|  |
| --- |
| 산학협력 프로젝트 수행계획서 |

|  |  |  |  |
| --- | --- | --- | --- |
| **과제명** | AI/ML 기반의 수요예측 | | |
| **협력기관명** | 구글코리아 | **과제멘토** | 정명훈 |
| **책임교수** | 정설영 | **소속** | 컴퓨터학부 |
| **참여인원** | (총 6명) 기업체 1명, 참여교수 1명, 학부과정 4명 | | |
| **수행기간** | 2019.09.02 ∼ 2019.12.01 | **유형** | 중기 |
| **추진배경** |  | | |
| ◯ 기업은 재화나 서비스 등의 여러 가지 매체를 고객의 수요에 맞게 공급하여 수익을 발생시킨다. 그러한 공급은 간단해 보이지만 다음 수요를 예측하지 못해서 수요에 대한 공급이 적어지거나 혹은 반대로 초과해버린다면 여러 가지 문제점(예: 매출 저하 또는 재고 관리 비용 상승)을 야기한다.  ◯ 고객의 수요에 대한 공급을 맞추기 위해서는 기존의 매출, 수요데이터를 이용하여 다음 기간(월, 분기, 연)에 대한 수요를 예측할 수 있으며 이를 통해 기업체의 생산 및 발주를 최적화 할 수 있다. | | | |
| **목표 및 내용** |  | | |
| * 이번 학기 수강중인 ‘데이터베이스’수업의 효율성을 높이고 심도 있는 학습을 유도한다. * 롯데 (한국 데이터의 장점 vs 매출액만 있음 단점) * Walmart (매출액과 함께 날씨, 경기지수, 매장정보 등이 포함) * 다양한 데이터 분석 도구 활용  1. 데이터베이스: Google BigQuery 2. 데이터 정제 도구: Excel or Google Dataprep 3. 시각화: Data Studio 4. 개발도구: Colab(Python) 5. ML Framework: Tensorflow, Keras  * 데이터 분석에 필요한 여러 모델을 이해하고 이를 통해 추후 공모전 참여도 제고한다. * 시계열 예측, Neural Network, Keras, Python 통계관련 패키지(scipy, pandas, numpy 등) * ARIMA 기반 예측, AutoML Tables 기반 예측, DNN 기반 예측   ※ 상세 학습 내용   |  | | --- | | **내용** | | **1.‘오프라인 미팅’으로 9월 20일 구글코리아에서 프로젝트 세부 방향과 구체적인 내용을 정함** | | **2.(AI/ML)/데이터 처리/데이터 분석을 맡을 팀원 분담** | | **3. 정기적인 진행 상태 확인·진행 방향 질문·기술적인 부분 질문** | | **4. 오프라인 회의 1회(예정)** | | **5. GITHUB 활용/협업 프로그램/개발 언어** | | **6. 유튜브로 데이터처리/분석/AI·ML 코드 학습** | | **7. 기타 여비·학습 현황·온라인 회의 등 문서화 작업** | | | | |
| **기대효과** |  | | |
| ◯ 데이터 분석의 기초적인 방법을 습득  ◯ 데이터 분석에 필요한 처리 과정을 이해하고 파이프라인을 구축  ◯ 실제 기업 비즈니스에 도움이 되는 수요예측 모델을 개발하여 생산, 발주 비용을 최적화 | | | |

**※ 프로젝트 깃허브(GitHub) 주소(URL)**

[**https://github.com/GoldenEastLight/AI\_ML\_Project**](https://github.com/GoldenEastLight/AI_ML_Project)

