Research and development report

1. Research project name

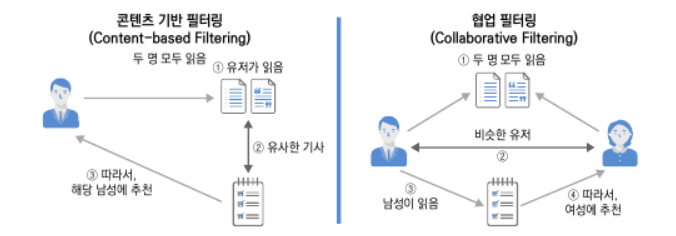
A Study on the Development of Government-funded R&D Project Recommendation and Search System Using Artificial Intelligence

1. R&D Task Overview

* research objectives

R&D support projects play a key role in supporting corporate growth and leap forward. In fact, companies such as "Inbody" and mobile real estate trading platform "Jikbang," which have the world's best body composition measurement technology, have provided an opportunity to leap forward through mid-term R&D. In Korea, the best way for companies to secure Seed Money is to participate in R&D government support projects as the angel investment environment is not developed, so they have no choice but to rely on loans in the early stages of start-ups. For the survival of domestic startups, it is important to keep in mind that participation in R&D support projects is a must, not an option. As such, it is very important for companies to participate in R& D support projects, but there are no services that help startups or small and medium-sized companies support public announcements yet. These needs and demands support the goals of this study.

Artificial intelligence (AI), a technology used in this study to "develop a government-backed R&D business recommendation and search system using artificial intelligence," is a computer engineering field that focuses on solving cognitive problems linked to human intelligence is an IT technology that is not limited to services and business. Among them, the algorithm to be used in this study is the recommended algorithm. The recommended algorithm is an algorithm that determines the user's input value with an artificial intelligence model and predicts and returns the result value that the user ultimately wants to see in advance. YouTube and Netflix, which we know, are also recommended algorithm-based services that recommend the following content according to each user's viewing record.

[]

[Figure 1] Example of content-based and collaborative filtering (source: Software Carpentry)

There are two typical filtering techniques for recommended algorithms (filtering technique: how to derive recommended result values ): "Content-based filtering" and "Collaborative filtering". Among them, the filtering technique we will use in this study is "content-based filtering" because unlike collaborative filtering techniques that require data (e.g., usage records) from other users in artificial intelligence model development and learning, content-based filtering is a technique that analyzes only user input values to find and filter similar documents. If various R&D government support project announcements are learned in the artificial intelligence model using Machine Learning techniques to find and show similar announcements in keywords entered by users, users will be able to easily and quickly decide which projects to participate in or wait for announcements from other projects, which will not only bring convenience to users (small and medium-sized companies).

* the main content

The research process was conceptually divided into four stages, and it analyzes users' ideas ( keywords) to participate in government support projects through each stage and proposes to establish a system that can provide users with a link to the announcement. Based on preliminary research on machine learning-based learning methods and data processing methods such as natural language processing, the algorithm was constructed in a way that allows users to easily and easily receive business recommendations through fast and light artificial intelligence models.

Background and necessity of this study

Purpose and scope of research

Analysis of R&D Support Project Data

Adopting artificial intelligence ( machine learning) algorithms

Data processing and artificial intelligence model learning (machine learning techniques)

Sample UI implementation for pilot operation

Confirmation of study application

Analyze effectiveness and directions based on implementation

**Purpose and scope of R&D**

**Research and Development Theoretical Consideration**

**Conduct R&D**

**Analysis of R&D results**

[Figure 2] Flowchart of the study

Artificial intelligence (AI) is being developed to provide convenient and efficient services to users, but as always, it always faces technical problems. Even if you have the data necessary to develop an artificial intelligence model, you need to adopt algorithms suitable for the amount, quality, and purpose of data, and model learning methods vary widely, so sufficient preliminary research and knowledge acquisition are needed for efficient performance. In addition, the presence of duplicate data negatively affects learning, so there are various sites to import the data, but it had to be selected as one. Fortunately, the amount and content of the data did not change even if the site was selected as one because the announcements shared by each site were similar.

