**QnA To KonanBot**

**version 1.0**

**Konan Technology**

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Subject** |
| 1.0 | 2019/12/09 | 신규 작성 (최연식) |

목 차

1. Overview……………………………………….……………………………………………………………………..3
   1. 개요……………………………………………………………………………………………………………..3
2. Services …………………………………………………………………………………………………………..…..4
   1. 실행 …………………….………………………………………………………………………………….…4
   2. 챗봇 메뉴…………………….………………………………………………………………………………5
   3. 프로그램 구성 …………………………….…………………………………………………………..…7

1. Overview

1.1 개요

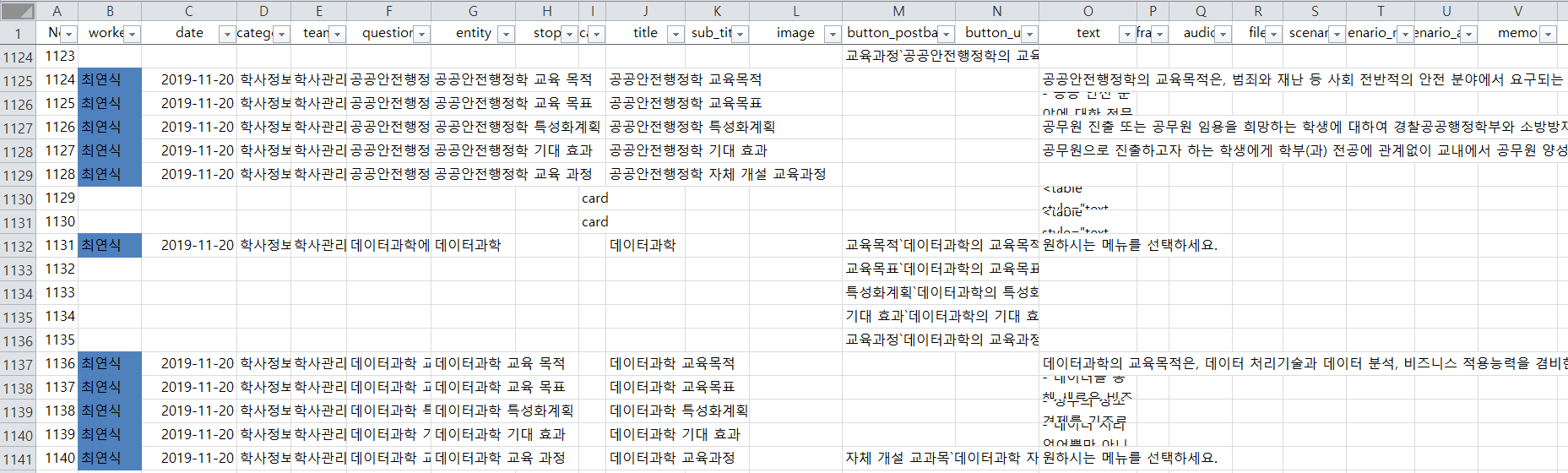
본 문서는 코난봇 데이터 구축 시, QnA데이터를 가지고 Konanbot에 넣을 데이터 형식으로 갖추어주는 프로그램에 대한 명세이다. 중간중간 수동으로 데이터를 검수해야 하므로, 챗봇 형태로 개발하였다. 생성되는 데이터로는, ‘ner-user.txt’, ‘ner-stop.txt’, ‘final\_data\_set.csv’가 있으며, ‘syn-user’의 경우 직접 생성해야 한다.

본 프로그램이 100% 자동화 프로그램은 아니다. 작업 중간에 생성된 데이터를 직접 확인해가며 데이터를 수정해야 한다. 이에 대한 내용은 뒷 부분에서 자세히 설명하도록 하겠다.

본 프로그램은 Linux서버용으로 개발되었으며, window 환경에서 사용할 시 code의 수정이 필요하다.

프로그램을 위한 데이터셋은 다음과 같은 형태여야 한다.

* 입력데이터
  + 열 이름이 No, worker, date, category, team, question, entity, stop, card, title, sub\_title, image, button\_postback, button\_url, text, frame, audio, file, scenario, scenario\_name, scenario\_action, memo와 같이 구성된 파일에 작성한다.
  + 파일 확장자는 '.xlsx’이어야 한다.