**1. 과제 목적 및 필요성**

**<선정 배경>**

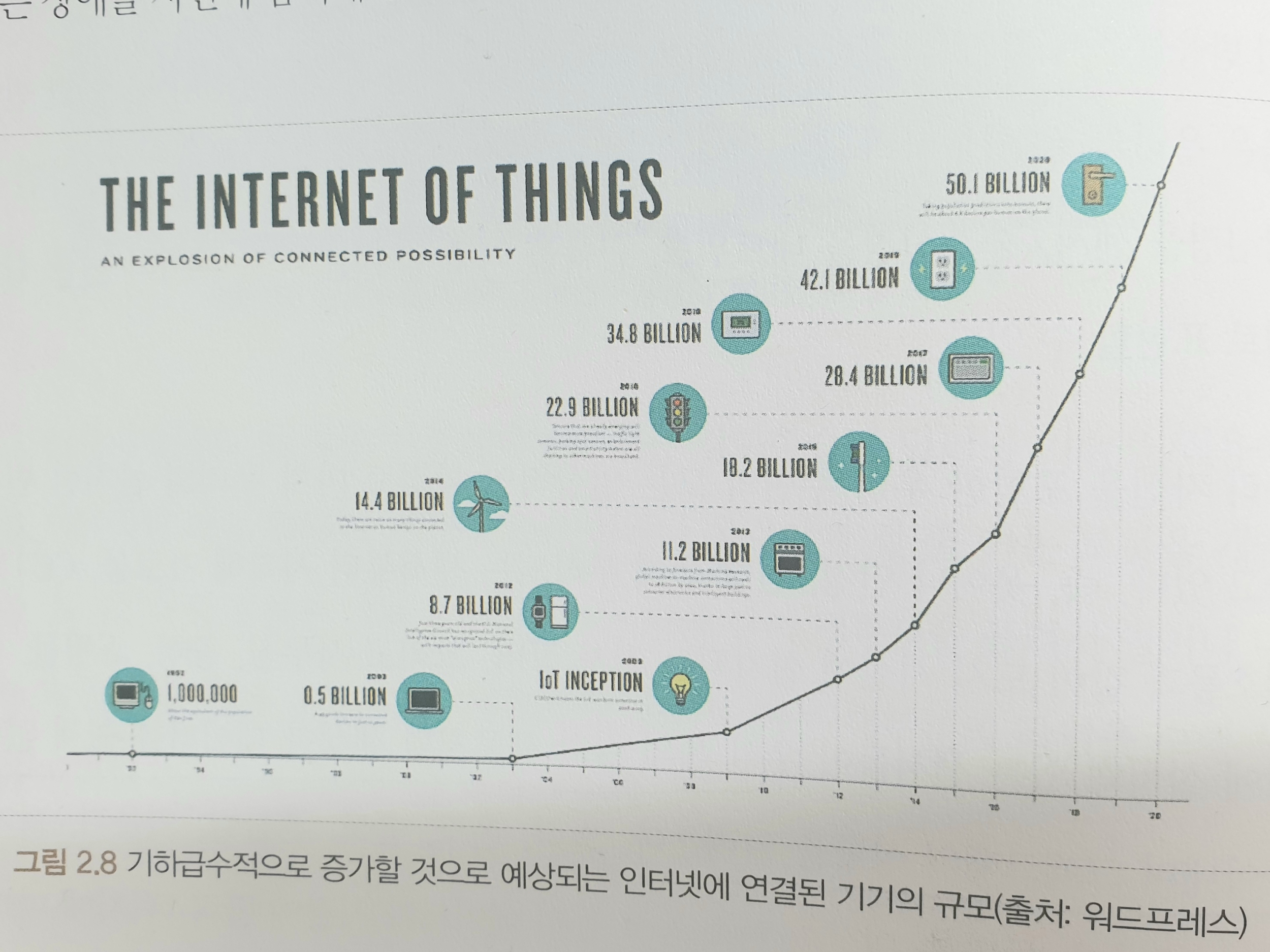
4차 산업혁명 시대가 도래하면서 관심사가 데이터 분석을 어떻게 할 것인가에 대해 높아졌고, 특히 2016년 3월 알파고와 이세돌 간의 대결을 통해 저희 4명의 학부생들도 AI/ML에 대한 관심이 집중되었다. 따라서, 이를 주도하는 구글의 프로젝트를 꼭 맡아보고 싶었고, 기존부터 기업의 시장 수요예측은 주가예측과 동일하게, 만약 우리가 이전의 데이터 자료를 최적의 모델로 완벽히 학습만 할 수 있다면 이는 구태여 이번 프로젝트뿐 아니라, 다른 오픈 데이터 공모전, AI 예측 기반 공모전에도 참여할 수 있는 토대가 될 수 있다고 생각하여 채택하였다.

**<제안 목적>**

1. 우리 팀원의 대다수는 이번 학기에 개설된 ‘데이터베이스’ 수업을 수강 중에 있다. 데이터베이스 수업에 흥미도 돋고, 이를 더 효율적으로 활용해보기 위해 데이터베이스 수업에서 많이 다루지 않는 데이터 분석을 세부적으로 배울 수 있음을 목적으로 하였다.

2. 우리 팀원 중 일부는 2019 AI 우주전파재난에 참가해본 경험이 있다. 현재 프로젝트와 최종 모델은 다르지만, AI 예측 공모전의 특징은 그 방향/기술은 비슷하다는 점에 있다. 따라서 이번 프로젝트를 발판으로 여러 AI 공모전에 대한 참여의식도 제고하며 데이터 분석/처리에 대한 흥미를 높일 목표를 가지고 있다.