텍스트이(가) 표시된 사진

자동 생성된 설명

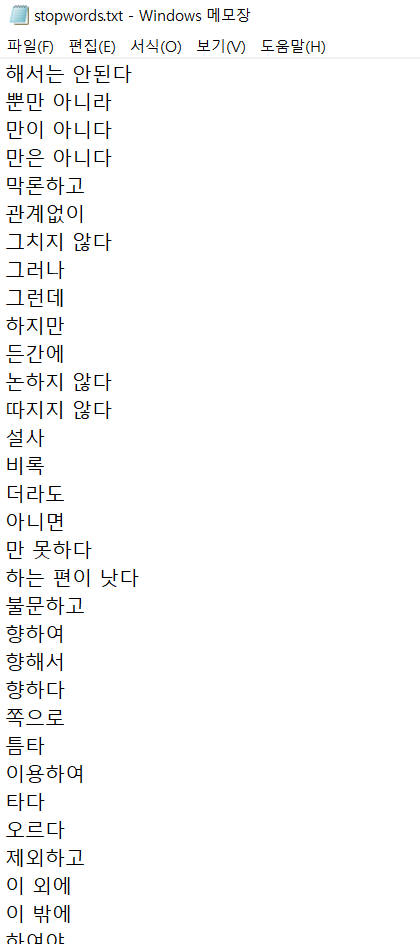
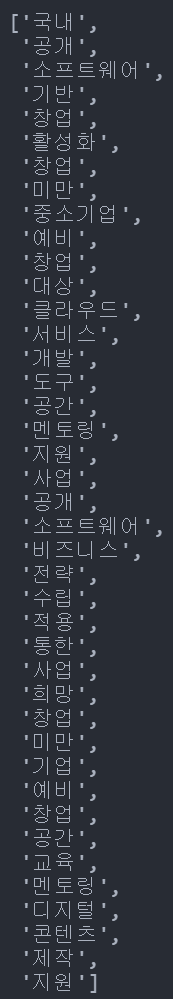
[Figure 3] Bizinfo category selection and download method

텍스트, 테이블이(가) 표시된 사진

자동 생성된 설명

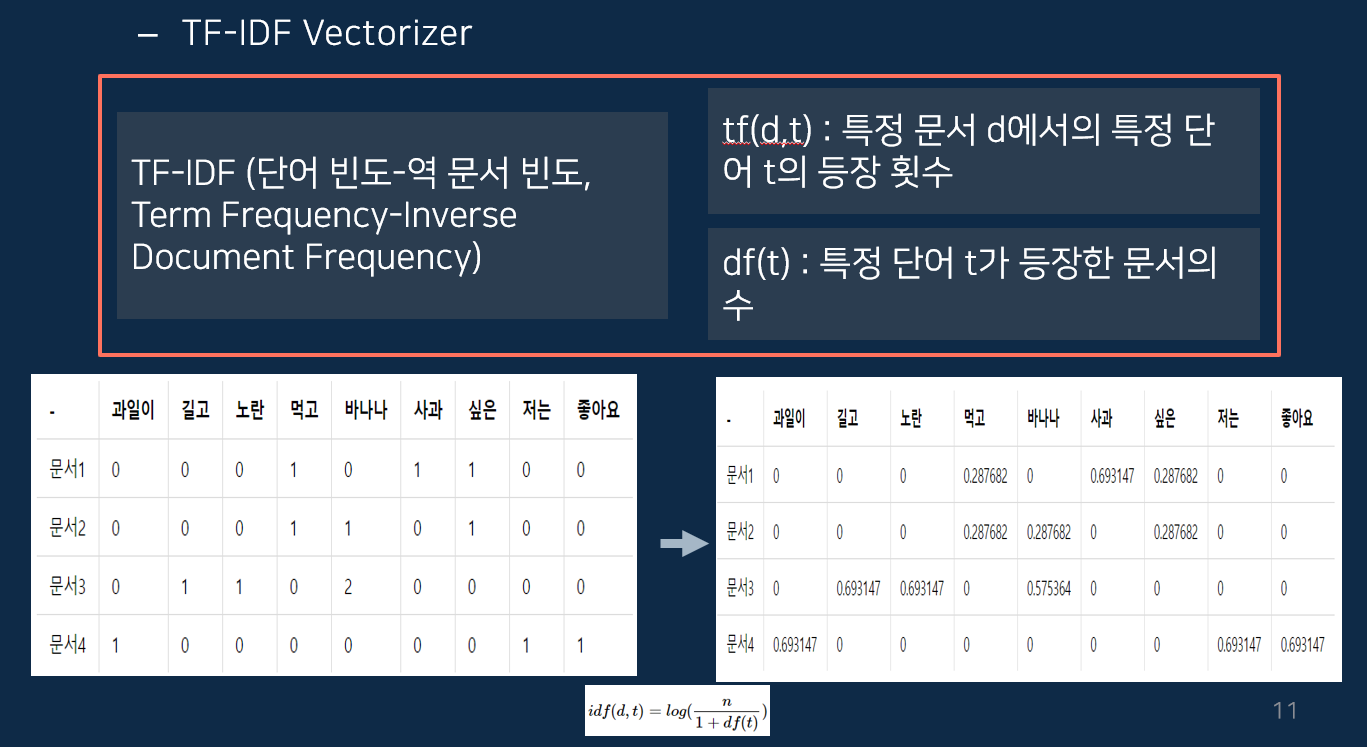
[Figure 4] Bizinfo Announcement of R&D Government Support Project between 2020 and 2021 (6642 cases)

