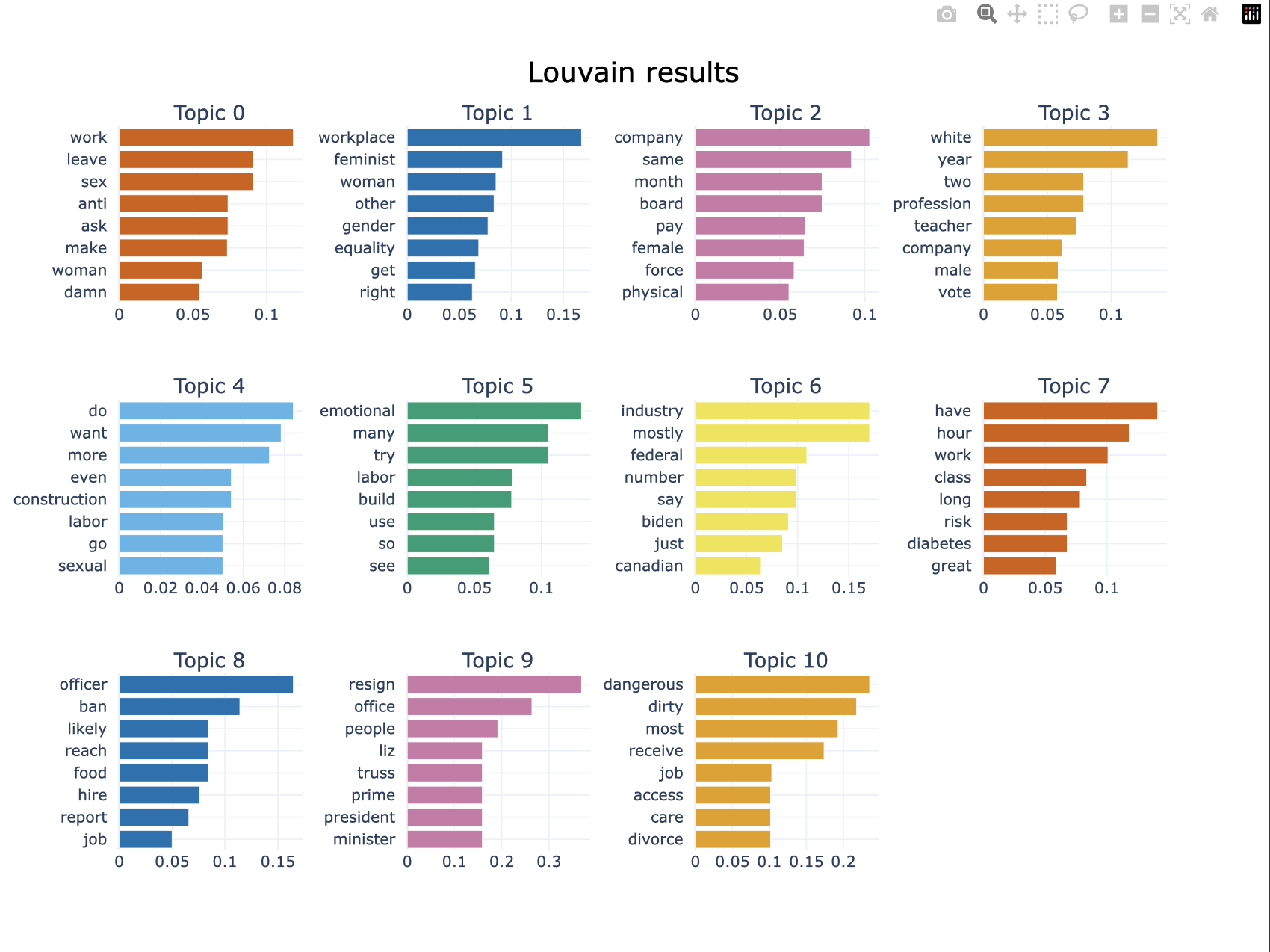
 

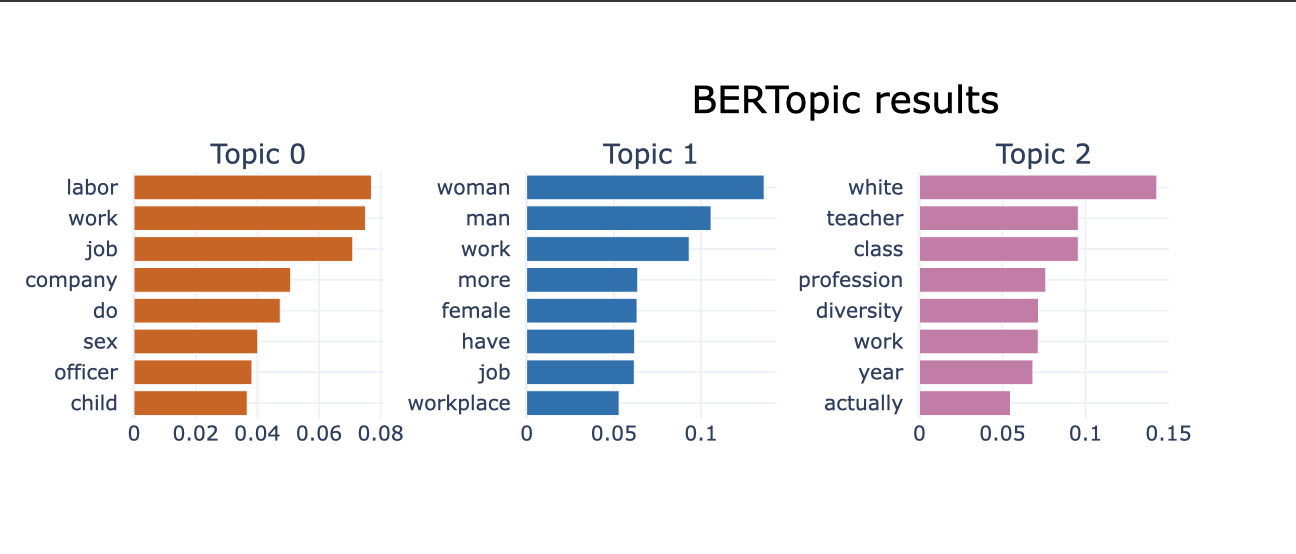


An array was created to compare three different community detection methods (hard Louvain, soft Louvain, and BERTopic) in terms of execution time, community count, Normalized Mutual Information (NMI), Modularity (Q), Normalized Cut (Ncut), and InfoMap (InfoMap) data. InfoMap ranking vector was found using the probability matrix for documents given documents.



**3.Obtained Results:**

****

****

The assessment and comparison of community detection methodologies—hard Louvain, soft Louvain, and BERTopic—involved the visualization of community structures, document allocations within each community, and the evaluation of relevant metrics. To accomplish this, WordCloud visualization in Python was employed as a preferred alternative to Gephi. The project specifically generated WordClouds to illustrate the communities identified by the Louvain algorithm, emphasizing the words and their respective weights. Word size was proportionally determined by word frequency, with more frequent words presented in larger sizes and less frequent ones in smaller sizes. This approach not only facilitated a nuanced exploration of community contents but also added an aesthetic layer through the application of color and shape enhancements, contributing to a more comprehensive understanding of the detected community structures.

