캡스톤 디자인 발표

IT 기술 정보 제공 챗봇

지도교수: 한혁수
팀　　명: K3SI
팀　　장: 김영민 201610980
팀　　원: 김민균 201610974
　　김상우 201610978
　　손성윤 201610999
　　이범희 201611015

목차

1.주제

- 선정 이유
  - 최근 인공지능 기술을 응용하는 대화형 시스템과 챗봇의 개발이 활발
  - 챗봇 등을 통해 고객에게 서비스를 제공하는 메시지 어플리케이션의 활용 증가
  - 비전공자들이 IT 기술에 대한 정보를 찾는 어려움 발생

머신러닝 + 챗봇 ➡ IT 기술 정보 제공 챗봇

1.주제

1 메시지 입력

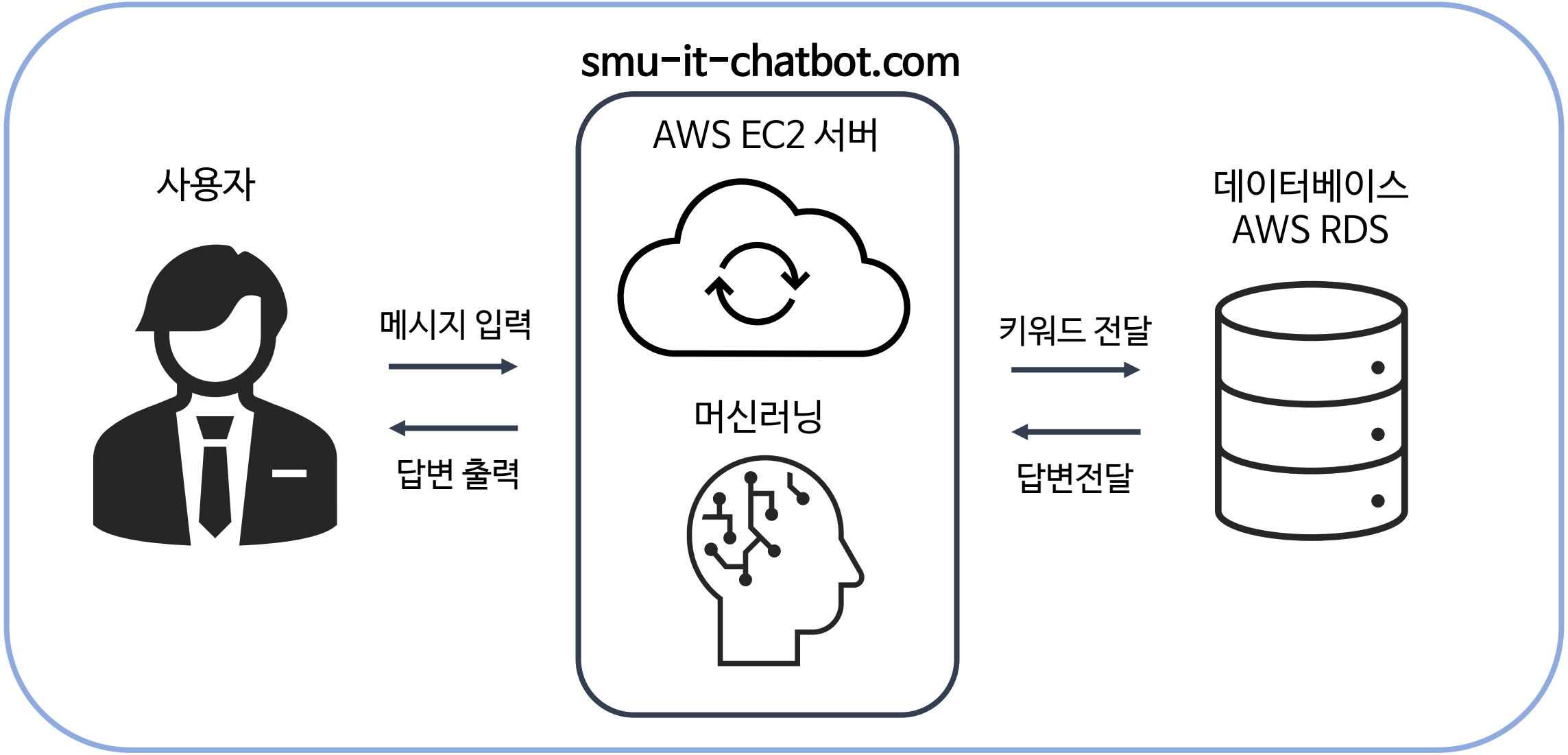
2 챗봇 답변

네트워크가 뭐야?

네트워크에 대한 답변입니다.

지리적으로 떨어져 있는 다른 위치에 있는 장치 간에 정보를 교환할 수 있도록 이들 장치를 상호 …

1.주제


smu-it-chatbot.com
AWS EC2 서버
사용자
메시지 입력
답변 출력
머신러닝
키워드 전달
답변전달
데이터베이스
AWS RDS

2. 팀원 소개

이름	역할
김영민	프로젝트 매니저, 백엔드 메인 개발
김민균	백엔드 서브개발, 서버 정보 조사
김상우	프론트엔드 서브개발, 챗봇 정보 조사
손성윤	프론트엔드 메인개발, 데이터베이스 설계, 산출물 작성 및 관리
이범희	머신러닝 개발

3. 프로젝트 구성

- 챗봇 특징
  - 챗봇 동작 원리

1. 메시지 확인
2. 메시지 전처리
3. 의도 분류
4. 개체명 인식

3. 프로젝트 구성

1

메시지 확인

랜섬웨어가 뭐야?

api 정의는?

2

메시지 전처리

랜섬웨어가 뭐야

API 정의는

1. 영어는 대문자로 통일
2. '?' 기호 제거

3. 프로젝트 구성

3

의도 분류

형태소 분석

1. 문장 안의 조사 제거
ex) 을,를,이,가,은,는…

2. 의도 판단

(랜섬웨어) (가) (뭐) (야)

(API) (정의) (는)

(랜섬웨어) (뭐)

(API) (정의)

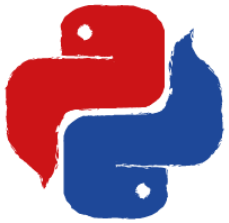
의도 : 질문

3. 프로젝트 구성

3

의도 분류

형태소 분석



KoNLPy

형태소 분석기

~~Komoran~~

~~Twitter~~

~~Hannanum~~

~~Kkma~~

MeCab

3. 프로젝트 구성

3

의도 분류

형태소 분석기

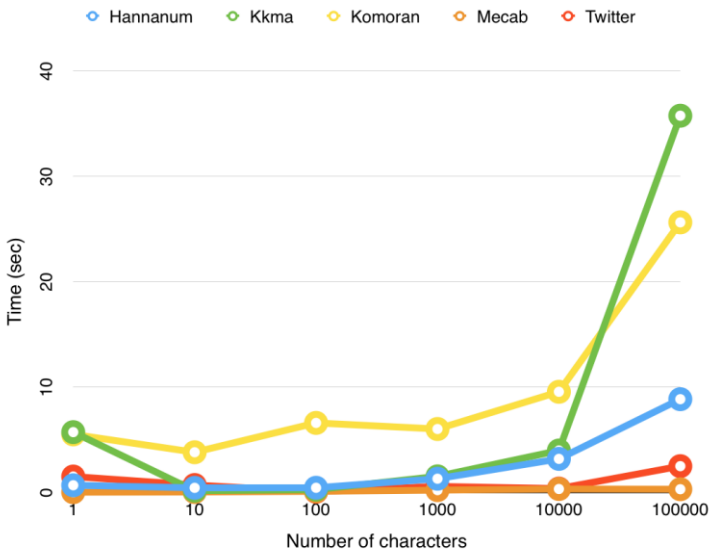
MeCab

Mecab 특징

1. 데이터 크기 증가에 크게 영향을 받지 않는 속도

- Kkma: 35.7163 *secs*
- Komoran: 25.6008 *secs*
- Hannanum: 8.8251 *secs*
- Okt (previous Twitter): 2.4714 *secs*
- Mecab: 0.2838 *secs*