The data to be used for model learning was easily available in bizinfo. Since it was convenient to download in the form of Excel (xlsx) and import data by linking API for free if necessary, it selected a corporate yard among the three candidate sites and downloaded about a year of R&D government support project data. 6,642 data were obtained, including technical fields, competent ministries, application start dates, and application deadline, and not only announcements in progress to improve the performance of the artificial intelligence model but also closed announcements were downloaded. At the beginning of the study, I thought about crawling data (business automation), but it was convenient to select and download the desired period and category in the corporate yard, and it saved a lot of time as I could skip the business automation stage.

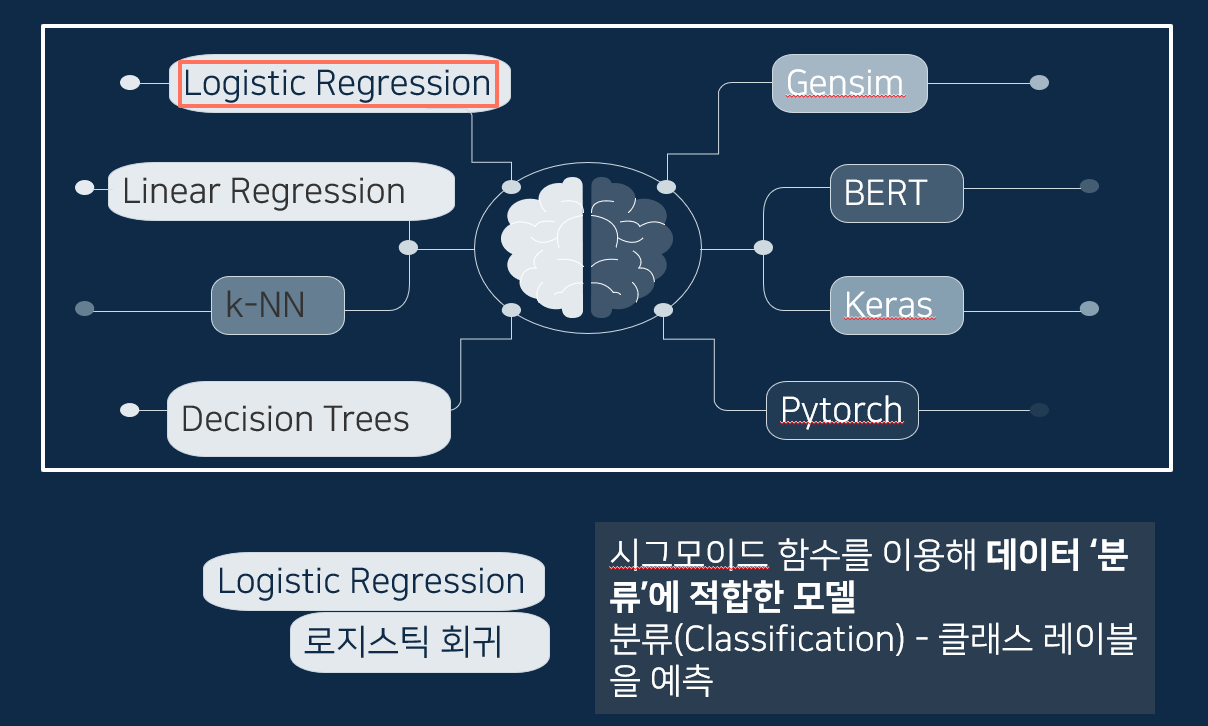
[Figure 5] Data processing using stopwords (pre-processing)