**<필요성 기술>**

1. 데이터 처리를 어떻게 할 것인가에 대한 중요성이 시간이 지날수록 점차 더 중요해지고 있다. 특히 기업들은 소비자의 편의와 인건비 절약을 위해 더 많은 마이크로프로세서를 탑재하고, 이러한 작은 컴퓨터를 통신에 의해 더 촘촘히 망을 만들고 이 컴퓨터들은 사용자의 데이터를 쉬지 않고 수집하여, 수집된 데이터는 더 거대한 데이터를 만들게 되고 이는 또다시 ‘패턴’을 만들게 된다. 우리는 이 빅 데이터를 통해 데이터 간 패턴을 파악해보고 규칙을 지정하며 이를 통해 만들어진 모델로 다음 분기/월/년의 물품 수요/공급을 예측하여 실제 오픈 데이터를 제공하는 기업을 (가상) 예측해 보고자 한다.

**2. 과제 내용 및 추진 방법**

**<전체적인 방법·방향 기술>**

1. 롯데마트 데이터 샘플과 월마트 데이터 샘플 자료를 이용

→ 단순히 매출 실적만 기록된 한국 대형마트 데이터 샘플과 매출 기록뿐 아니라 기타 데이터 자료까지 기입된 월마트 데이터 샘플 비교한다.

이를 통해 결과에 어떠한 영향을 주는지 비교해 보고자 한다.

2. 데이터 분석 도구로는 Google BigQuery, Google Dataprep, Data Studio 등을 사용

→ 구글 웹 분석 브라우저에서 지원하는 데이터 처리 플랫폼 BigQuery를 통해 데이터를 테이블화하고 보관하고자 한다. 또한 Google Dataprep를 통해 테이블화된 데이터를 매개변수에 따라 조정하고 이를 각 매개변수의 분포 정도를 Data Studio로 시각화한다.

3. 데이터 분석은 총 2가지 방법으로 분석하고 비교하는 것을 최고 목표로 한다.

3.1 : 시계열 예측 , Neural Network , Keras , Python Tensorflow 패키지 등 ANN 스터디.

3.2 : ARIMA 기반 예측으로 Graph 우선 데이터 예측

3.3 : AutoML Tables(DNN) 기반 예측으로 수치 데이터 예측

3.4 : Google Colab 을 통해 Parameter 조정으로 최적의 수치/Graph 값 도출

→ 3.2 와 3.3 예측 결과 비교와 데이터 샘플 Parameter 간의 차이 조사

11월 한국 방송 미디어 공학회 추계 학술대회 논문에 활용한다.

**<진행 방법·학습 방법>**

1. 멘토님과 메일을 주고받아 ‘오프라인 미팅으로’ 9월 20일 구글코리아 사무실에서 프로젝트 진행의 세부 방향과 구체적인 내용을 정할 예정이다.

→ 대구 - 서울간 여비(교통비) 지원 요청

2. 데이터 베이스(정제화 과정 포함)/데이터 분석(ARIMA, DNN, Colab)을 맡을 팀원 분담

→ 전체 팀원이 3개의 부분을 세부 학습하는 것은 짧은 일정 대비 효율성이 낮을 것으로 보인다. 프로젝트의 전반적 이해를 위해서 각 분담별로 공부는 ‘온라인 자료조사’를 통해 진행하고 데이터 처리/분석/문서·서류 등 은 분담을 통해 진행한다.

3. 정기적인 진행 상태 확인·진행 방향 질문·기술적인 부분 질문

→ 상시로 메일이나 카톡 혹은 화상회의를 통해 멘토님께 현재 프로젝트 진행 방향에서 진행 방향/기술적인 어려움을 극복하기 위해 도움을 받고자 한다.

4. 멘토님과 함께 오프라인 회의 1회(10월 예정)

5. GITHUB 활용/협업 프로그램/개발 언어

→ 프로젝트 진행 정도를 GITHUB를 활용하여 소스코드를 협업하고, 코드를 짜는 과정은 Team Viewer를 통해 협업하여 기술적인 어려움을 극복한다. Graph 기반 ARIMA 웹 브라우저 이용, DNN은 어려운 방식으로는 Python Tensorflow 쉬운 방법으로는 Keras 혹은 기타 DNN 툴 이용

6. 온라인 자료조사를 통해 데이터처리/분석 학습

→ 팀원들 간 1주 정기 온·오프라인 회의(학습 자료 점검) / 멘토님과 상시로 기술적인 어려움, 방향 커뮤니케이션(메일과 google docs 적극 활용)

7. 기타 여비·학습 현황·온라인 회의 등 문서화 작업

→ 담당 분야 자습이 되지 않으면, 전체 프로젝트에 차질을 빚으므로 일주일에 하루씩 정기 회의 외에도 팀원 간 항시 소통한다.