< Input data example >

2. Services

* 1. 실행
* 실행 및 결과

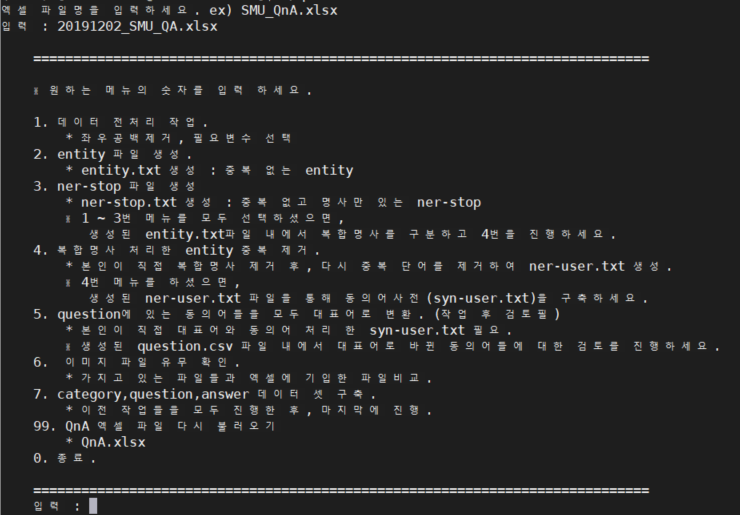
1. Linux 서버 내, /home/konan-isdt/konanbot-util/bin 디렉터리에서 다음을 입력.

* ./qnaToBot.sh



< 입력란에 위의 예시처럼 QnA 데이터를 확장자명까지 입력한다. >

* 1. 챗봇 메뉴



* 총 9개의 메뉴가 있으며, 해당 버튼을 누르면 설명대로 작업이 진행된다. 단, 전처리 작업 및 데이터 불러오기 등 선행작업이 필수적으로 요하므로, 반드시 1번부터 차례대로 실행해야 한다.

1. 데이터 전처리 작업.

* 기본적으로 사용할 변수를 자동적으로 선택하고, 오기입된 공백들을 제거한다. 또한, category와 team을 통해 category#team 형태를 만들어준다.

1. entity.txt 파일 생성

* input data의 entity열의 값들을 모두 가져와 중복을 제거한 후 ‘entity.txt’를 생성한다. 단, 생성된 entity.txt 파일을 통해 복합명사를 직접 구분해야 한다.

수정한 파일은 4번 메뉴에서 다시 사용된다.

1. ner-stop.txt 생성.

* Input data로부터 stop 열의 값들을 모두 가져오고, 형태소분석기를 통해 명사인 단어들만 추출하여 ner-stop.txt로 저장한다.

1. ner-user.txt 생성

* 2번 메뉴에서 생성한 entity.txt 파일에서 복합명사를 처리하고 난 뒤, 다시 한번 더 중복을 제거하여 ner-user.txt 파일을 생성한다.

※ 생성된 ner-user.txt 파일을 토대로 대표어 및 동의어를 정하고 syn-user.txt라는 이름을 가진 파일을 생성해야한다.(5번 메뉴에서 활용)

1. question 대표어 변환

* input data로부터 question 부분에 있는 동의어들을 모두 대표어로 변환시켜준다. 최종적으로 question.csv 파일을 생성한다.

※ 생성된 question.csv 파일에서 기존의 동의어가 포함되어 있던 질문들을 따로 필터를 걸어 검토해야한다.(7번 메뉴에서 활용)

1. 이미지 파일 확인

* Input data에서 image 열의 값들을 모두 불러오고 domain을 image 파일 앞에 붙여준다. 그 뒤, 서버 image 폴더에 올린 domain\_image.jpg or png 파일들과 이름을 비교하여 이미지 파일을 검수한다.(대소문자 구분 X)
* 최종적으로 파일 존재 여부를 출력한다.

1. 최종 데이터 구축

* 이전 작업들을 모두 진행하고 최종적으로 KonanBot에 넣을 input data를 생성한다.

1. 엑셀 파일 다시 불러오기

* QnA셋 데이터를 다시 불러온다.

1. 종료하기.

* 해당 값을 입력 시, 프로그램을 종료한다.

2.2 프로그램 구성

* 디렉터리 구성

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 항목1 | 항목2 | 항목3 | 구분 | 내용 |
| qna\_to\_bot |  |  | 디렉터리 | 최상위 디렉터리 |
|  | input |  | 디렉터리 | 입력 데이터 저장 디렉터리 |
|  | output |  | 디렉터리 | 출력 데이터 저장 디렉터리 |
|  | document |  | 디렉터리 | 사용 설명서 저장 디렉터리 |
|  | morph.py |  | python | 형태소분석을 위한 모듈(참조) |
|  | qnaToBot.py |  | python | QnA데이터로부터  KonanBot에 넣을  데이터를  생성하는 모듈 |

* + qnaToBot python 파일은 프로그램을 실행하기 위한 모듈이다.
  + morph python 파일은 형태소분석을 위한 참조모듈이다.
  + 모듈 속에서 conf 파일을 불러오고 수정을 하는데, 이는 transform\_answer 디렉터리 내에 있는 conf를 참조하는 것이다.
* qnaToBot.py
  + 입력 데이터를 읽고 챗봇 메뉴결에 따라 작업을 진행하며, KonanBot에 넣을 데이터를 생성한다.