The first task after data collection is natural language preprocessing. Since an algorithm of a technique for measuring document similarity through keywords was adopted, words such as adverbs and verbs other than nouns were not needed in the R&D government-supported project announcement data. The most basic preprocessing technique, the word Tokenizing after removing the term, was used to remove the term first, and only the necessary words were extracted from the content. The terminology dictionary (txt file) [used the terminology](https://www.ranks.nl/stopwords/korean) dictionary of https://www.ranks.nl/stopwords/korean , which is widely used. Among the word tokenization libraries, "Okt" and "Kkma" were applied and tested once, and "Okt" that looks good with the naked eye was adopted. As shown in [Figure 5], key nouns such as 'software', 'service', and 'cloud' were successfully extracted.



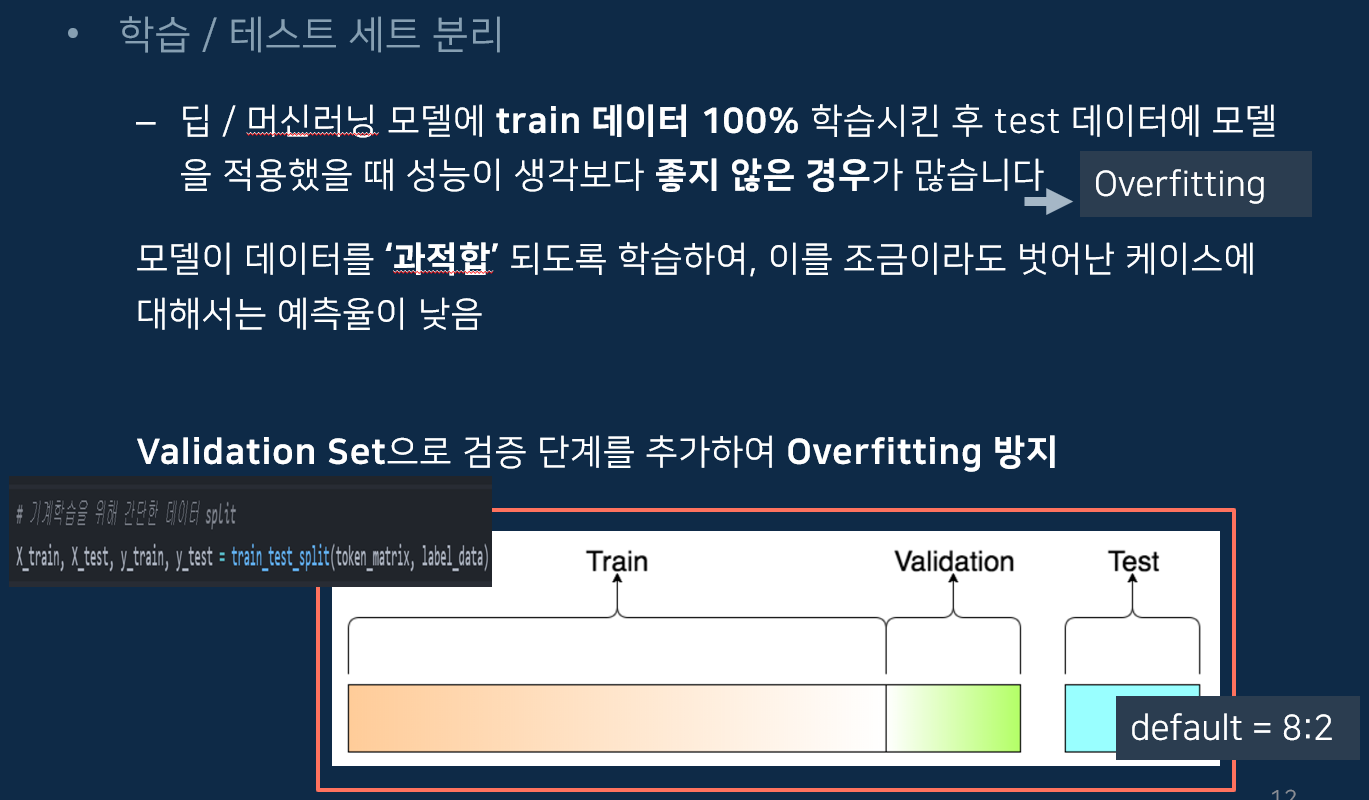
[Figure 6] TF-IDF Vector Algorithm

After extracting noun-oriented words from the contents of the data, the work performed is a document similarity measurement work. Typically, there are TF-IDF Vector and Counter Vectorizer in the library that measures similar scores for words, and in this study, TF-IDF Vector, which measures similar scores for each word according to the number of appearances and association of words, was used. In this way, the similarity of each word was measured and the score was pulse, and finally, a column was added to the dataset to check the similarity between documents, and each score was stored. Since the field of R&D government support project announcements was diverse, the TF-IDF Vector score also showed a wide distribution.



