제10회
국가통계방법론 온라인 심포지엄
국가통계의 현재와 미래

온라인가격정보 수집 · 정제에 대한 AI(인공지능) 적용 가능성 검토

김현 · 정영섭 · 윤행근

Ⅰ. 연구배경 및 목적

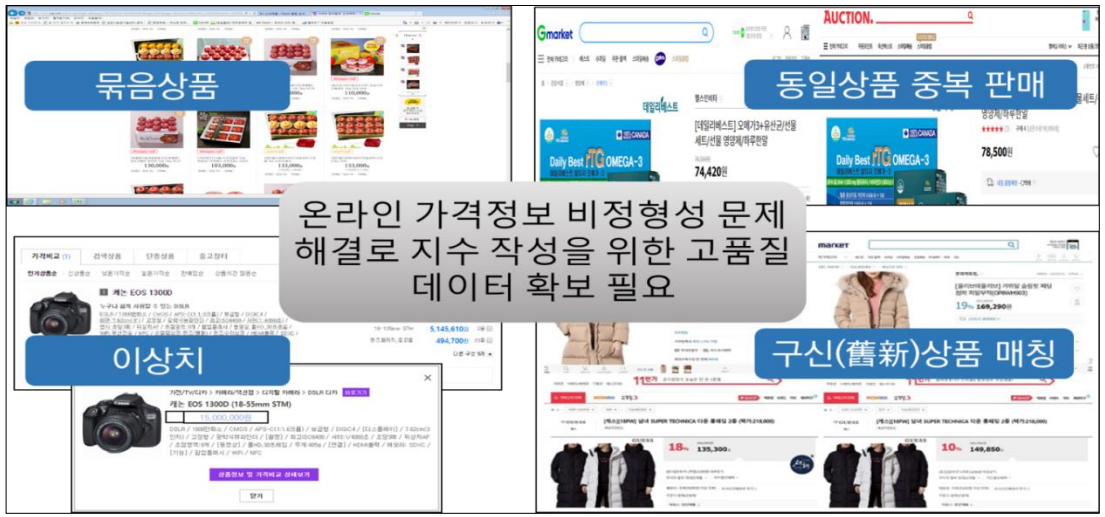
연구배경

- **온라인 쇼핑시장 성장**
  - 국내 온라인쇼핑 거래액은 2019년 기준 약 135.3조원으로 2017년 대비 약 43.6% 증가 (통계청)
  - 소매판매 중 온라인쇼핑 상품거래 비중 : 2017년 16.6% ➔ 2019년 21.4%
- **온라인물가지수를 통한 기존 물가지수 대체·보완 활발**
  - 온라인물가지수 : 온라인쇼핑 사이트에서 수집하여 작성되는 물가지수
  - 미국, 영국, 네덜란드, 독일 등에서 기존 물가지수 대체·보완을 위해 활용
- **통계청의 '일일 물가' 수집 및 작성**
  - '온라인 일일 물가 작성 시스템' 구축('14년 2월)
  - '20년 기준 매일 약 250만개 온라인 가격정보 수집

* 자료 : MIT

* 출처 : 통계청

- **온라인가격정보에 대한 높은 관심에도 비정형성에 의한 데이터 품질 저하는 물가지수 대체·보완의 장애로 작용**
  - 웹스크래핑(web scraping)으로 수집비용은 낮은 반면, 대용량의 온라인 자료 특성상 비정형자료 혼재로 정제에 상당한 시간과 비용 발생



연구목적

- **소비자물가지수 작성을 위한 온라인가격정보 처리 작업 분석 및 AI 적용가능성 평가**
- **온라인가격정보 작업별 AI 적용 방법 및 사례 제시**

Ⅱ. 온라인물가지수 사례 및 관련 동향

주요국의 기존 물가지수 대체·보완 사례

- **미국 노동통계국(BLS)** : 현장조사 외 웹스크래핑 데이터 등 데이터 원천별 지수 산출
- **영국 통계청(ONS)** : 빅데이터 기반 온라인 물가지수 작성
- **네덜란드 통계청(CBS)** : 웹스크래핑, 스캐너 등 데이터원천에 따른 지수 작성
- **독일 통계청(FSO)** : 스캐너, 웹스크랩 데이터 이용 CPI 개발 방법론 연구, 대상 범위 품목 확대 중