\* 자동으로 생성되는 파일 :

entity.txt, ner-user.txt, ner-stop.txt, question.csv, final\_data\_set.csv

\* 직접 생성해야 하는 파일 :

syn-user.txt

* + 해당 모듈의 실행을 위해 pandas, sys, os, collections, configparser 라이브러리의 import가 필요하다.
  + 또한, morph, worked\_file\_to\_rich 모듈의 import가 필요하다.

|  |  |
| --- | --- |
| 전역변수명 | 내용 |
| inputPath | Input data를 불러올 디렉터리를 가리킨다. |
| outputPath | output data를 저장할 디렉터리를 가리킨다. |

|  |  |
| --- | --- |
| 함수명 | 내용 |
| choice(data) | Input data를 입력 받아 필요한 변수만 선택하고, 질문을 기준으로 결측치를 제거 후, 각 셀마다 좌우공백을 제거한다. 또한, 모든 데이터를 문자열 타입으로 변환한다. 결과로 DataFrame 형태로 반환한다.  필요한 변수 : category, team, question, entity, stop  \* image 변수는 데이터를 불러올 때 따로 정제한다. |
| category\_team(data) | Choice 함수에서 정제된 데이터를 입력 받아 category열의 단어 사이에 띄어쓰기가 있을 경우 ‘\_’문자로 변환하고, category와 team을 ‘#’으로 join한다.  Ex) SMU\_story#세명대학교  결과적으로 해당 열만, DataFrame 형태로 반환한다. |
| entity(data) | Choice 함수에서 정제된 데이터를 입력 받아 entity열만 추출한다. 띄어쓰기로 구분된 entity들을 split하고 영문을 모두 소문자로 변환한 한다.  최종적으로 중복 단어를 제거한 후, entity.txt를 생성한다. |
| stopword(data) | Choice 함수에서 정제된 데이터를 입력 받아 stop 열만 추출한다. 결측치를 모두 제거하고, 띄어쓰기와 “ , “로 구분된 단어들까지 모두 split하였으며 중복제거를 위해 모든 영문을 소문자로 변환하였다.  stopword에는 명사만 나타내면 됨으로, morph 모듈의 형태소분석을 통해 명사만 추출하도록 하였다.  최종적으로 ner-stop.txt를 생성한다. |
| set\_word(entity) | 실행 전, entity 함수에서 생성한 entity.txt 파일을 사용자가 직접 복합명사를 처리해서 수정해야 한다.  수정된 entity.txt의 중복 단어들을 제거하여 ner-user.txt를 생성한다. |
| replace\_question(data, syn) | 실행 전, ner-user.txt로부터 사용자가 직접 동의어 사전을 구축하여 syn-user.txt로 output 폴더에 저장하여야 한다.  Choice 함수로부터 정제된 데이터와 syn-user.txt를 불러온다.  syn-user의 동의어들을 모두 key값으로, 대표어들을 모두 value 값으로 갖는 dictionary를 생성한다. 그 dictionary를 통해 question 열의 모든 동의어들을 대표어로 변환한다.  결과파일은 question.csv로 저장되며, dataframe으로 반환한다.  ※ 복합명사 속 글자도 대표어로 변환되는 경우가 있으므로 해당 파일을 열어, 대표어로 바뀌지 않아야 하는 문자들을 직접 검토해서 수정해야 한다.  Ex) 대표어 : 방법 // 동의어 : 방식  한방식품 -> 한방법품(x) |
| replace\_all(text, dic) | replace\_question 함수에 포함되어 있는 sub함수로, 동의어를 대표어로 바꾸기 위한 함수이다. replace\_question 함수의 question들과, 동의어사전을 입력받아, 동의어에서 대표어로 replace된 문자를 반환한다. |
| final\_data\_set(data,question) | Choice 함수로부터 정제된 데이터와 복합명사에 대해 사용자가 직접 검토한 question.csv를 입력데이터로 받는다.  최종적으로 category, question, answer로 구성된 final\_data\_set.csv를 생성한다. answer의 경우, work\_file\_to\_rich 모듈을 통해, 정리된 answer format을 가져온다. 최종적으로, 해당 데이터를 KonanBot에 넣으면 된다. |
| check\_image(image\_df) | Input data에서 image열을 가져온다.  먼저, domain을 입력 받으며, 대소문자 상관없이 해당 값을 대문자로 변환한다. Ex) SMU  서버 내 이미지 파일 디렉터리(/home/konan-isdt/konanbot-data/attachments/images/)내에서 해당 도메인이 들어간 이미지 파일들과, input data의 image열의 파일 이름들과 비교하여 image들의 존재 여부를 확인한다.  단, 서버에 올린 이미지 파일 앞에는 ‘domain\_’이 붙어야 하며 확장자는 소문자여야 한다. 만약 확장자가 대문자일 경우, cmd에서 해당 디렉터리로 이동한 뒤 다음 코드를 입력한다.   * ren \*.PNG \*.png * ren \*.JPG \*.jpg   Ex) SMU\_지원절차.jpg (O) 지원절차.JPG (X) |
| chatbot() | 본 모듈의 main함수 기능을 한다. 모듈 특성상 중간에 사용자가 직접 데이터를 검토하고 수정해야 하므로 챗봇 형태로 개발하였다. |