[Figure 7] Adopting an artificial intelligence model – Logic Regression

After securing the dataset obtained through the previous data purification and preprocessing, artificial intelligence model learning should be conducted. However, the artificial intelligence model, which is open for free, is limited and has different performance depending on the purpose and usage, so it is necessary to measure the performance once in a while to select the best model for this study. About 6,000 cases of data were determined to compare performance by using a small amount of machine learning techniques to proceed with deep learning techniques, reducing selection, and then selecting the logic regression, linear regression, and decision tree algorithms of the scikit-learn library.



[Figure 8] Artificial Intelligence Model Prediction Performance Test

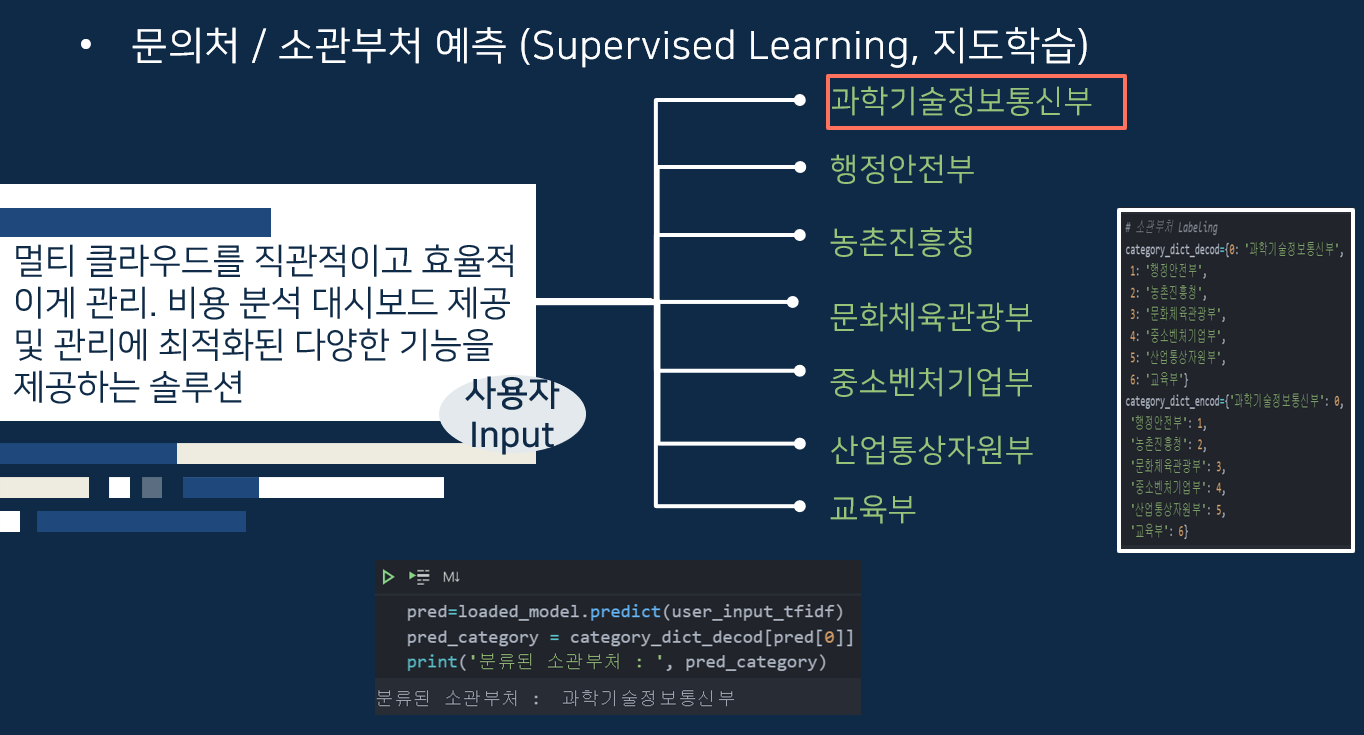
When learning and testing an artificial intelligence model, data is usually separated by a certain amount. On average , Train/Validation is conducted at a ratio of 8 (learning) : 2 (test). The reason for dataset separation is that if artificial intelligence model learning is carried out using all data, performance is degraded because it cannot provide data to be tested. Therefore, in this study, we also conducted a predictive performance test for each algorithm by keeping the ratio of 8:2.

