**Interim Report for Term Project**  
BTM539 Ryu, Sungyoung (20229007 / K-School)

This interim report presents an analysis of Korea's National Research and Development (R&D) proposals from the period of 2006 to 2020, related to No.9 of the United Nations Sustainable Development Goals (SDGs): Industry, Innovation and Infrastructure. The study utilizes the concatenation function from the Python library such as Np, Pd, Gensim to merge the individual proposal datasets into a single dataframe, comprising a total of 755,654 attributes and 141 variables. Despite the challenges posed by the large dataset size and the complexity of processing Korean language text, the availability of a consolidated and well-organized dataset sourced from a single database (NTIS) provides a valuable opportunity for analysis. However, it should be noted that certain years may contain missing data due to the inclusion of additional variables during annual revisions. The analysis primarily focuses on word preprocessing and word embedding techniques for further exploration.

**Introduction:**

The No.9 Industry, Innovation and Infrastructure of the United Nations SDGs highlights the crucial role of research and development in promoting sustainable industrial growth, fostering innovation, and enhancing infrastructure. This is the reason why this project aims to investigate and analyze Korea's R&D proposals during a 15-year period, spanning from 2006 to 2020.

**Methodology:**

The analysis involves merging individual R&D proposal datasets using the concatenation function from the Pandas library. By combining these datasets, a comprehensive and consolidated dataframe is created, containing a significant number of attributes and variables for further analysis. The study focuses on employing word preprocessing and word embedding techniques to extract meaningful insights from the textual content of the proposals.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

**Data Description**:

The dataset used in this study is sourced from the NTIS database, ensuring the reliability and organization of the data. The merged dataframe consists of 755,654 attributes, representing the number of R&D projects, and 141 variables capturing relevant information related to each project. It is important to note that some years may exhibit missing data due to the inclusion of additional variables during annual revisions.

스크린샷, 텍스트, 폰트이(가) 표시된 사진

자동 생성된 설명

**Current Status:**

The analysis aims to gain insights into the Korean National R&D proposals, specifically in the context of the Industry, Innovation and Infrastructure. Additionally, word embedding methods can be utilized to capture semantic relationships and patterns within the text data, enabling a deeper understanding of the key concepts and themes prevalent in the R&D proposals.

|  |  |
| --- | --- |
|  |  |
| Wordcloud Before pre-processing (Noise) | Wordcloud Post pre-processing on project title |

**Intermediate Conclusion:**

The analysis of Korea's National R&D proposals from 2006 to 2020 provides valuable insights into the country's efforts. Despite the challenges posed by the large dataset size and the complexity of processing Korean language text, the availability of a consolidated and well-organized dataset sourced from the NTIS database offers a unique opportunity for analysis. The application of word preprocessing and word embedding techniques allows for a more comprehensive exploration of the textual content and facilitates the identification of significant themes and patterns.

Based on the findings of this interim report, the following project is suggested for further research:

- Topic Modeling: Apply topic modeling techniques what we learnt from BTM539 class to identify the main topics and subtopics present in the R&D proposals, providing a structured overview of the research areas and priorities within the scope of the SDGs. (Hopefully I could do some forecast or extra)

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

The dataset used in this study encompasses a substantial volume of data, posing challenges in terms of its size and the complexity associated with natural language processing (NLP) tasks for accurately processing Korean text. Specifically, the task of identifying a suitable language model capable of effectively handling Korean characters presented a significant hurdle. However, the study has the advantageous circumstance of sourcing the dataset from a singular database, specifically the National Science and Technology Information Service (NTIS). This ensured the data's reliability and facilitated a streamlined and comprehensive analysis.