AI 적용 검토 및 연구 사례

- **데이터 변환 및 필터링, 품목 분류** 등 주로 비정형정보 정제단계에서 실무 적용 검토 및 연구 진행 중
- **주요 사례** : 미국 노동통계국, 영국 통계청, 네덜란드 통계청

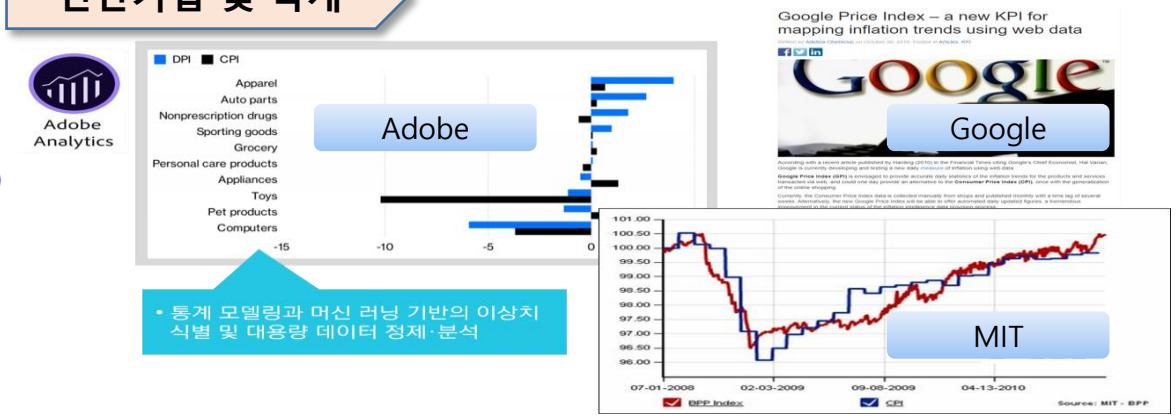
민간기업 및 학계

- **민간기업** : Adobe, Google 등을 중심으로 온라인물가 지수 작성 및 제공
  - **Adobe Analytics** : **통계 모델링과 머신러닝 기반 이상치 식별, 대용량 데이터 정제·분석 기법**을 통해 **온라인물가지수 산출에 필요한 데이터 질을 제고**
- **학계** : BPP를 진행하고 있는 MIT 등이 주도

주요국 사례

주요 내용	AI 연구
• **노동통계국(BLS)** – 현장조사 이외 웹스크래핑 데이터 등 데이터 원천별 지수 산출 및 평가 – 품목 분류를 위한 AI 적용 검증 연구 • **MIT** : BPP(Billion Price Project)를 통해 22개국 온라인물가지수 제공	• 품목 분류(BLS)
• **영국(Office for National Statistics, ONS)** – 빅데이터 기반의 온라인 물가지수 작성 * CLIP : 큰 데이터셋을 군집화하여 물가지수에 반영하는 기법 – 웹스크랩 데이터의 필터링, 분류에 대한 머신러닝 적용 검증 연구	• 필터링: 극단값 및 결측치 • 분류: 분류 오류 식별, 품목 분류
• **네덜란드(CBS)** – 로봇 툴, 웹스크래핑, 스캐너 등 데이터 원천에 따른 지수 작성 – 데이터 변환, 품목 분류 및 평가에 AI(머신러닝, 딥러닝) 적용 검증 연구	• 데이터 변환 • 품목 분류 및 평가
• **독일** – 스캐너 및 웹스크랩 데이터를 이용한 CPI 개발 방법론 연구 – 온라인물가지수 적용 대상 품목 확대 중	• 데이터 변환 및 품목 분류

민간기업 및 학계



ㅍ. AI(인공지능) 개요

AI(인공지능) 개념과 활용사례

- **개념** : 사람의 생각과 관련된 활동, 즉 의사결정, 문제 해결, 학습 등의 활동을 자동화하는 것 (Bellman, 1978)
- **발전과정** : 인지·학습·추론 등 인간지능영역 전과정에서 혁신적 진화를 가져왔고 향후 행동 영역으로 확장 기대
- **활용사례**: 전문가시스템, 자연어처리(Natural Language Processing; NLP), 데이터마이닝, 음성 및 컴퓨터 비젼 등

발전 과정

* 출처 : KISTEP, "인공지능(SW)", KISTEP 기술동향브리프, 2018-16호.