텍스트이(가) 표시된 사진

자동 생성된 설명

[Figure 9] Adoption of artificial intelligence model and parameter tuning – Logistic Regression

In each algorithm performance test, the prediction rate of the Logic Regression algorithm was about 70%, ranking first among three machine learning algorithms. Finally, after adopting the logistic regression algorithm, hyperparameter tuning was conducted to further improve performance. Hyperparameter tuning is a technique that improves performance by narrowing the prediction of the model by modifying the values of the parameters provided by the scikit-learn library to suit the data. In the case of the dataset we prepared, we showed optimal performance when we modified the values of Max\_iter (number of learning) and solver (selecting detailed algorithms).



[Figure 10] Addition of prediction function of competent ministries

carry out research and development A new idea came up and I decided to add a feature. In addition to using the function of providing links to recommendations to users, the government will add a prediction function of the relevant department so that users can easily and easily know which department to contact if they have questions or inquiries. The addition of this feature was simple. In the dataset, there was a 'subordinate ministries' column, and seven ministries were already divided into each publication. As a result of conducting the test after learning the artificial intelligence model once more by labeling it with each number (0 to 6), the prediction rate was higher than when the announcement link was predicted.

텍스트이(가) 표시된 사진

자동 생성된 설명

[Figure 11] Refine user-entered data

After learning and testing the artificial intelligence model, a function for processing user input data was developed. An automatic correction function was added to the user 's input as if the announcement was refined and preprocessed so that even if the user enters a sentence other than a keyword, it can be predicted normally. Therefore, even if a user enters a keyword or sentence that is not related to the R&D government support project, the program will modify it accordingly. After receiving user input, key keywords are extracted through natural language preprocessing and similar announcements are searched through artificial intelligence models to show the most similar TOP 10 supportable announcements. In addition, 10 similar announcements are shown in the already closed announcements for reference to the past announcements.

텍스트이(가) 표시된 사진

자동 생성된 설명

[Figure 12] Results of public announcement recommendation based on user input

1. Contents and results of R& D performance

* R&D implementation period

Performance period: about 3 months from March to June 2021

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | March. | | | | April. | | | | May. | | | | | June. | | | |
|  | One week. | Two weeks. | Three weeks. | Four weeks. | Five weeks. | Six weeks. | Seven weeks. | Eight weeks. | 9 weeks. | 10 weeks. | Week 11 | 12 weeks | 13 weeks | 14 weeks | 15 weeks | 16 weeks | 17 weeks |
| Learning the necessary knowledge for AI development of artificial intelligence |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| A Study on the Data Distribution and Announcement Site for R&D Government-funded Projects |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Installing libraries and programs for model learning and testing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Collection of necessary data (R&D government-supported project announcement data) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Testing and selecting artificial intelligence algorithms |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Learning and Developing Artificial Intelligence Models |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Develop sample UI/UX for demonstration |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Demonstration and Testing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

[Table 1] Task Performance WBS

* R&D personnel

Lee [Jung-ho](mailto:jhlee@didim365.com) (jhlee@didim365.com ) Didim365 Co., Ltd., a researcher at the Corporate Affiliated Research Institute.

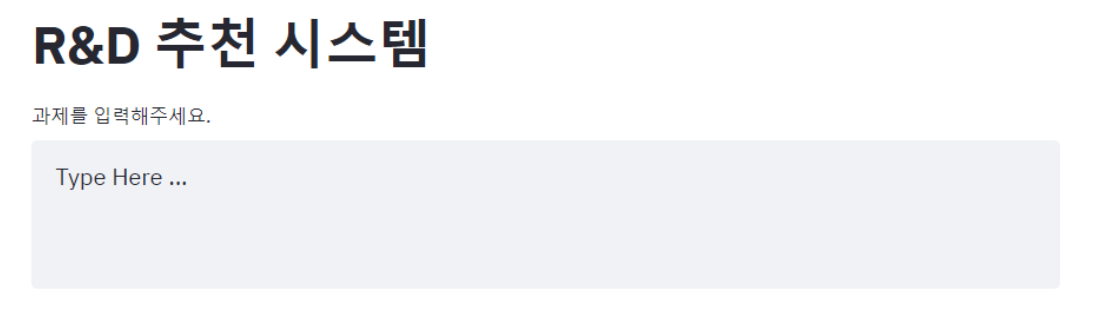
Lee [Nam-young (](mailto:jkll@didim365.com)jkll@didim365.com ) Didim365 Co., Ltd., a researcher at the Corporate Affiliated Research Institute.

Kim [Shin-gu](mailto:ksg@didim365.com) (ksg@didim365.com ) Didim365, a researcher at the Corporate Research Institute.

Hwang Mi-young (popoo@didim365.com ) Didim365, a researcher at the Corporate Affiliated Research Institute.

* Major R&D achievements

This study focuses on the development of an artificial intelligence model that searches and recommends announcements through keywords so that SMEs can easily participate in R&D government support projects. It is designed to improve participation in R&D support projects.



[Figure 14] User input of R&D government-supported project recommendation system

테이블이(가) 표시된 사진

자동 생성된 설명

[Figure 14] R&D Government Support Project Recommendation Output

When you enter interests, sentences, or keywords in the input box of the R&D government-supported project recommendation system, the program operates in real time to refine the user's input and send it to the locally stored artificial intelligence model to calculate document similarity. When the document similarity measurement is completed, the scores are stored in each data row and sorted in high order. After that, 10 announcements that have not passed the deadline and 10 of the total announcements are sent to the UI to show them to the user. If you click on the link, you can go to the appropriate announcement.

The process operates in an environment of the Intel i5-10400, 16GB RAM, and GTX 1650, and takes an average of 3 seconds for the user to see the results. Using high-end cloud computing could also be reduced to an average speed of 1 second. The recommendation results were much more satisfactory than the results of searching for keywords on Google or Naver, not the R&D government-supported project recommendation system.

* Limitations of Research and Future Research Directions

First, the amount of data to be used for artificial intelligence model learning was about 6,000 cases, making it difficult to improve prediction performance. In particular, even if the ongoing announcement is less than the closed announcement, the recommended result value was completely different from the keyword except for one to three announcements.

Second, I think it will be popular with more users if a service is implemented to help them select words and express ideas that are advantageous when filling out support documents, not only recommending announcements. Even if a public announcement is recommended, most of the project participants are said to be on the decline because they cannot fill out proper.

Nevertheless, for most keywords, it was found to have a significant impact, such as recommending satisfactory R&D government support project announcements. This means that it has resolved unnecessary search and wire inquiries and has become a technology that helps small and medium-sized companies commercialize their ideas. In order to ensure the accuracy of the research results, it is necessary to make efforts to conduct future research, such as collecting more public announcement data.



1. research evidence

* estimated cost of research

TODO: Register at the Management Support Office

Documents certifying the current status of the organization and employees of the dedicated department and the qualifications of research personnel.

Statement of Salary Payment for Research Personnel, etc.

Specification of purchase of samples, parts, raw materials, reagents used in R&D work

Other specifications related to research and human resource development expenses

1. References
2. Introduction to Natural Language Processing Using Deep Learning
3. News-Simpler 1nsidewill Github

https://github.com/1nsidewill/News-Simpler