Journal of the Korea Academia-Industrial cooperation Society
Vol. 25, No. 8 pp. 623-627, 2024

https://doi.org/10.5762/KAIS.2024.25.8.623
ISSN 1975-4701 / eISSN 2288-4688

# 생물다양성 정보 제공을 위한 인공지능 챗봇 시스템 구축

**장동석, 홍윤식***
**인천대학교 컴퓨터공학과**

# Development of an AI-Based Chatbot System for Providing Biodiversity Information

Dong-Seok Jang, Youn-Sik Hong*
Dept. of Computer Science & Engineering Incheon National University

**요 약** 생물다양성은 생태계의 건강과 지속 가능성에 중요한 요소로, 생물종의 다양성과 분포에 대한 정보는 생태계 보호 및 관리에 필수적이며, 한반도는 다양한 기후와 지형적 특성 덕분에 생물다양성이 풍부한 지역 중 하나이다. 이에 따라 한반도의 생물다양성을 체계적으로 관리하고 보호하기 위한 데이터 통합 및 분석 시스템의 필요성이 대두되고 있으며, 본 연구는 환경부 국립생물자원관에서 운영 중인 한반도 생물다양성 웹사이트에서 제공하는 데이터를 전처리하고, 인공지능(AI)[1] 기반의 챗봇 시스템을 구축하여 사용자에게 효율적인 정보 검색을 제공하는 것을 연구 목표로 한다. 수집된 데이터는 TF-IDF(Vectorizer)를 사용하여 벡터화한 후, 코사인 유사도(Cosine Similarity)를 활용하여 사용자 질문과 가장 유사한 결과를 도출하며, 챗봇은 OpenAI의 GPT-3.5-turbo 모델을 활용하여 높은 수준의 대화 능력을 갖추고 있으며, 추가적으로 로컬에서 Sentence-Transformer[1] 모델(all-MiniLM-L6-v2)[2]을 사용하여 임베딩을 생성함으로써 과금 및 입력 제한을 예방하였다. 또한, Streamlit과 Streamlit-chat을 활용하여 사용자 친화적인 인터페이스를 구현하였다. 본 논문에서는 챗봇[3] 시스템의 개발 과정, 데이터 전처리 방법, TF-IDF 벡터화 및 코사인 유사도를 이용한 유사도 계산 방법 등을 상세히 설명하며, 실제 구현된 챗봇 시스템의 성능을 평가한다.

**Abstract** Biodiversity is a crucial factor for the health and sustainability of ecosystems. Information on species diversity and distribution is essential for conservation and management of ecosystems, and the Korean Peninsula with its diverse climate and geographic features is a region rich in biodiversity. Consequently, there is a growing need for an integrated data management and analysis system to systematically manage and protect the biodiversity of the Korean Peninsula. This study preprocesses data from the Korea Biodiversity Information System (KBIS) operated by the National Institute of Biological Resources to develop an AI-based chatbot to facilitate efficient information retrieval. The collected data are vectorized using term frequency-inverse document frequency (TF-IDF), and cosine similarity is utilized to derive the most relevant results to user queries. The chatbot employs OpenAI's GPT-3.5 Turbo to provide high-level conversational capabilities. Additionally, local use of the all-MiniLM-L6-v2 sentence-transformers model generates embeddings to prevent cost and input limitations. Furthermore, a user-friendly interface is implemented using Streamlit and Streamlit-chat. This paper details the development process of the chatbot system, data preprocessing methods, TF-IDF vectorization, and the use of cosine similarity for similarity calculations, and evaluates the performance of the implemented chatbot system.

**Keywords :** Artificial Intelligence, Biodiversity, Transformer, Embedding, Chat-bot System

*Corresponding Author : Youn-Sik Hong(Incheon National Univ.)
email: yshong@inu.ac.kr
Received May 20, 2024 Revised June 25, 2024
Accepted August 2, 2024 Published August 31, 2024

# 1. 서론

## 1.1 연구 배경

### 1.1.1 생물다양성이란?

생물다양성은 생태계의 건강과 지속 가능성에 중요한 요소로, 지구상의 모든 생물체의 다양성과 복잡성을 나타낸다. 이는 인간의 생존과 직결된 문제로, 생물다양성의 감소는 생태계 서비스의 악화 및 경제적 손실을 초래한다. 한반도는 다양한 기후와 지형적 특성으로 인해 풍부한 생물다양성을 자랑하나, 도시화, 산업화, 기후변화 등 여러 요인으로 인해 생물다양성이 위협받고 있으며, 이를 보호하고 관리하기 위한 체계적인 접근이 요구되고 있다.