**3. 과제 추진 일정 및 예산 활용 계획**

|  |  |  |  |
| --- | --- | --- | --- |
| 날짜 | 내용 | 필요 예산 | 비고 |
| 9월 4일 | 팀원 1차 정기 회의 |  | 담당업무 분할 |
| 9월 20일 | 멘토 1차 오프라인 회의 | 왕복 교통비 | 구글코리아 사무실 |
| 9월 25일 | 팀원 2차 정기 회의 |  |  |
| 10월 2일 | 팀원 3차 정기 회의 |  |  |
| 10월 9일 | 팀원 4차 정기 회의 |  |  |
| 10월 16일 | 팀원 5차 정기 회의 |  |  |
| 10월 26일 | 멘토 2차 오프라인 회의 |  | 미정 |
| 10월 30일 | 팀원 6차 정기 회의 |  | 논문 요약 최종 |
| 11월 1일 | 한국방송미디어공학회 추계학술대회 요약문 제출 |  | 논문 제출(요약문) |
| 11월 6일 | 팀원 7차 정기 회의 |  |  |
| 11월 13일 | 팀원 8차 정기 회의 |  |  |
| 11월 20일 | 팀원 9차 정기 회의 |  | 논문 최종 수정 |
| 11월 22일 | 한국방송미디어공학회 추계학술대회 논문제출 |  | 최종 논문 제출 |
| 11월 27일 | 팀원 10차 정기 회의 |  | 학술대회 중심 |
| 11월 30일 | 한국방송미디어공학회 추계학술대회 참가 | 왕복 교통비 | 논문 통과  해야 가능 |
| 12월 1일 | 팀원 11차 정기 회의 |  | 최종 회의 |
| 12월 2일 | 종합설계프로젝트 발표 |  | 최종발표 |

*\* 본 일정은 추후 변동 있을 수 있음.*

**4. 기대효과 및 활용방안**

**<교육적 효과>**

현재와 다가올 미래에 있어서 AL/ML 기술은 과학문명의 발전에 빼놓을 수 없는 핵심적인 기술로 자리 잡았다. 프로젝트 수행을 통하여 학부 수업만으로는 습득하기 어려운 AI/ML 기술에 대하여 기초지식 습득과 여러 분야에서 적용되는 방식뿐 아니라, 현재 수강 중인 데이터베이스 과목에서 배우는 데이터 처리에 대해서도 깊은 이해가 가능하다.

**<활용 방안>**

그동안 우리 주변의 사회현상을 예측하기 위해 모집단을 대표할 수 있는 표본을 추출하여 표본에서 생산된 통계량으로 추정해 왔다. 이러한 방식은 기존의 이론 모형이나 연구자가 결정한 모형에 근거하여 예측하기 때문에 제한된 결과만 알 수 있고, 다양한 관계를 파악하는 데는 한계가 있다. 이에 반하여 4차 산업혁명을 맞이한 지금 인공지능을 개발하는 머신러닝 방법이 다양한 변인들의 관계를 보다 정확히 예측할 수 있다. 이번 프로젝트는 AI/ML 기술을 이용한 기업의 수요예측 프로그램이지만 이 프로그램을 통하여 여러 가지 예측 모델을 학습해보고, 각 학습 모델을 이해하며 다양한 데이터 분석에 적합한 모델을 적용해 볼 수 있다.

**5. 예상되는 주요 과제성과**

한국 방송 미디어 공학회 추계 학술대회에 11월 1일 1차 논문제출(요약문)

11월 22일에 최종논문 제출

11월 30일 학술대회에 참가할 예정

※추가적인 시제품, 상용화 계획은 멘토님과 1차 오프라인 회의 후 더 자세히 정할 예정

**6. 참여인력(세부)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 지도교수 | 소속 | 경북대학교 컴퓨터학부 | | 성명 | 정설영 |
| 참여인력  (산업체) | 기업명 | 성명 | 직위 | 전화 | Email |
| 구글코리아 | 정명훈 | 부장 | 010-7287-0851 | javalove93@gmail.com |
| 과  제  참  여  학  생 | 소속(학과) | 학위과정  (성별) | 학번 | 성명 | 담당업무 |
| 생태환경시스템학부 | 학사과정  (남) | 2013088009 | 김동빈 | 총괄 관리 |
| 컴퓨터학부 | 학사과정  (남) | 2015115112 | 김보성 | (정)데이터처리, (부)데이터분석 |
| 컴퓨터학부 | 학사과정  (남) | 2013105086 | 장혁 | (정)데이터분석, (부)데이터처리 |
| 컴퓨터학부 | 학사과정 (남) | 2017111751 | 장현준 | (정)모델테스트, (부)데이터처리 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |