|  |
| --- |
| **[소프트웨어학과]** |
| [식자재 데이터의 자연어 처리] |
| **[결과보고서]** |
| **Version 1.0** |
| **2020. 06 . 26** |

학번: 201723272

이름: 안창희

지도교수: 최영준 교수

목 차

[요약 3](#_Toc44106663)

[1 개요 4](#_Toc44106664)

[1.1 프로젝트 개요 4](#_Toc44106665)

[1.1.1목표 4](#_Toc44106666)

[2 관련 선행 연구 조사 4](#_Toc44106667)

[2.1 선행 연구 조사 4](#_Toc44106668)

[2.1.1 표준 코드 사전 정의 4](#_Toc44106669)

[2.1.2 딥러닝 기반 자연어 처리 4](#_Toc44106670)

[2.1.3 지향점 4](#_Toc44106671)

[3 연구 결과 5](#_Toc44106672)

[3.1 문제 정의 5](#_Toc44106673)

[3.1.1 기존 문제점 분석 5](#_Toc44106674)

[3.2 제안하는 기법 혹은 소프트웨어 구조 5](#_Toc44106675)

[3.2.1 기존방식 5](#_Toc44106676)

[3.2.2 프로젝트방식 5](#_Toc44106677)

[4 성능 분석 6](#_Toc44106678)

[4.1 성능 분석 환경 6](#_Toc44106679)

[4.2 성능 분석 결과 6](#_Toc44106680)

[5 결론 7](#_Toc44106681)

[6 참고자료 7](#_Toc44106682)

[7 Git 7](#_Toc44106683)

<표 목차>

[표 1 성능분석표 6](#_Toc44106599)

[표 2 참고자료1 7](#_Toc44106600)

[표 3 참고자료2 7](#_Toc44106601)

<그림 목차>

[그림 1 표준화 방식(기존) 5](#_Toc44106567)

[그림 2 표준화 방식(개선방안) 5](#_Toc44106568)

# 요약

The core properties of standard information on food materials were weight and two producers. The regular expression formula, which is operated by extracting the number before the unit, such as kg or ml, was prepared and separated, and the producer name was to be separated only by the producers in the dictionary while updating the manual dictionary.  
The probability of extracting weight information was 91%, or 6,2491 out of the total 68830 codes, while the probability of extracting producer names was 94%, or 63910 codes.  
  
We were able to learn a variety of small and medium business data outside of the standard language processing module.  
The existing standardization data acquisition work was able to improve the repetitive tasks.

# 1 개요

## 1.1 프로젝트 개요

### 1.1.1목표

한가지 상품에 대한 정보를 얻기 위해 업체의 개수만큼 검색을 실행하는 불편함 때문에 이 프로젝트를 시작하였다. 프로젝트의 초안은 6개 표본업체에 대하여 각각 1만품목 내외의 총7만개상당의 원본데이터를 얻고, 그 중에서 동일 품목에 대한 표준정보를 최대한 많이 제시하는 것이다.

# 2 관련 선행 연구 조사

## 2.1 선행 연구 조사

### 2.1.1 표준 코드 사전 정의

표준 코드 정의는 수집된 용어로부터 코드를 선별하여 현 코드의 코드값을 조사한다. 현 코드를 바탕으로 통합 요구 사항과 통합 필요성에 따라 통합 대상을 파악하고 표준 코드를 정의하고 현 코드와 매핑 설계를 한다. 정의된 표준 코드별로 종속 관계를 정의하여 향후 해당 코드에 대한 수정, 삭제에 대한 권한을 관리할 수 있도록 한다.

표준코드는 기존의 코드사용에 대해 다음의 장점을 가진다.

1. 재사용성
2. 일관성
3. 정보 분석성

### 2.1.2 딥러닝 기반 자연어 처리

자연어 처리 분석 단계는 다음과 같다.

1. 형태소 분석
2. 구문 분석
3. 의미 분석
4. 담화 분석

### 2.1.3 지향점

본 프로젝트는 기존의 표준코드 사전 정의 방식과 자연어 처리 방식을 참고하여

식자재 데이터 코드 처리의 어려움을 해소해 보고자 한다.

# 3 연구 결과

## 3.1 문제 정의

### 3.1.1 기존 문제점 분석

본 프로젝트의 발단이 되는 식자재 데이터 처리의 어려움은 다음과 같다.

1. 정보 분석성 부재  
   같은 상품이라 할지라도 분류, 정의 방식이 업체마다 다르기 때문에 식자재   
   유통업체 시장 전체에 대한 분석자료를 내기가 어렵다.
2. 재사용성 부재  
   기업체 특성상 계약을 체결하여 유통사와 식당점주 사이의 데이터운용이 필요한데  
   운용 플랫폼을 구축하는데 있어서 애로사항이 생긴다.

## 3.2 제안하는 기법 혹은 소프트웨어 구조

### 3.2.1 기존방식

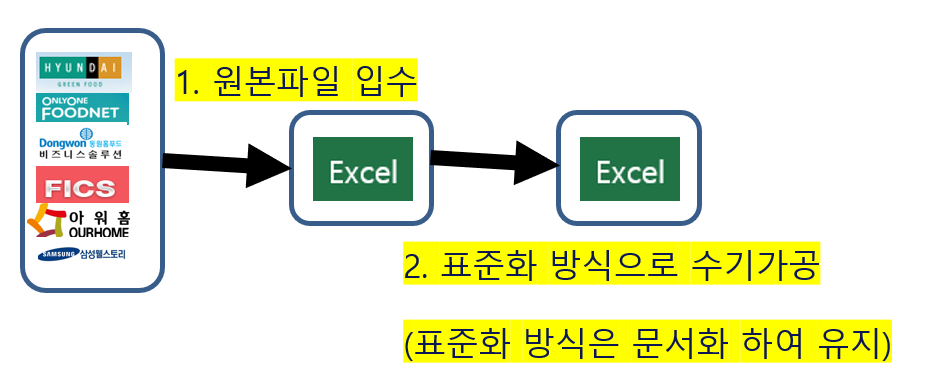


그림 1 표준화 방식(기존)

### 3.2.2 프로젝트방식

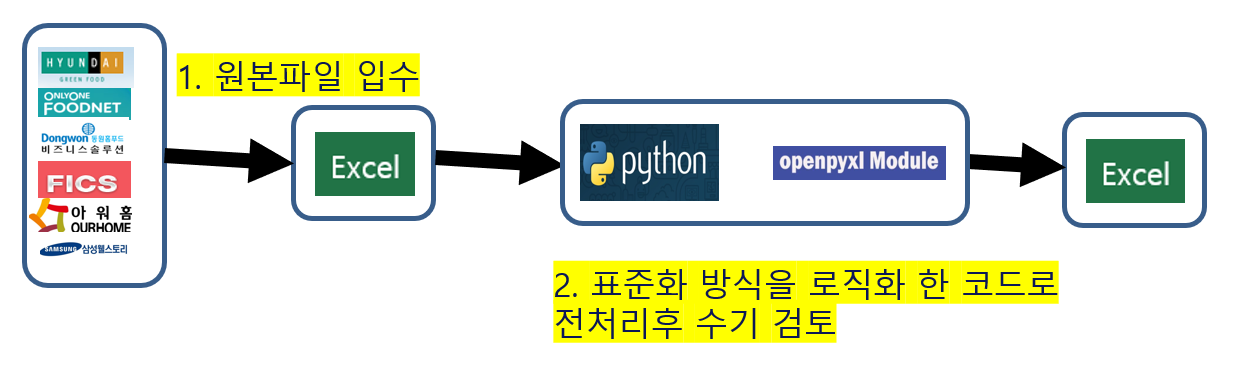


그림 2 표준화 방식(개선방안)

# 4 성능 분석

## 4.1 성능 분석 환경

본 프로젝트의 성능 분석에 쓰여야 할 기준은 전처리된 데이터의 정확성이다.

표준화 데이터의 가장 중요한 두가지 속성인 생산자명과 중량의 추출 확률을 분기별로 조사하여 성능을 분석하였다.

## 4.2 성능 분석 결과

전체 68830개 원본코드를 가지고 작하였다.

최종 매칭확률은 중량은 62491개인 91%였고, 생산자명은 63910개인 94%였다

즉 원본코드 68830개 중에서 약 63000개의 코드는 수기로 작성할 필요 없이

검토만 거치면 회사 내에서 표준화 데이터베이스에 입력해도 되는 값이 되는 것이다.

표 1 성능분석표

|  |  |  |
| --- | --- | --- |
| 주차 | 매칭확률(생산자명) | 매칭확률(중량) |
| 1 | 60% | 80% |
| 2 | 60% | 83% |
| 3 | 60% | 87% |
| 4 | 60% | 88% |
| 5 | 60% | 80% |
| 6 | 60% | 91% |
| 7 | 65% | 91% |
| 8 | 68% | 91% |
| 9 | 73% | 91% |
| 10 | 75% | 91% |
| 11 | 81% | 91% |
| 12 | 90% | 91% |
| 13 | 94% | 91% |

# 5 결론

기존의 업무 방식이 인력을 통해 유지되고 있었으며 본 프로젝트는 이를 개선하는 방향으로

본인 역량에 맞추어 가능한 만큼만 진행하는 것이었으므로 충분히 프로젝트 목표를 만족했다고 볼 수 있다.

또한 현재 회사 내에서 해당 프로젝트를 시발점으로 최종적으로 인공지능 모델을 구축하는

방향으로 나아가기 위한 프로젝트가 기획되었다. 이번 여름방학 인턴활동을 통해 해당 기업체와 다시금 협업할 기회를 가지게 되었으며 앞으로도 본인의 성장을 위해 많은 도움을 얻을것을 기대한다.

# 6 참고자료

표 2 참고자료1

|  |  |
| --- | --- |
| 자료 이름 | 데이터 표준 정의 |
| 자료 유형 | 웹사이트 |
| 자료 일부분 | 표준 코드에는 각 산업별로 법·제도적으로 부여하여 공통적으로 사용되는 코드뿐만 아니라 기업 내부에서 정의하여 사용하는 코드도 포함된다. |
| 저자 | DBGuide.net |
| 규격자료 | IEEE Criteria for Class IE Electric Systems, IEEE Standard 308, 1969. |
| 링크 | http://www.dbguide.net/db.db?cmd=view&boardUid=12 867&boardConfigUid=9&categoryUid=216&boardIdx= 44&boardStep=1 |

표 3 참고자료2

|  |  |
| --- | --- |
| 자료 이름 | 딥러닝 기반 자연어 처리 |
| 자료 유형 | 파워포인트(.ppt) |
| 자료 일부분 | 형태소분석: 문장을 형태소 단위로 분리 |
| 저자 | 강원대학교 이창기 교수 |
| 규격자료 | IEEE Criteria for Class IE Electric Systems, IEEE Standard 308, 1969. |
| 링크 | https://www.slideshare.net/LGCNSairesearch/ss-173607437 |

# 7 Git

https://github.com/leaping96/self\_directed\_project.git