### 1.1.2 연구 목적 및 필요성

1.1.1에서 정의한 생물다양성과 관련하여, 방대한 양의 생물다양성 정보를 효율적 검색·활용하기 위해 발전된 정보 제공 시스템의 필요성이 대두되고 있다. 인공지능 기술을 적용하여 이러한 문제를 해결할 수 있는 방법을 제시하며, 본 연구에서 적용한 기법은 인공지능 기반[1]의 챗봇 시스템 구축을 통해 사용자가 자연어 형태로 질문을 하더라도, 방대한 데이터베이스를 바탕으로 적절한 답변을 제공함으로써 정보 접근성을 향상시키는데 목적이 있다. 최종적으로 본 연구의 주요 목적은 환경부 국립생물자원관에서 운영 중인 한반도의 생물다양성 웹서비스의 데이터를 전처리하고 데이터셋을 구축하여 인공지능 기반의 챗봇 시스템을 통해 사용자에게 효율적인 정보 검색을 제공하고 생물다양성 정보의 효율적 관리 및 활용을 도모하고 생태계 보호 및 관리에 기여하고자 한다.

## 1.2 관련연구

### 1.2.1 한반도의 생물다양성

한반도의 생물다양성 웹서비스는 환경부 국립생물자원관에서 운영하는 정보제공 플랫폼으로,연구자·일반 사용자 들에게 한반도의 풍부한 생물다양성 정보를 제공하며, 생물다양성에 대한 이해를 높이고 보존 활동에 참여할 수 있도록 한다.

주요 기능으로는 생물다양성 DB 구축·관리, 생물종 분포 및 특성 정보 제공, 생물종 검색, 교육 및 홍보 자료 제공, 연구 및 정책 지원 등이 있으며 한반도의 생물다양성 웹서비스는 정확하고 신뢰할 수 있는 정보를 제공함으로써 생물다양성 보존을 위한 중추적인 역할을 수행하고 있다. 이를 통해 생태계의 건강과 지속 가능성을 유지하는 데 기여한다.

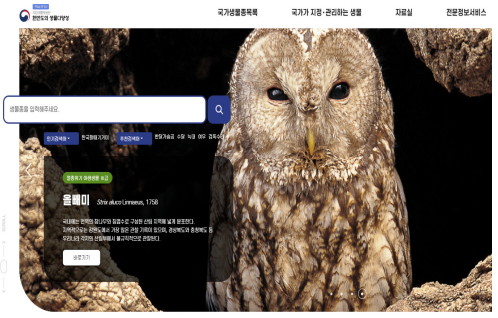


Fig. 1. Web Service Page : https://species.nibr.go.kr

### 1.2.2 인공지능 기반 챗봇

인공지능 기반 챗봇은 자연어 처리 기술[1]을 활용하여 사용자와 상호작용하는 소프트웨어 애플리케이션이다. 본 연구에서는 한반도의 생물다양성 웹서비스에서 제공하는 데이터를 효율적으로 검색하고 제공하기 위해 AI 챗봇 시스템을 개발하였으며, 이 챗봇은 사용자 질문에 대해 신속하고 정확한 답변 제공과 생물다양성 정보의 접근성을 높이고 활용도를 극대화하는 것을 목표로 한다.

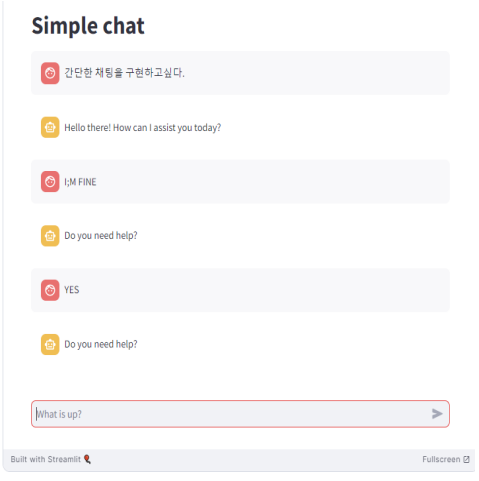


Fig. 2. Example : Streamlit Chatbot application

챗봇의 기능·처리 절차는 자연어 처리 텍스트 임베딩, 유사도 계산을 통해 사용자 친화적인 UI를 제공한다. 데이터 전처리 과정을 통해 텍스트를 임베딩으로 변환한

후, 인공지능 모델을 사용하여 이를 벡터로 변환하며, 이후, 데이터베이스 내의 임베딩과의 유사도를 계산하여 가장 관련성이 높은 답변을 선택하는 로직으로 구현하였다. 이러한 AI 기반 챗봇은 다양한 정보를 신속하고 정확하게 제공함으로써 사용자의 의사결정 과정에 중요한 기여를 할 것으로 기대된다.