(문장 100,000개를 처리하는 데 걸리는 형태소 분석기별 시간)



2. 사용자 사전을 이용하여 단어 우선순위 조절

```
핀터레스트,,,0,NNP,*,F,*,*,*,*,*,*
플러딩,,,0,NNP,*,T,*,*,*,*,*,*
포직스,,,0,NNP,*,F,*,*,*,*,*,*
인터리빙,,,0,NNP,*,T,*,*,*,*,*,*
스니퍼,,,0,NNP,*,F,*,*,*,*,*,*
파이어웍스,,,0,NNP,*,F,*,*,*,*,*,*
파이어워크,,,0,NNP,*,F,*,*,*,*,*,*
파이어와이어,,,0,NNP,*,F,*,*,*,*,*,*
투피시,,,0,NNP,*,F,*,*,*,*,*,*
투플,,,0,NNP,*,T,*,*,*,*,*,*
매트릭스,,,0,NNP,*,T,*,*,*,*,*,*
```

(사용자 사전의 일부)

→ IT 용어를 최우선으로 인식

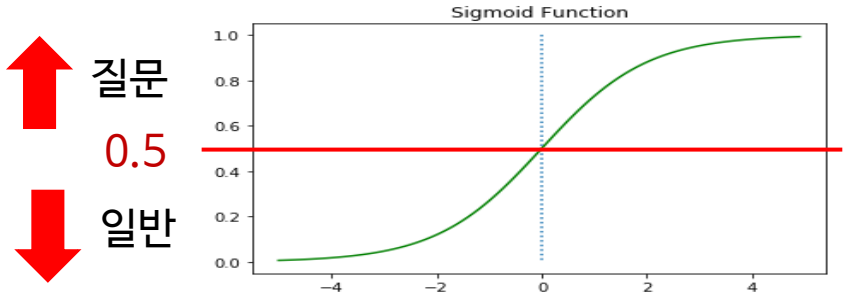
3. 프로젝트 구성

3

의도 분류

의도분류 머신러닝 기법

- 분류 종류: 이중 클래스 분류
- 출력 활성화 함수: sigmoid
- 손실 함수: binary_crossentropy
- 옵티마이저: RMSprop
- 인공 신경망: GRU



(시그모이드 함수)

1	자네는 혹시 신의 한 수를 본 적 있나?
2	내 말이 틀려?
3	크롤링는 어떤건가요?
4	다중 제어 장치 확인
5	에일리어스 검색해주세요
6	색채 해상도가 뭔지 알려주시기 바랍니다
7	하이퍼바이저의 정보를 알려주시기 바랍니다
8	아날로그가 뭔지 답변해줘
9	네, 저희쪽도 슬슬 움직이겠습니다.
10	다중 가상 기억 장치 알아?
11	정말....미안해요. 하지만 오해하지 말아주세요.
12	세션 개시 프로토콜의 정보 확인하고 싶어
13	아이콘의 정보를 확인하고 싶어
14	마이크로컴퓨터 아는지?
15	서명 뭐에요
16	그럼, 지금 당장 전화해.
17	블랙 햇 찾아줘
18	네서스를 설명해주세요
19	멀티 스레드의 정보를 알려줘
20	가자
21	널의 정보 원합니다.
22	스니핑의 정보를 확인하고 싶어요
23	다운로더 확인

대략 23,000개의 문장

(의도 분류 모델을 학습 시키는데 필요한 데이터)

의도 분류 모델 생성

```
4128/5413 [=================>.......] - ETA: 0s - loss: 0.0102 - acc: 0.9985
4384/5413 [==================>......] - ETA: 0s - loss: 0.0096 - acc: 0.9986
4640/5413 [===================>.....] - ETA: 0s - loss: 0.0092 - acc: 0.9987
4928/5413 [====================>...] - ETA: 0s - loss: 0.0090 - acc: 0.9986
5216/5413 [=====================>..] - ETA: 0s - loss: 0.0085 - acc: 0.9987
5413/5413 [==============================] - 1s 237us/sample - loss: 0.0083 - acc: 0.9987

테스트 정확도: 0.9987
```

모델 정확도: 99% 이상

3. 프로젝트 구성

4

개체명 인식

1. 개체명 모델 이용 =〉 개체명 인식

2. IT 용어 태그가 붙은 형태소 이어 붙이기

3. 사용자가 질문한 IT 용어 추출

3. 프로젝트 구성

4

개체명 인식

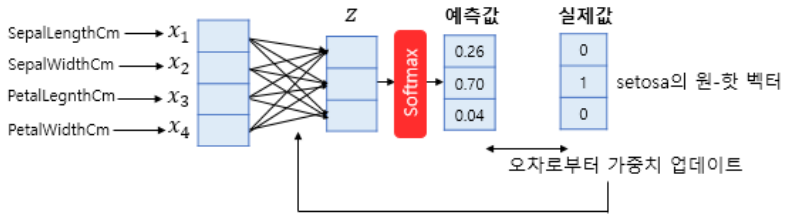
159	sentence: 28	다중	NNG	B_DATA
160		홉	NNG	I_DATA
161		무선	NNG	I_DATA
162		네트워크	NNG	I_DATA
163		정보	NNG	O
164		가르쳐	VV+EC	O
165		주	VX	O
166		세요	EP+EF	O
167	sentence: 29	분산	NNG	B_DATA
168		시스템	NNG	I_DATA
169		의미	NNG	O
170		궁금	XR	O
171		해요	XSA+EC	O
172	sentence: 30	폴링	NNG	B_DATA
173		의미	NNG	O
174		궁금	XR	O
175		해	XSA+EC	O

(개체명 인식 모델을 생성하는데 필요한 데이터)

개체명 인식 - 머신러닝 기법

- 분류 종류: 다중 클래스 분류(BIO 표현)
- 출력 활성화 함수: softmax
- 손실 함수: categorical_crossentropy
- 옵티마이저: Adam
- 인공 신경망: LSTM

대략 13,000개의 문장



(softmax 함수)

개체명 인식 모델 생성

```
 - f1: 98.38
              precision    recall  f1-score   support

       _DATA       0.98      0.98      0.98      2871

   micro avg       0.98      0.98      0.98      2871
   macro avg       0.98      0.98      0.98      2871
weighted avg       0.98      0.98      0.98      2871

f1_score did not improve from 0.984674
```

모델 정확도: 98 % 이상

3. 프로젝트 구성

키워드 : 랜섬웨어

데이터베이스 검색

159	단말 감시 프로그램	Terminal Monitor Program, TMP 시분할 기능(TS...	https://terms.tta.or.kr/d
160	단말 계산기 패드	terminal calculator pad 모든 데이터를 주 컴퓨터...	https://terms.tta.or.kr/d
161	단말 기기	Terminal Equipment, TE ①시스템 또는 통신망의...	https://terms.tta.or.kr/d
162	단말기	terminal ①디지털 자료 전송 시스템에서 자료를 ...	https://terms.tta.or.kr/d
163	라우터	router 근거리통신망(LAN)을 연결해주는 장치...	https://terms.tta.or.kr/d
164	라우팅	Routing LAN등의 데이터 네트워크에서 메시지...	https://terms.tta.or.kr/d
166	랜섬웨어	Ransomware Ransom(몸값)과 Ware(제품)의 합...	https://terms.tta.or.kr/d
168	레지스트리	Registry 윈도즈 95나윈도즈 98및윈도즈 NT와 ...	https://terms.tta.or.kr/d
169	로그	log 시스템 사용에 관련된 전체의 기록, 즉 입출...	https://terms.tta.or.kr/d
170	롤백	roll back 트랜잭션 처리중 시스템에 장애가 발생...	https://terms.tta.or.kr/d
171	리소스	resource 요구되는 연산을 수행하기 위하여 필...	https://terms.tta.or.kr/d
172	릴리즈	release 공식적인 변경 승인 과정과 테스트를 거...	https://terms.tta.or.kr/d

챗봇 답변

3. 프로젝트 구성

- 부가 기능

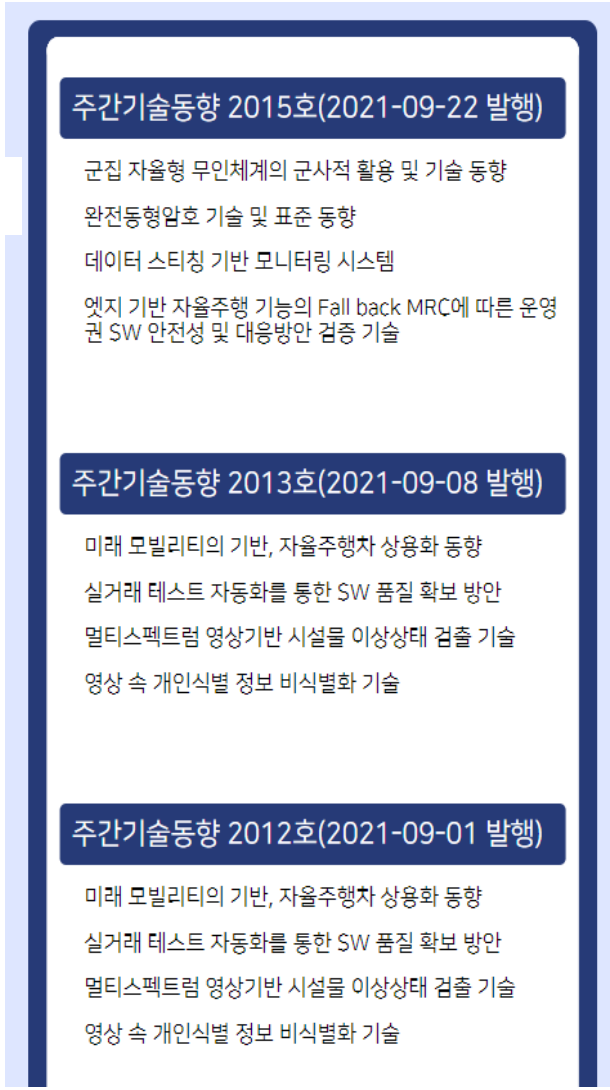
  파이썬 BeautifulSoup를 이용한 웹 크롤링

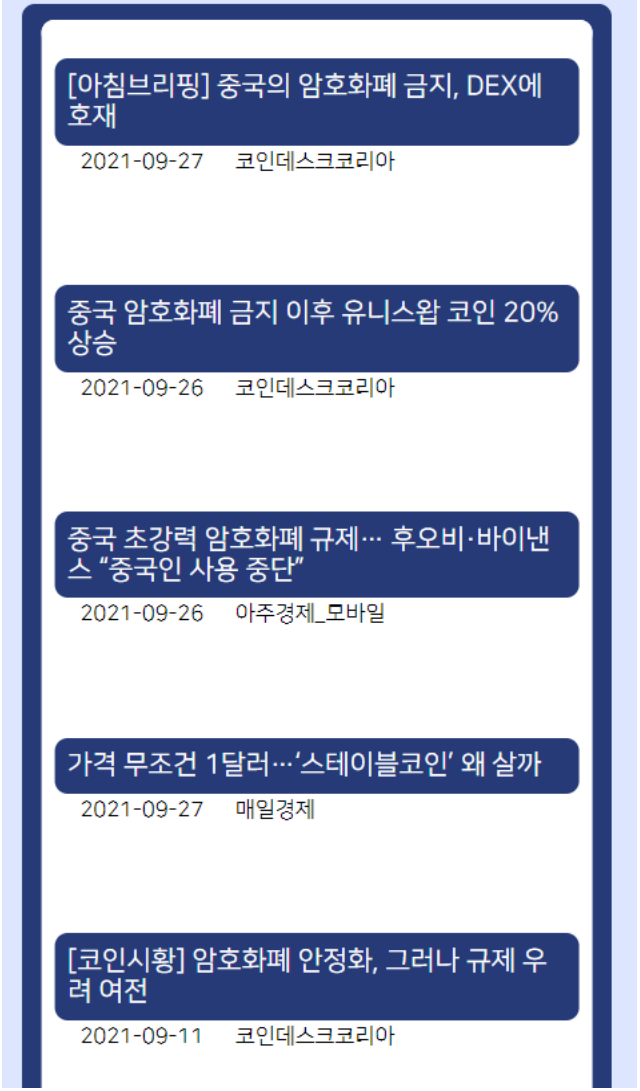
ITFIND

- IT지식 포털 주간 기술 동향

Google news

- 키워드로 구글 뉴스 검색



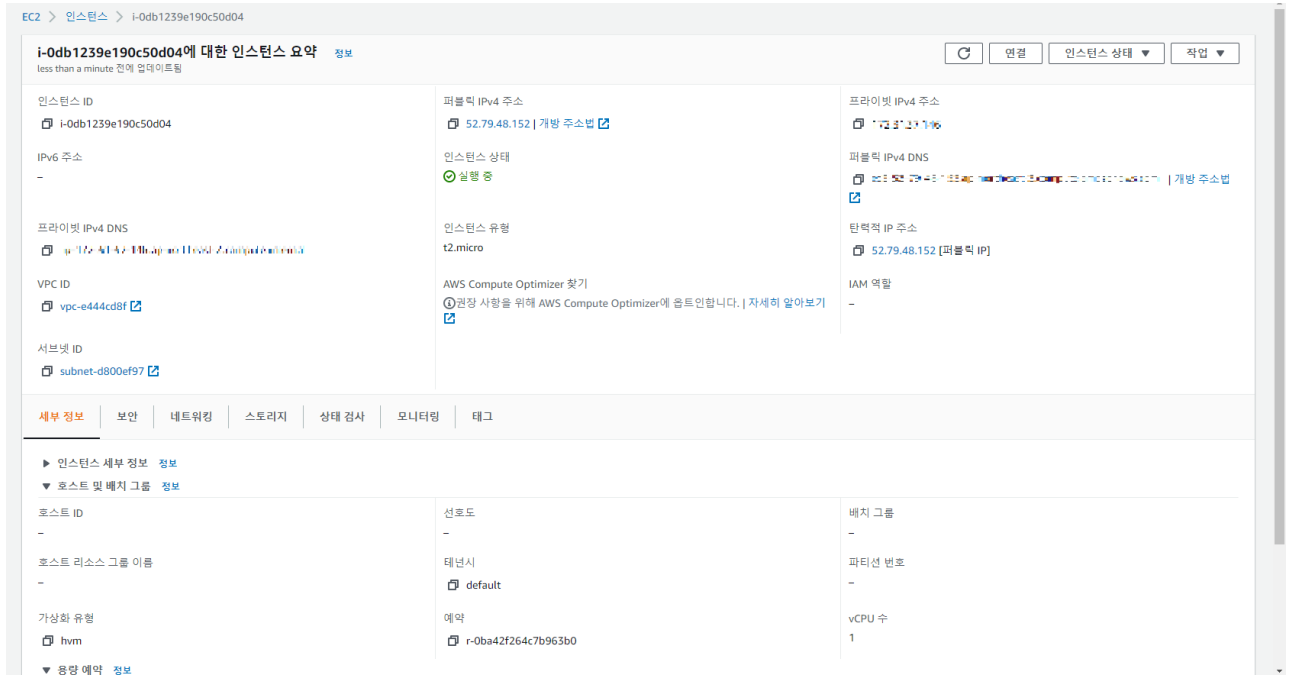


3. 프로젝트 구성

- 시스템 특징

Amazon EC2

AWS EC2(Elastic Compute Cloud)서버 =〉챗봇 서버 운용



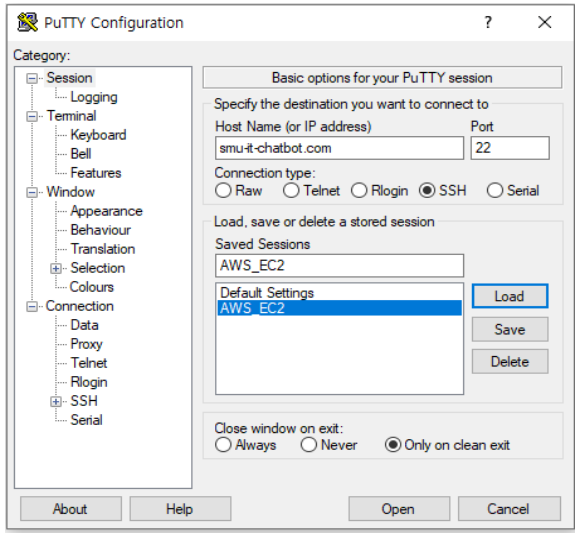
django

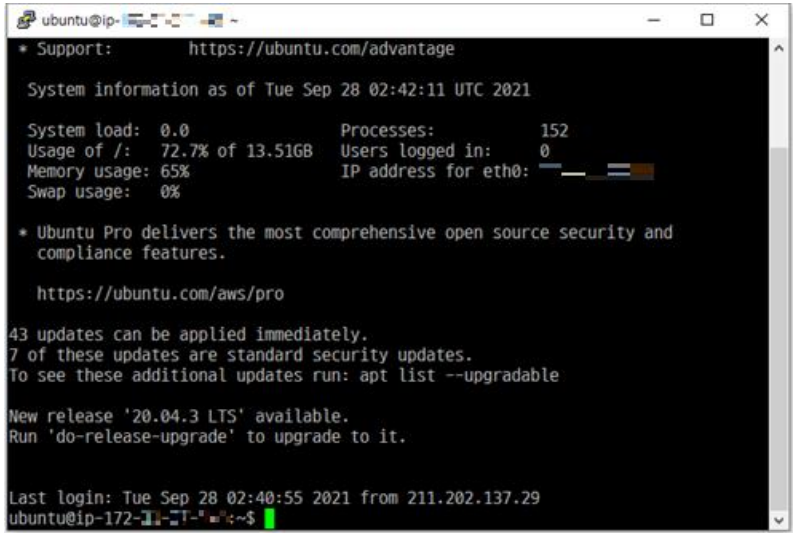
웹 페이지 개발

3. 프로젝트 구성

Putty

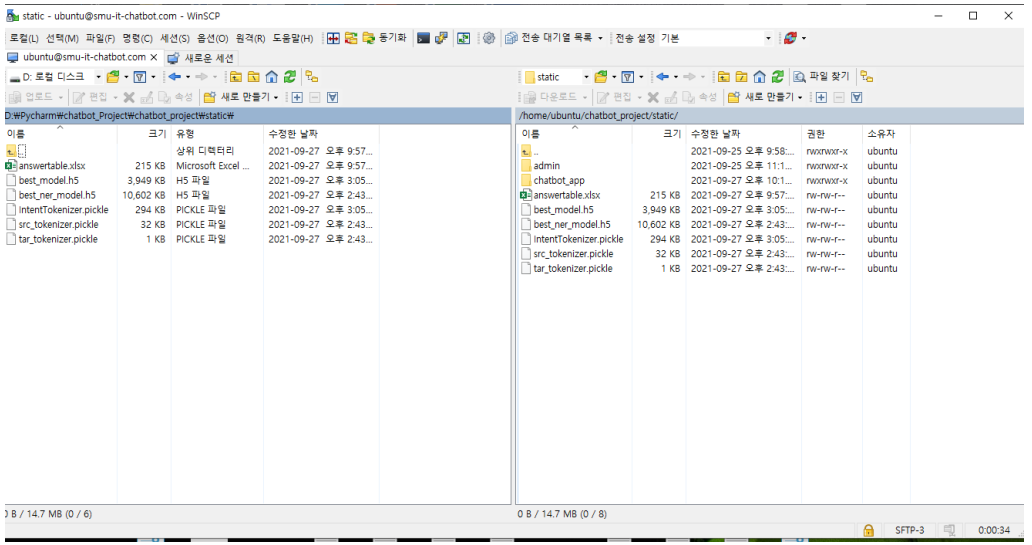
- Putty로 AWS EC2서버 접속
- CLI





WinSCP

- 머신러닝 모델 파일 관리



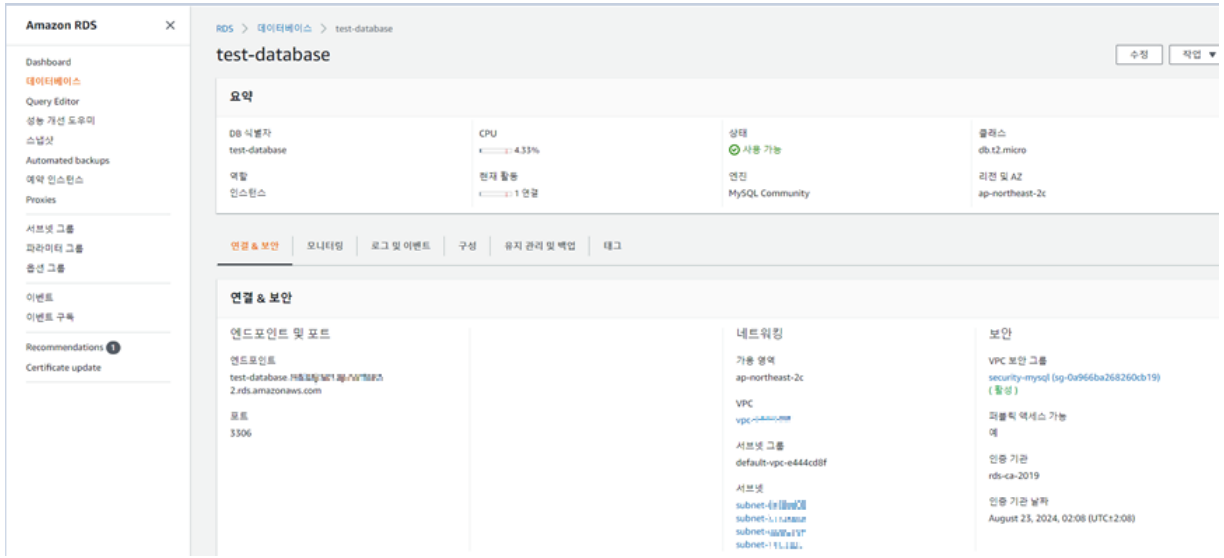
3. 프로젝트 구성

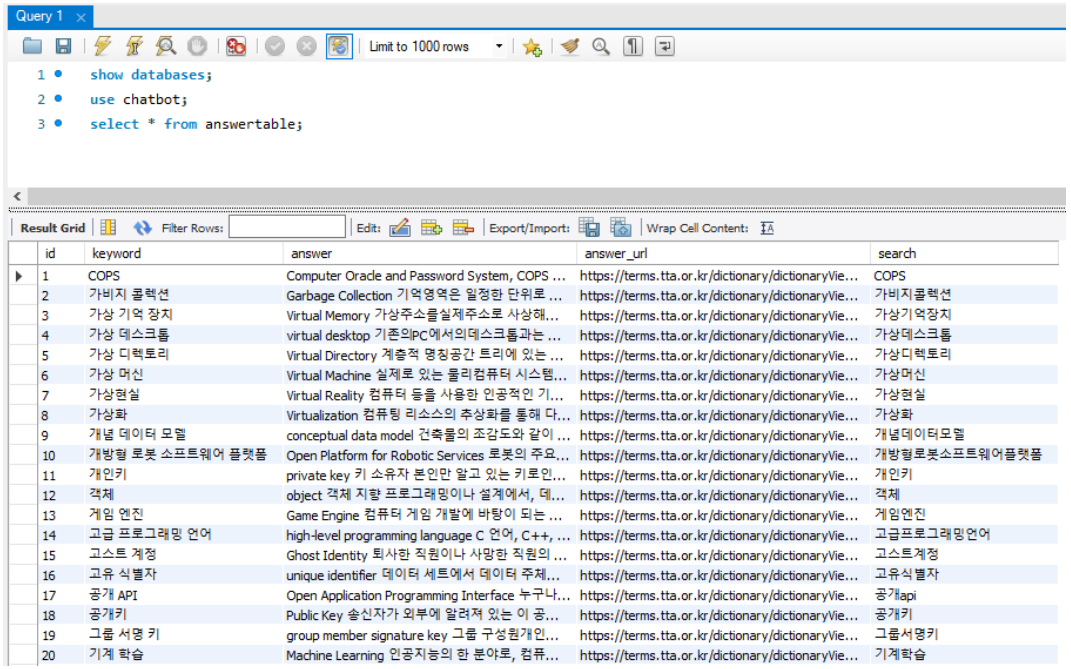
Amazon RDS

AWS RDS (Relational Database Service)
- 데이터베이스 사용

MySQL

MySQL workbench
- 데이터베이스 관리



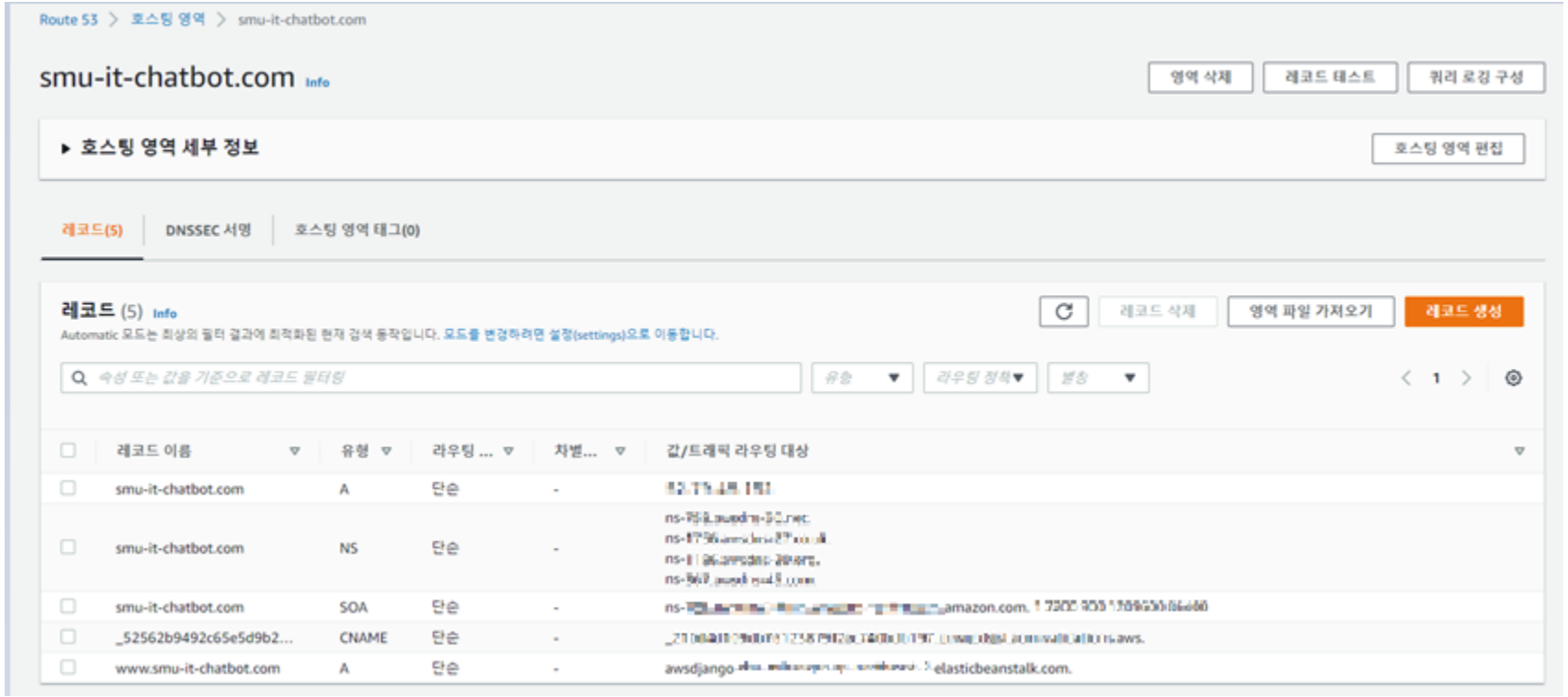


id	keyword	answer	answer_url	search
1	COPS	Computer Oracle and Password System, COPS ...	https://terms.tta.or.kr/dictionary/dictionaryVie...	COPS
2	가비지 콜렉션	Garbage Collection 기억영역은 일정한 단위로 ...	https://terms.tta.or.kr/dictionary/dictionaryVie...	가비지콜렉션
3	가상 기억 장치	Virtual Memory 가상주소를실제주소로 사상해...	https://terms.tta.or.kr/dictionary/dictionaryVie...	가상기억장치
4	가상 데스크톱	virtual desktop 기존의PC에서의데스크톱과는 ...	https://terms.tta.or.kr/dictionary/dictionaryVie...	가상데스크톱
5	가상 디렉토리	Virtual Directory 계층적 명칭공간 트리에 있는 ...	https://terms.tta.or.kr/dictionary/dictionaryVie...	가상디렉토리
6	가상 머신	Virtual Machine 실제로 있는 물리컴퓨터 시스템...	https://terms.tta.or.kr/dictionary/dictionaryVie...	가상머신
7	가상현실	Virtual Reality 컴퓨터 등을 사용한 인공적인 기...	https://terms.tta.or.kr/dictionary/dictionaryVie...	가상현실
8	가상화	Virtualization 컴퓨팅 리소스의 추상화를 통해 다...	https://terms.tta.or.kr/dictionary/dictionaryVie...	가상화
9	개념 데이터 모델	conceptual data model 건축물의 조감도와 같이 ...	https://terms.tta.or.kr/dictionary/dictionaryVie...	개념데이터모델
10	개방형 로봇 소프트웨어 플랫폼	Open Platform for Robotic Services 로봇의 주요...	https://terms.tta.or.kr/dictionary/dictionaryVie...	개방형로봇소프트웨어플랫폼
11	개인키	private key 키 소유자 본인만 알고 있는 키로인...	https://terms.tta.or.kr/dictionary/dictionaryVie...	개인키
12	객체	object 객체 지향 프로그래밍이나 설계에서, 데...	https://terms.tta.or.kr/dictionary/dictionaryVie...	객체
13	게임 엔진	Game Engine 컴퓨터 게임 개발에 바탕이 되는 ...	https://terms.tta.or.kr/dictionary/dictionaryVie...	게임엔진
14	고급 프로그래밍 언어	high-level programming language C 언어, C++, ...	https://terms.tta.or.kr/dictionary/dictionaryVie...	고급프로그래밍언어
15	고스트 계정	Ghost Identity 퇴사한 직원이나 사망한 직원의 ...	https://terms.tta.or.kr/dictionary/dictionaryVie...	고스트계정
16	고유 식별자	unique identifier 데이터 세트에서 데이터 주체...	https://terms.tta.or.kr/dictionary/dictionaryVie...	고유식별자
17	공개 API	Open Application Programming Interface 누구나...	https://terms.tta.or.kr/dictionary/dictionaryVie...	공개api
18	공개키	Public Key 송신자가 외부에 알려져 있는 이 공...	https://terms.tta.or.kr/dictionary/dictionaryVie...	공개키
19	그룹 서명 키	group member signature key 그룹 구성원개인...	https://terms.tta.or.kr/dictionary/dictionaryVie...	그룹서명키
20	기계 학습	Machine Learning 인공지능의 한 분야로, 컴퓨...	https://terms.tta.or.kr/dictionary/dictionaryVie...	기계학습

3. 프로젝트 구성

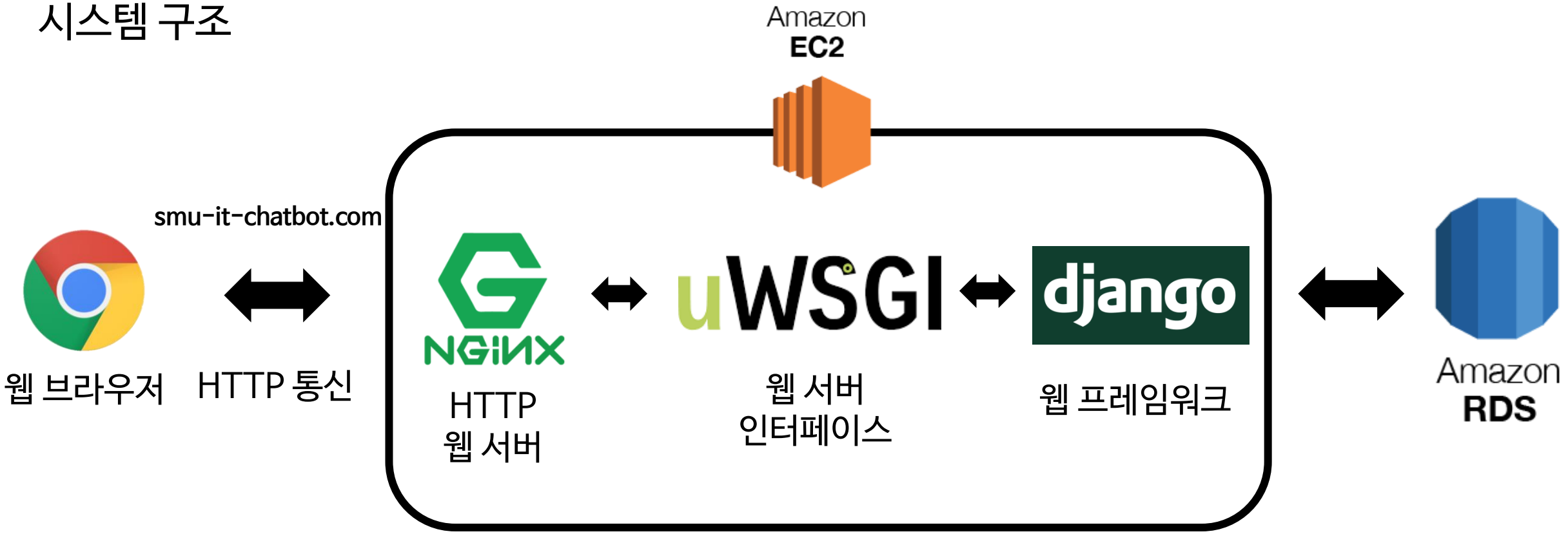
Amazon Route 53

Amazon Route 53을 이용한 도메인 주소 등록



3. 프로젝트 구성

시스템 구조



3. 프로젝트 구성

- 강점

IT 분야 정보 집중

서버와 웹페이지 배포

언제든지 접속 가능

IT Find 주간 기술 동향

구글 뉴스 / 구글 검색

최신 기술 동향 파악 가능

3. 프로젝트 구성

- 제약사항

해결

기본적인 챗봇 기능 – 문답 형식 → 기능 추가 – 주간 기술 동향, 구글 뉴스, 구글 검색

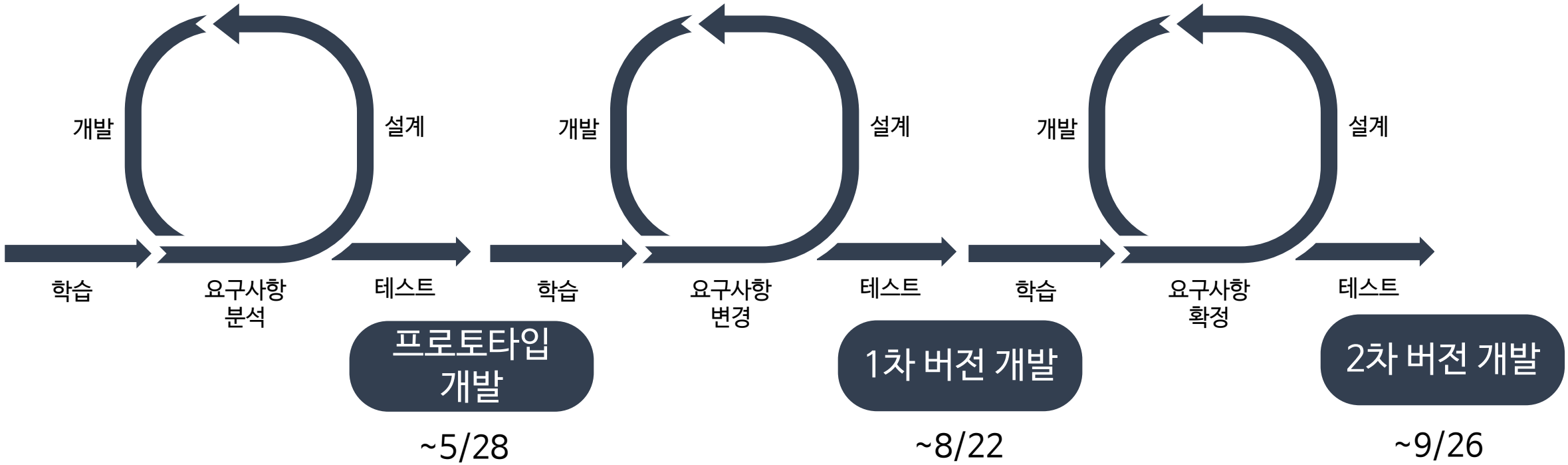
개선 방안

IT 정보에 대해 정의정보만 제공
IT 용어의 정의를 물어보는 질문만 가능
– 질문 형식 한정적
– 구체적인 질문 불가능 → 기계가 인식할 수 있는 의도 종류 추가.
답변에 머신러닝 기술 이용
=〉 다양한 의도에 대해 자연스럽게 답변할 수 있도록

답변할 수 있는 IT 정보의 개수 665개
데이터베이스에 기록된 단어만 제공 → 공인된 IT 정보 사이트 참조.
웹크롤링과 답변을 생성하여 답변테이블 구성

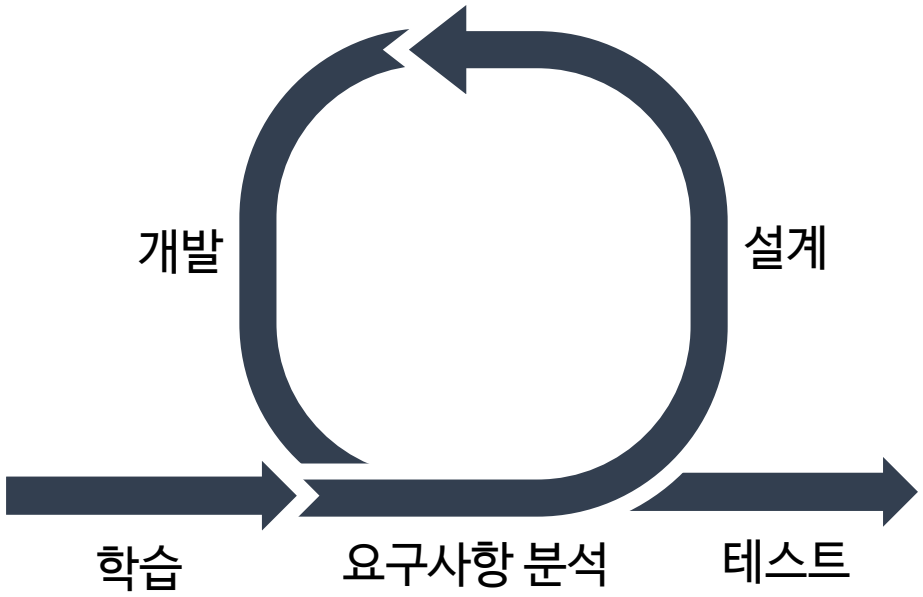
4. 프로젝트 관리

- 라이프 사이클
  - 애자일 방법



4. 프로젝트 관리

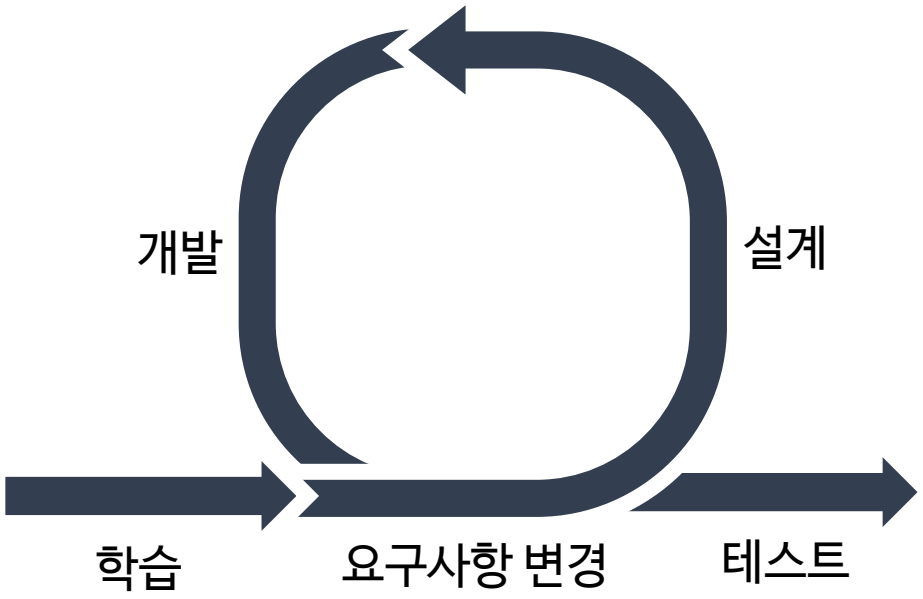
1. 프로토타입 개발



- 기간 : 1/8 ~ 5/28
- 학습 : Python, 챗봇 조사, 머신러닝 이론
- 요구사항 : 사용자 채팅 입력, 챗봇 응답, 머신러닝
- 어려움 : 파이썬 패키지 라이브러리간 버전 충돌 발생
- 해결방안 : 공식문서 확인을 통해 버전을 조율

4. 프로젝트 관리

2. 1차 버전 개발 (장고, 기본적인 챗봇 기능 제공)

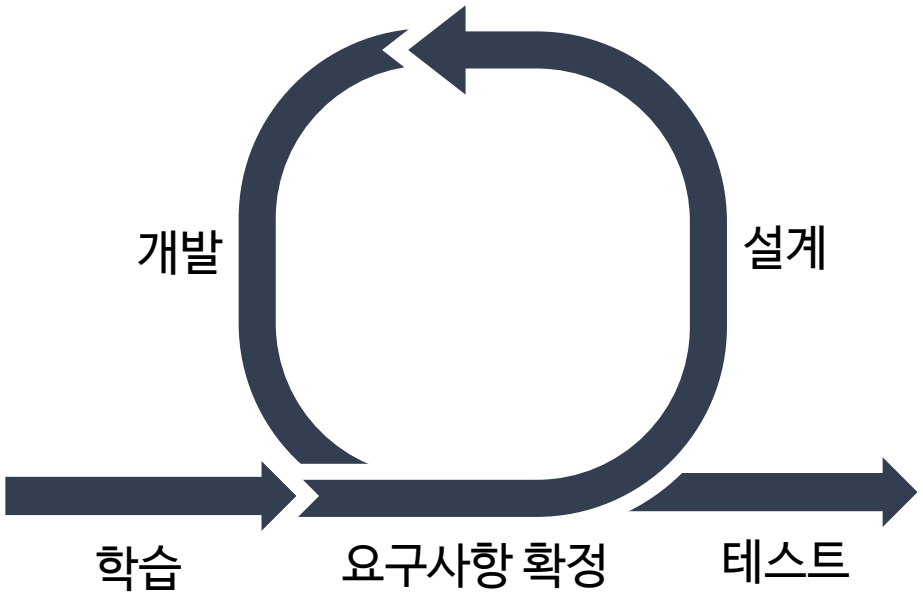


- 기간 : 6/24 ~ 8/22
- 학습 : 웹 개발, AWS EC2, 머신러닝 패키지
- 요구사항 : 머신러닝 정확도 개선,

  AWS EC2 서버를 통한 웹페이지 접속
- 어려움 : 웹서버 운용관련 문제 발생, 데이터베이스 연결 오류
- 해결방안 : AWS 문의 및 공식 문서, 자료조사를 통해서 해결

4. 프로젝트 관리

3. 2차 버전 개발 (부가 기능 및 웹 배포)



- 기간 : 8/23 ~ 9/26

- 학습 : 웹개발, 웹 배포

- 요구사항 : 머신러닝 처리 속도 개선,

  웹크롤링을 통한 주간기술동향, 구글 뉴스 항목 추가,

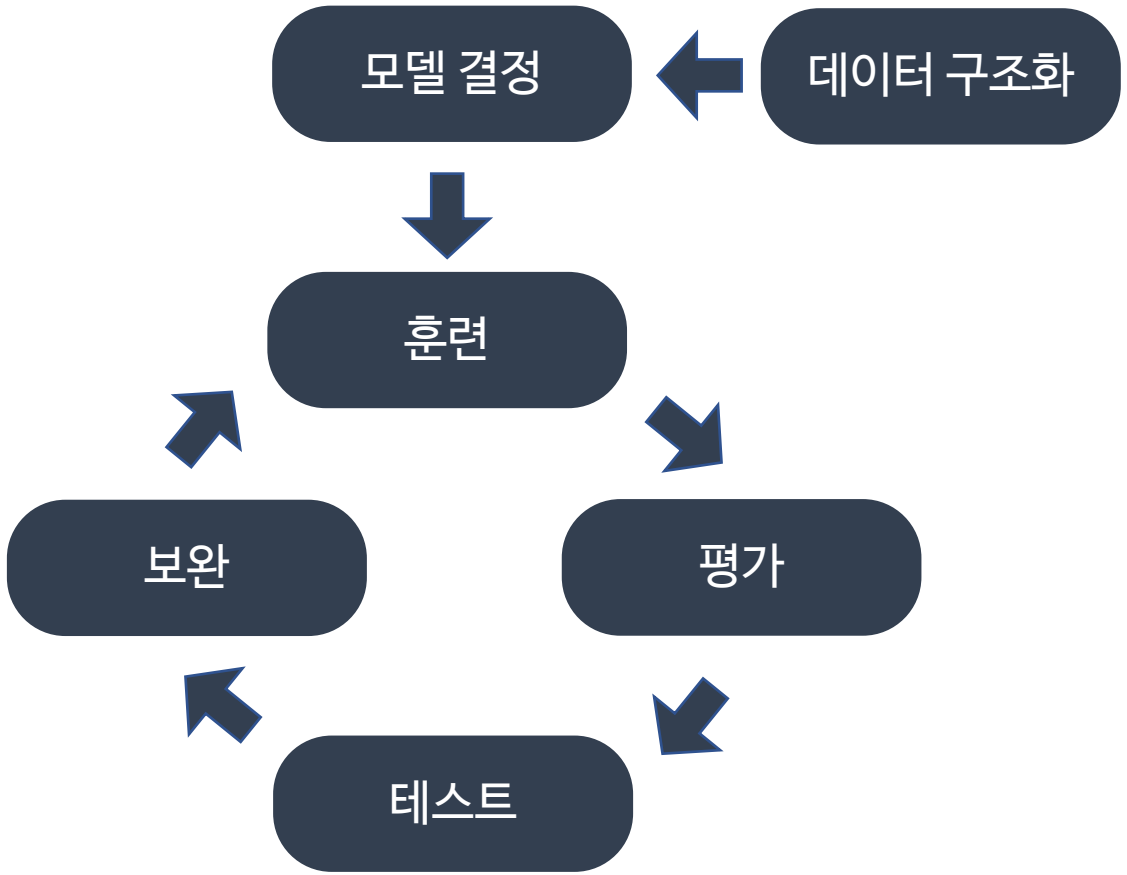
  도메인을 통한 웹페이지 접속, 웹디자인 개선

- 어려움 : 머신러닝 지연 발생, 웹 개발 코드 문제, 팀원 건강문제

- 해결방안 : 추가 학습과 자료조사를 통한 해결 및 웹 개발 언어 숙련

4. 프로젝트 관리

- 머신러닝 라이프 사이클



- 36,000개의 학습 데이터
- IT용어 질문이라는 특정 목적의 코퍼스를 구하는데 어려움 발생

➡ 학습데이터 직접 생성

- 10만개 이상의 데이터가 적절.
- 질문문은 틀이 어느정도 정해져 많이 생성하는데 한계 발생
- 특정 목적을 가진 문장 =〉 특정 형식
- 데이터가 적어도 IT질문을 잘 구별

➡ 새로운 IT 용어 추가

➡ 학습 데이터 크기 증가

5. 시연

시연

https://www.youtube.com/watch?v=dMtnAkEyKXA

6. 결론

- 각 팀원들의 한 마디

김영민	프로그래밍 언어만으로 혼자서 코드를 잘 쓰는 것만으로 끝이 아니라 팀원들과 함께 다양한 지식들을 잘 통합하여 프로젝트를 개발, 관리하는 것을 배울 수 있었습니다.
김민균	팀 프로젝트 경험을 통해 일방적으로 배우는 게 아닌 스스로 알아가고, 또 서로의 지식을 쌓아가는 이상적인 배움을 할 수 있었습니다
김상우	실제로 프로젝트하면서 새로운 언어를 배울 수 있어서 좋은 경험이 되었습니다. 최근 백신접종으로 프로젝트 일정이 밀려 조율하는게 어려웠습니다
손성윤	프로젝트 경험을 통해 평소 접하지 못했던 분야에 대하여 지식을 습득 할 수 있었습니다. 또한 새로운 분야를 시작하는 발판이 되었습니다.
이범희	머신러닝이라는 새로운 분야에 도전해 볼 수 있어서 좋았습니다. 머신러닝, 딥러닝이란 것이 데이터만 넣으면 해결되는 만능이 아니란 것을 알게 되었습니다.

6. 결론

- 정리

프로젝트 경험

- 챗봇, 웹 개발/웹페이지 배포
- AWS EC2 서버
- 머신러닝

문제 해결 능력

- 수많은 시행착오와 에러
- 구글링, 책을 통한 다양한 지식 습득

감사합니다

팀 K3SI