활용 사례

구분	내용	사례
전문가 시스템	- 특정 문제 영역에 대해 전문가 수준의 해법 제공(복잡한 추론, 계산) - 작업 중요도 높은 분야에 추천정보 활용 (원자력, 항공 분야)	의료진단 고장진단 추천시스템
자연어 처리(NLP)	사람이 쓰는 일반 언어로 작성된 문서를 처리하고 이해하는 분야	구문·의미 분석 개체명 인식 기계번역 질의 응답
데이터 마이닝	실제 대용량 데이터에서 이전에 알려지지 않은, 잠재적으로 유용한 정보 추출	통계학 머신러닝
음성 인식	사람의 음성 언어를 컴퓨터가 해석하여 그 내용을 문자 데이터로 전환하는 처리	AI스피커
컴퓨터 비젼	컴퓨터를 이용하여 시각적 기능을 갖는 기계장치를 만들려는 분야	안면인식

Ⅳ. 수집 · 정제에 대한 AI 적용가능성 평가

온라인가격정보의 종류와 비정형성 비교

- **온라인가격정보의 종류**
  - 텍스트 : 카테고리, 제품 식별번호(ID), 상세정보(상품명 등)
  - 이미지 : 상품 이미지
- **비정형성 비교(텍스트)**
  - 상품명 등 상세정보는 물가지수 산출에 유용한 정보가 다수 포함됨에도 카테고리, 식별번호 등 다른 정보보다 비정형성이 큼

포함된 정보량이 많고, 비정형성 큰 상세정보 (상품명 등)에 대해 우선 검토

구분	내용	포함된 정보량	비정형성
텍스트	제품 카테고리	적음	다소 작음
	제품 식별번호(ID)	적음	작음
	상세정보(상품명 등)	**많음**	**큼**
이미지	상품 이미지	보통	보통

온라인가격정보 처리 작업별 내용(텍스트)

- **각 작업별로 세부작업 존재, 작업별로 처리방식도 상이**
  - 검출 : 묶음상품, 이상치, 동일상품 중복판매, 구신상품 매칭
  - 정보추출 : 단위 정보 추출
  - 품목분류 : 특정 상품이 속한 품목군 매칭

작업 내용 및 특성을 고려해 AI 선별 적용 검토

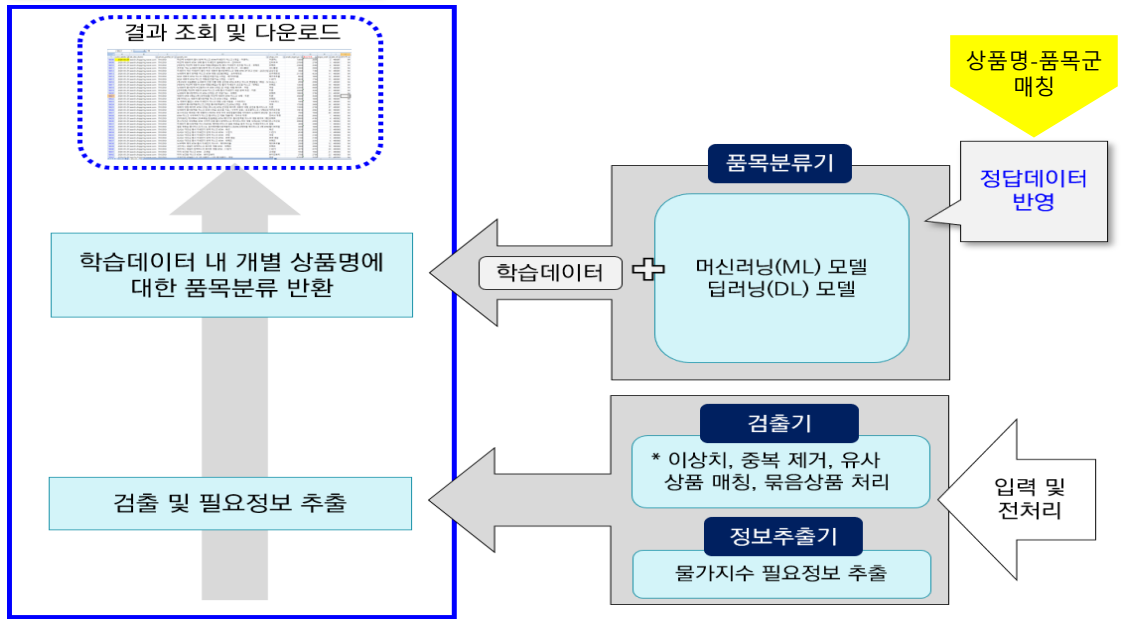
작업	세부작업	문제 정의	입력	출력	예시
검출	묶음상품	상품에 복수 개 상품 포함 여부	상품명 (텍스트)	{T, F} 중 1가지	마스크 : "KF94 마스크 1매 특가! 세정제 1개!" → T
	이상치	이상제품 여부	상품명 (텍스트)	{T, F} 중 1가지	전복 품목: "해신탕 해천탕 4-5인분 …" → T
	동일상품 중복판매	중복상품 여부	복수 개 상품명 (텍스트)	{T, F} 중 1가지	예시 없음
	구신상품 매칭	주어진 2개 제품 {T1, T2}이 같은 제품인데 새롭게 등장했는지 여부	복수 개 상품명 (텍스트)	{T, F} 중 1가지	예시 없음
정보 추출	단위 정보 추출	상품 개수에 대한 정보 추출	상품명 (텍스트)	숫자 형태	"바른고기집 당일 가공 냉장 생닭 10호 1kg + 1kg 총 두 마리" → {'kg': ['1', '1'], '마리': ['2']}
품목 분류	품목분류	품목 분류	상품명 (텍스트)	품목분류명	"갈치 2미 1.5kg …" → "갈치"

Ⅳ. 수집 · 정제에 대한 AI 적용가능성 평가 (계속)

처리 작업별 AI 적용가능성 평가

- **AI 적용방법 선택** : 자연어처리(NLP)를 통한 텍스트마이닝
  - 온라인가격정보의 다수를 차지하는 비정형 텍스트 처리에 유용
- **작업별 적용가능성 평가** : NLP 필요성 정도에 따라 평가
  - 동일상품 중복판매, 구신상품 매칭, 품목분류 처리에 적용 가능

예상 처리 흐름도



적용가능성 평가

작업	세부작업 (내용)	AI 적용가능성
검출	묶음상품 (복수 개 상품 포함 여부)	보통
	이상치 (이상제품 여부)	보통
	동일상품 중복판매 (동일상품 복수채널 판매 여부)	높음
	구신(舊新)상품 매칭 (특정상품의 지속 판매 여부)	높음
정보추출	물가지수 작성에 필요한 정보 추출 (가격, 용량, 무게, 개수 등)	낮음
품목분류	물가지수 대상품목 분류	매우 높음

V. AI 활용 예시

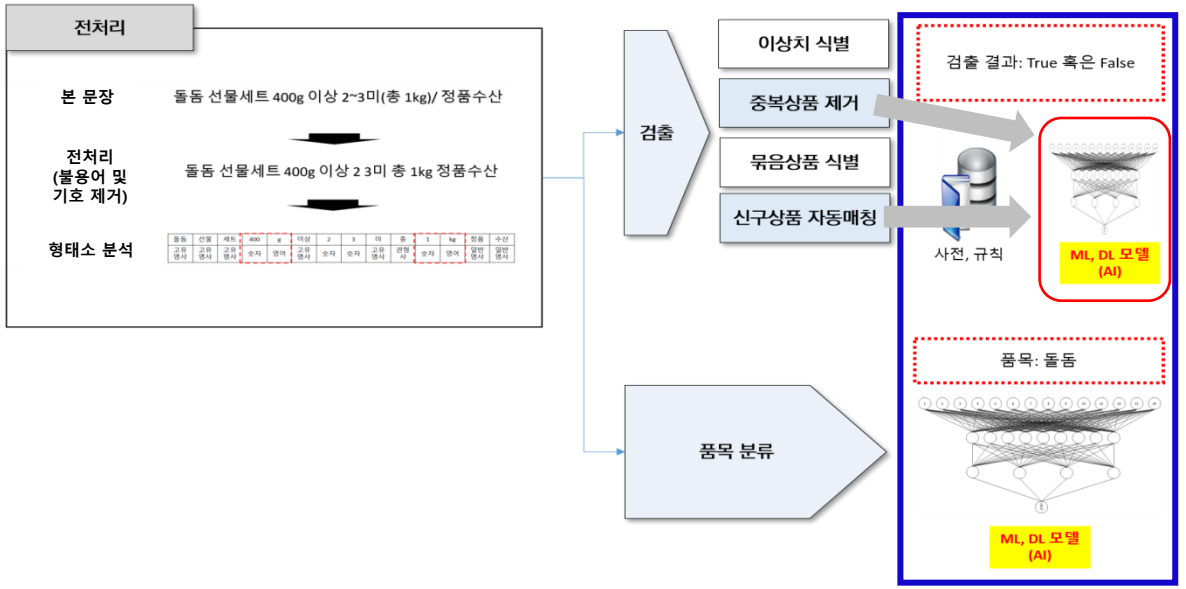
적용가능한 AI 기법 및 모델 (작업별)

- **검출** : 중복상품 제거, 구신상품 자동매칭 ➔ KNN(최근접이웃법) 등 단순 머신러닝 기법 적용
  - **입력 대상 데이터** : 상품명 등 각종 정보(텍스트)
- **품목분류** : 상품명-품목군이 매칭된 정답데이터(training data)를 활용한 딥러닝 기반 언어모델 적용
  - **입력 대상 데이터** : 상품명(텍스트)

		대상	형태소 분석	검출/추출 방법	AI 적용 시 방법	
					특성(feature) 추출	적용 알고리즘/ 모델 (계획)
검출	이상치 식별	가격 등	○	확률분포 이용 및 조건 설정	-	-
	중복상품 제거	상품명 등 각종 정보	○	-	○ (필요시)	최근접 이웃 분석 *동일날짜 비교
	묶음상품 식별	상품명	○	단위정보 이용 및 조건 설정	-	-
	신구상품 자동매칭	상품명 등 각종 정보	○	-	○ (필요시)	최근접 이웃 분석 *과거/현재 비교
정보추출	단위 정보 추출	상품명	○	규칙 및 조건 설정	-	-
	기타 정보 추출	상품명	○	규칙 및 조건 설정	-	-
품목 분류		상품명	○	-	○ (모델에 따라 상이)	언어모델(CNN, RNN, Attention, BERT 등)

AI 활용 정보 처리 과정 (작업별)

- **[STEP 1] 공통(데이터 전처리)**
  - 상품명 입력 ➔ 전처리(불용어 및 기호 제거) ➔ 형태소 분석
- **[STEP 2] 작업별 AI 알고리즘 및 모델 적용**
  - **중복상품, 구신상품 자동매칭** : 머신러닝(machine Learning), 딥러닝(deep learning) 모델
  - **품목분류** : 머신러닝(machine Learning), 딥러닝(deep learning) 모델

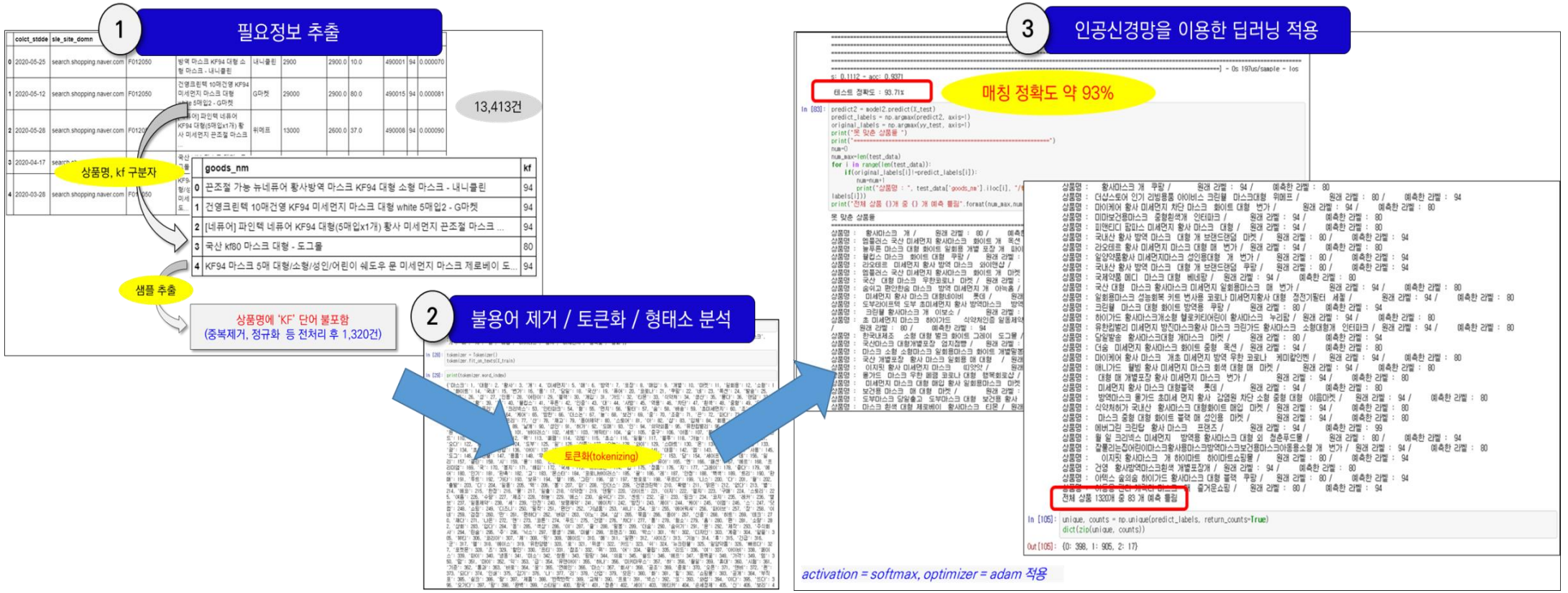


V. AI 활용 예시 (계속)

[사례] 실제 데이터(마스크) 이용 품목 분류

- **마스크 데이터 내 상품명(텍스트)을 입력으로 준 후에 80/94/99 품목 분류의 정확도 검증**
  - AI 트레이닝을 위한 마스크내 품목 80/94/99에 대한 정답데이터는 통계청 '물가동향과' 제공

- 데이터를 읽어 간단히 KF 80/94/99를 분류하는 딥러닝 모델 구축
  - KONLPY에서 제공하는 형태소 분석기를 이용
  - 불용어 사전은 일반적으로 사용하는 조사들을 임의대로 작성
  - 분류 테스트 : 상품명에 포함된 영어와 숫자를 지운 채 분류



연구 요약

- **비정형텍스트 형태의 상세정보(상품명 등)에 우선적 검토**
  - 상품명 등 상세정보는 포함정보가 많음에도 비정형 데이터 형태로 되어 있어 기존방식으로는 처리가 다소 어려운 상황
  - 비정형의 상세정보에 대해 AI 활용 자연어처리(NLP)가 가능함
- **세부작업별로 AI의 선택적 적용이 현실적 대안**
  - 작업별로 데이터 처리 내용 및 방식이 다르게 나타남
  - 이에 작업별 특성에 따라 AI를 선택적으로 적용하도록 함
- **세부작업별로 AI 적용가능성이 다르고, 비록 AI 적용이 가능하다고 해도 적용가능한 AI모델 및 알고리즘 상이**
  - **검출**: 중복상품 제거, 구신상품 자동매칭 ⇨ KNN(최근접이웃법) 등 단순 머신러닝 기법 적용
  - **품목분류**: 상품명-품목군 매칭된 정답데이터(training data)를 활용한 딥러닝 기반 언어모델 적용

연구의 한계점과 향후 과제

- **연구의 한계점**
  - **데이터를 이용한 다양한 분석 사례 제시 및 보완 필요**
    - 품목분류에 대한 AI 적용 사례 외 다양한 처리 작업(특히, 검출)에 대한 AI 적용 사례 제시 필요 (현재 연구 중)
    - 본 연구는 온라인가격정보 수집·정제에 대한 AI 적용의 기초 연구로서 향후 추가 연구 예정
  - **작업별 선택적 적용에 집중한 관계로 AI 적용에 대한 통합적 방안 및 아이디어 제시 필요**
    - 향후 추가연구를 통해 AI 적용을 위한 통합된 아이디어 마련 및 구체화
- **향후 과제**
  - **온라인가격정보 수집·정제 작업에 대한 AI 적용의 통일된 기준·합의 및 관련 논의 필요**
  - **모델 성능 평가·검증을 위한 정답 데이터 구축 방안 마련**
    - 지도학습 특성상 미지의 데이터 학습 전 AI 모델의 트레이닝 위한 정답 데이터 확보 필요