## 2. 본론

### 2.1 연구 방법

#### 2.1.1 데이터 수집 및 전처리

데이터 수집은 웹사이트 크롤링을 통해 CSV, JSON 형식으로 정보를 1차로 수집하였다. 수집된 정보는 데이터 정제 및 클렌징 작업을 통해서 누락·중복·오류 데이터를 식별하여 수정·제거하였으며, 특수문자, 불필요한 공백 등 제거 후 데이터를 정규화하였다. 정규화된 데이터는 임베딩백터 변환 후 유사도 검색 시 구문오류를 최소화하기 위해 문단 혹은 줄 단위로 변환하여 텍스트 파일과 CSV 형태의 데이터베이스로 구성하였으며, 본 연구에서 전처리 완료 후 생성된 데이터 수는 종목록 정보 60,010정보와 전처리된 데이터 120개를 포함하여 를 포함하여 60,130개이다.

| 파일명 | 입력내용 | embedding |
|---|---|---|
| 1.txt | 국가생물종목록은 우리나라가 관리하고 있거나 | [0.0015523418551310897, |
| 10.txt | 국가생물종목록 구축현황을 보면 조류는 1문 1 | [0.07548536360263824, 0. |
| 100.txt | 멱는 것이 금지되는 야생동물 지정현황으로는 | [0.01904090307652950 3, - |
| 1000.txt | 120000011840 phacellium episphaerium desn | [-0.01987597718834877, 0 |
| 10000.txt | 120000055003 protoopalina hexasoma yagiu | [-0.002654678653925657 3 |
| 10001.txt | 120000055007 goniomonas amphinema larse | [0.04950559884309769, -0 |
| 10002.txt | 120000055008 goniomonas pacifica larsen pa | [0.04241507872939 11, -0.( |
| 10003.txt | 120000055012 bicosoeca conica lemmermanr | [-0.01160509418696165, 0 |
| 10004.txt | 120000059773 cystopteris fragilis l bernh 한들 | [-0.01477186102420091 6, |
| 10005.txt | 120000059776 deparia conilli franch sav m ka | [-0.00569940917193889 6, |
| 10006.txt | 120000034904 eupsilia silla kononenko and a | [-0.00042614230187609 75 |
| 10007.txt | 120000034905 eupsilia strigifera butler 1879 산 | [-0.00555416569113731 4, |
| 10008.txt | 120000034906 eupsilia transversa hufnagel 17 | [0.03610876202583313, -0 |
| 10009.txt | 120000033437 ectropis aigneri prout 1930 연 | [0.01764944568276405 3, - |
| 1001.txt | 120000011841 phacellium stephanandricola h | [-0.01904194802045822, 0 |
| 10010.txt | 120000033438 ectropis crepuscularia denis ar | [-0.00758832227438688 3, |
| 10011.txt | 120000012126 puccinia convolvuli pers castac | [0.00409027142450134 6, - |
| 10012.txt | 120000012127 aspergillus cumulatus d han ki | [-0.02161425533610725, |
| 10013.txt | 120000012128 alternaria peucedani s h yu 20 | [-0.01585014350712299 3, |
| 10014.txt | 120000012129 puccinia crepidis japonicae lin | [0.03512608259916305 5, - |
| 10015.txt | 120000012133 puccinia dioscoreae kom 1899 | [0.00169095105957239 87, |
| 10016.txt | 120000012134 puccinia diplachnicola dietel 1 | [0.01851961761713028, -0 |
| 10017.txt | 120000022361 pseudolimnophila pseudolimn | [-0.03129119798541069, - |
| 10018.txt | 120000022368 rhabdomastix rhabdomastix us | [-0.04357380792498588 6, |
| 10019.txt | 120000022396 dichaetomyia bilbax wiedeman | [0.05127412080764770 5, - |
| 1002.txt | 120000011843 passalora robiniae shear s hug | [-0.03312772512435913, - |
| 10020.txt | 120000037828 enderleinellus tamiasis fahrenh | [0.00793548673339134 22, - |

Fig. 3. Created Embedded Values

#### 2.1.2 Sentence-Transformer 모델

Sentence-Transformer는 Transformer[2] 아키텍처를 기반으로 하며, BERT(Bidirectional Encoder Representations from Transformers)와 같은 사전 학습된 언어 모델을 사용한다. 트랜스포머 모델은 셀프 어텐션(self-attention) 메커니즘을 통해 문장의 문맥을 이해하고, 각 단어의 의미를 벡터로 표현하며 Sentence-Transformer는 BERT, RoBERTa, DistilBERT 등 다양한 사전 학습된 모델을 기반으로 하여 문장 임베딩을 생성한다. 본 연구에서는 'all-MiniLM-L6-v2' 모델을 사용하였으며, 이는 BERT 기반의 경량 모델로 높은 성능과 효율성을 제공한다. 또한 Sentence-Transformer 모델인 'all-MiniLM-L6-v2'를 사용하여 한반도의 생물다양성 웹서비스에서 수집된 텍스트 데이터를 임베딩 벡터로 변환하였다. 생성된 임베딩 벡터는 데이터베이스에 저장되며, AI 기반 챗봇이 사용자 질문에 대해 관련성이 높은 답변을 제공하는 데 사용된다. 이를 통해 사용자는 생물다양성 정보에 쉽게 접근할 수 있으며, 효율적인 검색 및 정보 제공이 가능해진다.

#### 2.1.3 임베딩 생성

임베딩 생성은 텍스트 데이터를 고차원 벡터로 변환하여 의미적 유사성을 비교할 수 있게 하는 과정이다. 본 연구에서는 Sentence-Transformer[2] 모델인 'all-MiniLM-L6-v2'를 사용하여 한반도의 생물다양성 웹서비스에서 수집한 텍스트 데이터를 임베딩 벡터로 변환하였다. 아래는 임베딩 생성 과정을 단계별로 표현하였다.

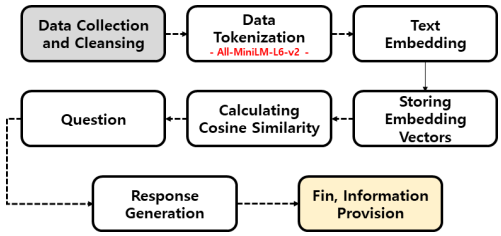


Fig. 4. Embedded generation procedure

#### 2.1.4 유사도 계산

2.1.3 임베딩 생성과 관련하여, 임베딩 벡터를 활용하여 생성된 임베딩 간의 유사도를 벡터 간의 거리나 각도를 계산하여 측정할 수 있다. 유사도 계산방법은 다양하지만 본 연구에서 사용한 임베딩 벡터 간의 유사도를 측정하는데 사용한 방법은 코사인 유사도(Cosine Similarity)를 설정하여 구현하였다. 코사인 유사도는 두 벡터 간의 코사인 각도를 계산하여 유사도를 측정하는 방법이며, 벡터의 크기는 무시하고 방향만 고려하므로, 텍스트 데이터의 유사도 계산에 많이 사용된다. 즉 코사인 유사도를 나타내는 수식은 다음과 같이 표현할 수 있다.

$$Cosine\ \ Similarity(A,B) = \frac{A \cdot B}{\| A \| \ \| B \|} \qquad (1)$$

$A$ and $B$ are embedding vectors.
$A \cdot B$ is the dot product.
$\|A\|$ and $\|B\|$ are the magnitudes of $A$ and $B$
여기서 $A$와 $B$는 각각 두 임베딩 벡터를 나타내며, 코사인 유사도의 값은 -1에서 1 사이의 범위를 가지며, 1에 가까울수록 두 벡터의 방향이 유사하고, 0에 가까울수록 유사도가 낮으며, -1에 가까울수록 반대방향임을 나타낸다는 것을 알 수 있다. 이 수식을 통해 벡터 간의 각도를 구하고, 이를 통해 두 벡터의 방향이 얼마나 유사한지를 측정할 수 있으며, 이러한 과정을 통해 코사인 유사도는 텍스트 데이터 간의 유사도를 정량적으로 비교할 수 있는 강력한 도구로 활용된다.

### 2.2 챗봇 시스템 구현

#### 2.2.1 OpenAI GPT-3.5-turbo Model

OpenAI의 GPT-3.5-turbo 모델은 OpenAI가 개발한 고도화된 자연어 처리(NLP) 모델로, 다양한 언어 이해 및 생성 작업을 수행할 수 있도록 설계되었으며, GPT-3.5-turbo는 이전 버전인 GPT-3의 개선된 버전으로, 더 나은 성능과 효율성을 제공한다. 모델의 기본 구조는 Transformer 아키텍처에 기반한 모델이며, Transformer는 자연어 처리 모델에서 널리 사용되는 구조로, 특히 언어 모델링 작업에서 뛰어난 성능을 발휘한다. 본 논문에서 적용한 Transformer 모델은 수십억 개의 매개변수로 구성되어 있고, 대규모 데이터 셋을 통해 학습되어 다양한 언어 작업에서 높은 성능을 발휘한다. 주요 특징으로는 대규모 데이터 학습, 뛰어난 문맥 이해 능력, 효율성, 그리고 다양한 언어지원이 대표적이며, 본 연구에 적용한 챗봇 시스템 구축[3-6]과 같이 자연스러운 대화 생성과 사용자 문의에 대한 신속한 응답을 제공한다.

#### 2.2.2 Streamlit을 활용한 UI 구현

본 연구에서는 자연어 처리(NLP) 모델과의 상호작용을 위한 사용자 인터페이스(UI)를 구현하기 위해 Streamlit과 Streamlit-chat을 활용[7]하였으며, Streamlit은 데이터 애플리케이션을 간편하게 개발할 수 있는 오픈 소스 파이썬 라이브러리이다.

Streamlit-chat은 streamlit과 함께 챗봇 인터페이스를 간단히 추가하는 라이브러리이며 Streamlit은 빠르고 직관적인 데이터 애플리케이션 개발을 지원하는 도구로 다음과 같은 특징을 갖는다. 첫 번째 간편한 설치와 사용이 가장 대표적인 특징이다. 파이썬 환경에서 간단히 설치하여 사용할 수 있으며, API를 활용하여 사용자가 쉽게 인터페이스를 구축[3]할 수 있고, 두 번째 실시간 업데이트에 따른 코드 변경시 애플리케이션이 자동으로 업데이트 되며 그 결과를 기다릴 필요 없이 실시간으로 결과확인이 가능하다.

## 3. 결론

본 연구에서는 한반도 생물다양성 정보를 효율적으로 제공하기 위한 인공지능(AI) 기반 챗봇 시스템[3-6]을 개발하였다. 이를 통해 사용자는 자연어 질문을 통해 생물다양성 관련 정보를 신속하고 정확하게 검색할 수 있다. Fig. 5에서는 최종적으로 구현된 챗봇 시스템[3-6]의 실제 실행 결과를 보여주며, 사용자가 입력한 질문에 대해 챗봇은 관련성 높은 답변을 제공하였으며, 이는 국립생물자원관의 데이터를 기반으로 하여 신뢰성과 정확성을 갖추고 있음을 확인하였다. 또한 챗봇의 대화 능력과 정보 제공 능력을 통해 사용자 친화적인 인터페이스가 구현되었음을 확인하였고, 인공지능 모델의 API를 활용하여 추가 검색정보를 사용자에게 제공하여 부족한 정보 전달을 보완함으로써 웹서비스의 목적을 달성하였다.

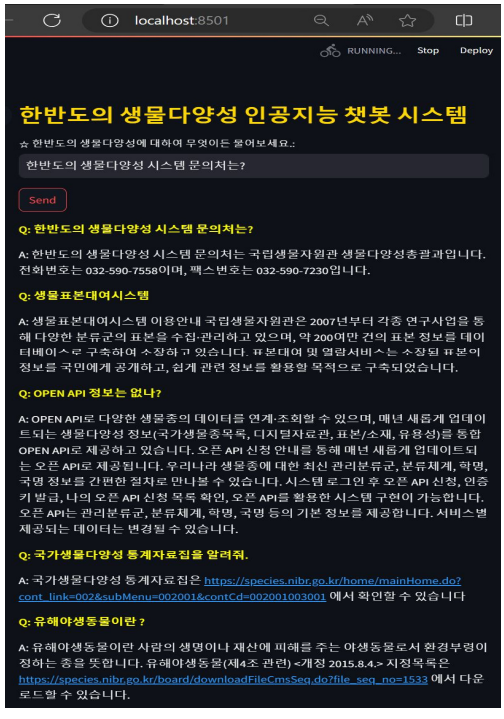

Fig. 5. AI-Based Chatbot System

Fig. 6에서는 임베딩 벡터의 시각화를 PCA 그래프로 모델의 성능을 표현 및 검증하였다. 사용자 입력(파란색 점)과 가장 유사한 답변(빨간색 점)의 임베딩 벡터를 주성분분석(PCA)를 통해 2차원 공간에 시각화하였으며, 각 사용자 입력과 가장 유사한 답변이 시각적으로 근접해 있음을 확인할 수 있고 이는 코사인 유사도를 활용한 챗봇 시스템의 유사도 계산이 효과적으로 작동함을 검증한 것이다.

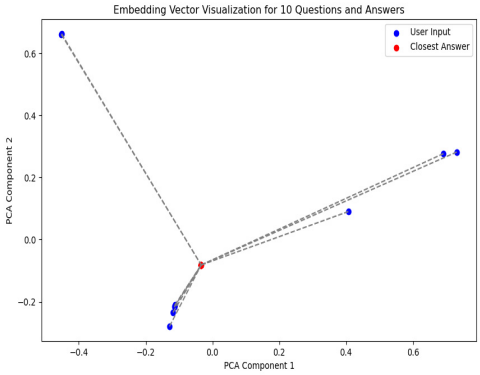


Fig. 6. Embedding Vector Visualization for 10 Q&A

최종적으로 본 연구에서 개발한 인공지능 기반의 챗봇 시스템은 생물다양성 정보를 신속 · 정확하게 제공함으로써 사용자의 의사결정 과정에 중요한 기여를 할 것으로 기대한다. 데이터셋 전처리는 이 챗봇 시스템 성능에 중요한 역할을 하며 향후 연구에서는 모델 성능 향상, 다양한 데이터 소스의 통합, 그리고 사용자 피드백 반영을 통해 확장할 수 있을 것으로 사료되며, 이를 통해 생물다양성의 체계적인 관리와 보호에 크게 기여하고자 한다.

## References

[1] G. Caldarini, S. Jaf, and K. McGarry, "A Literature Survey of Recent Advances in Chatbots," Information, vol. 13, no. 1, p. 41, Jan. 2022. DOI: https://doi.org/10.3390/info13010041

[2] J.K. Cage, Learn Deep Learning with 101 Problems: Hugging Face Transformer with PyTorch - Transformer Model Practice that Anyone Can Easily Follow, Ruby Paper, Aug. 2023, Chaps. 8, 10, 12, 14.

[3] S. Raj, "Creating a chatbot with Python: Using natural language processing and machine learning | Complete everything from chatbot design to implementation and deployment in one step," youngjin.com, Nov. 2020. pp29-66

[4] S. Lee, Creating a Chatbot Using the ChatGPT API: Master Python, Prompt Engineering, OpenAI API, Agent, and Vector DB in 5 Days, Hanbit Media, Mar. 2024, pp. 261-272, 282-297.

[5] J.S. Kim, W.J. Yoo, and S.J. Ahn, How to Use the Real ChatGPT API: From ChatGPT API-Based Voice Assistant to KakaoTalk/Telegram Chatbot Production, Langchain Utilization, and Fine Tuning, Wikibooks GAI Series 1, Aug. 2023, pp. 73-74, 235-238, 245-252.

[6] Y. Ogiwara, S. Furukawa, and Y. Choi, Create Your Own ChatGPT with OpenAI API and Python: From ChatGPT Basics to OpenAI API and Service App Creation Using Langchain, Wikibooks Generative AI Programming Series 5, Apr. 2024, pp. 101, 104-125.

[7] https://docs.streamlit.io/ [Internet]

장 동 석(Dong-Seok Jang) [정회원]

- 2022년 2월 : 인천대학교 컴퓨터공학 (공학석사)
- 2022년 9월 ~ 현재 : 인천대학교 컴퓨터공학과 박사과정

〈관심분야〉
인공지능, 가상화, 빅데이터

홍 윤 식(Youn-Sik Hong) [정회원]

- 1983년 2월 : 한양대학교 전자공학과 (공학석사)
- 1989년 2월 : 한국과학기술원 (KAIST) 전기 및 전자공학과 (공학박사)
- 1989년 3월 ~ 1991년 7월 : LG전자(주)우면연구소 선임연구원
- 1998년 3월 ~ 1999년 12월 : LG정통신(주)단말연구소 자문교수
- 1991년 8월 ~ 현재 : 인천대학교 컴퓨터공학부 교수

〈관심분야〉
모바일 컴퓨팅, 사물인터넷, 헬스케어