2020 빅데이터 경진대회 본선과제 #2

수비드 데이터 팀

대한민국 유통 활성화를 위한 적요 표준화

목차

1. 수비드 데이터 팀

“

수비드(Sous Vide)는 육류를 저온에서 장시간 조리해

육질을 연하게 만드는 요리 기법.

의미를 한눈에 파악하기 어려운 데이터를 가공 및 분석해

사용자가 쉽게 소화하도록 도와주는 서비스를 제공.

”

위희주, 김현경, 황지영

2. 과제 정의 및 목표

적요 Data Regularization

- 적요 데이터는 회사의 회계 기록을 수기로 작성한 데이터로, **물품 흐름을 파악**할 수 있는 자료
- 수기로 작성하는 적요 데이터의 특성상 작성자의 **주관적 판단**에 달려있어 표준화가 어려움

2. 과제 정의 및 목표

적요 Data Regularization

- 적요 데이터는 회사의 회계 기록을 수기로 작성한 데이터로, **물품 흐름을 파악**할 수 있는 자료
- 수기로 작성하는 적요 데이터의 특성상 작성자의 **주관적 판단**에 달려있어 표준화가 어려움

과제 목표

- **Word Vector, Google Search** 등의 방법을 통해 빅데이터를 이용하여

  적요 데이터으로부터 **주관적 판단을 최대한 배제**하며, **물품분류목록에 Mapping**하는 프로세스를 제안.
- **데이터 기반의 방법론**을 통해 대한민국의 유통 데이터 전산화 및 유통 활성화를 돕는

  효율적인 적요 표준화를 목표로 함.

2. 과제 정의 및 목표

품목분류코드.csv

category ↑		
가디건/조끼	니트/스웨터	목걸이/펜던트
가디건/조끼	니트/스웨터	목욕용품
귀걸이	드레스셔츠/캐쥬얼셔츠/남방	바디케어
기능성/체형보정/섹시속옷	런닝	바지&팬츠
기능화/실내화	런닝/팬티세트	반지
남성가방	레깅스	벨트
남성구두	마사지/팩/시트	분유/기저귀/물티슈
남성캐주얼화	모유/분유수유/이유용품	뷰티소품/소용량세트
내의/잠옷	모자/가발/귀마개	브라
네일케어	모자/모자세트	브라/팬티세트

- 주어진 품목분류코드는 243개 항목 중 90개 이상이 의류 패션 용품으로, 특정 품목이 **과도하게 자세히 분류**되어 있음.

- 반면 기업 간 물품 흐름이 세세하게 분류되어야 할 물류장비 및 공구는 분류가 충분히 되어있지 않음.

주어진 '품목분류코드.csv' 가
기업 적요 표준화 품목 분류 코드에
적합하지 않다고 판단

2. 과제 정의 및 목표

물품안내지도

물품분류

상품정보시스템 Product Information System

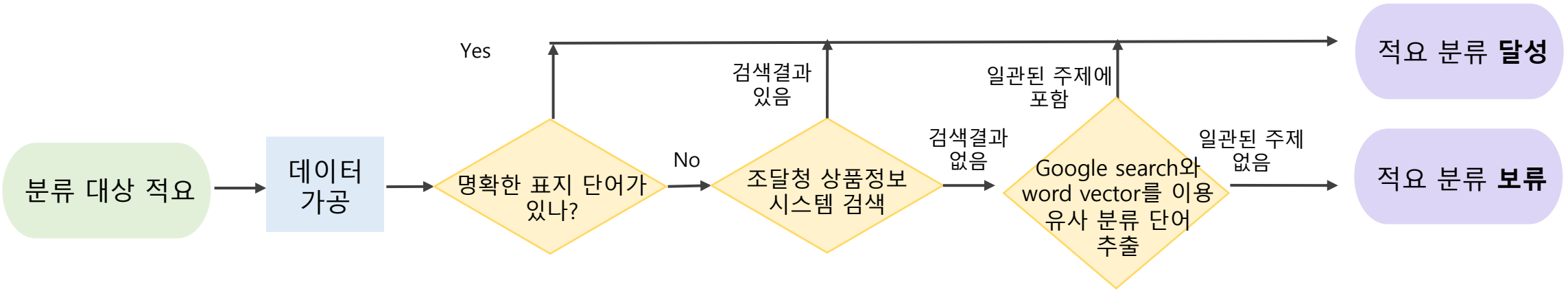
분류

[10]산동식물및동식물성생산품	설명보기	[11]광물,직물및비식용동식물자원	설명보기
[12]화학제품	설명보기	[13]수지,고무,탄성중합체	설명보기
[14]종이원료및종이제품	설명보기	[15]연료,연료첨가제,윤활유및방부식제	설명보기
[20]광산기계및액세서리	설명보기	[21]농,수,임,축산용기계	설명보기
[22]건축건설기계및보조용품	설명보기	[23]산업용제조가공기계및액세서리	설명보기
[24]물품취급,조정,저장기계,액세서리및소모품	설명보기	[25]상용,군용,개인용운송기구및액세서리와부품	설명보기

- 조달청에서 제공하는 물품분류는 국제 물품 분류 체계 UNSPSC를 기반으로 만든 한국 분류체계로, **우리나라 실정에 맞게 수정**되어 **한국 기업의 적요 표준화 분류에 적합**
- 주어진 적요 데이터가 100% 반영될 수 있게끔 **조달청의 상품정보시스템 + 몇 가지 항목을 더한 목록**을 적요 표준화에 이용함

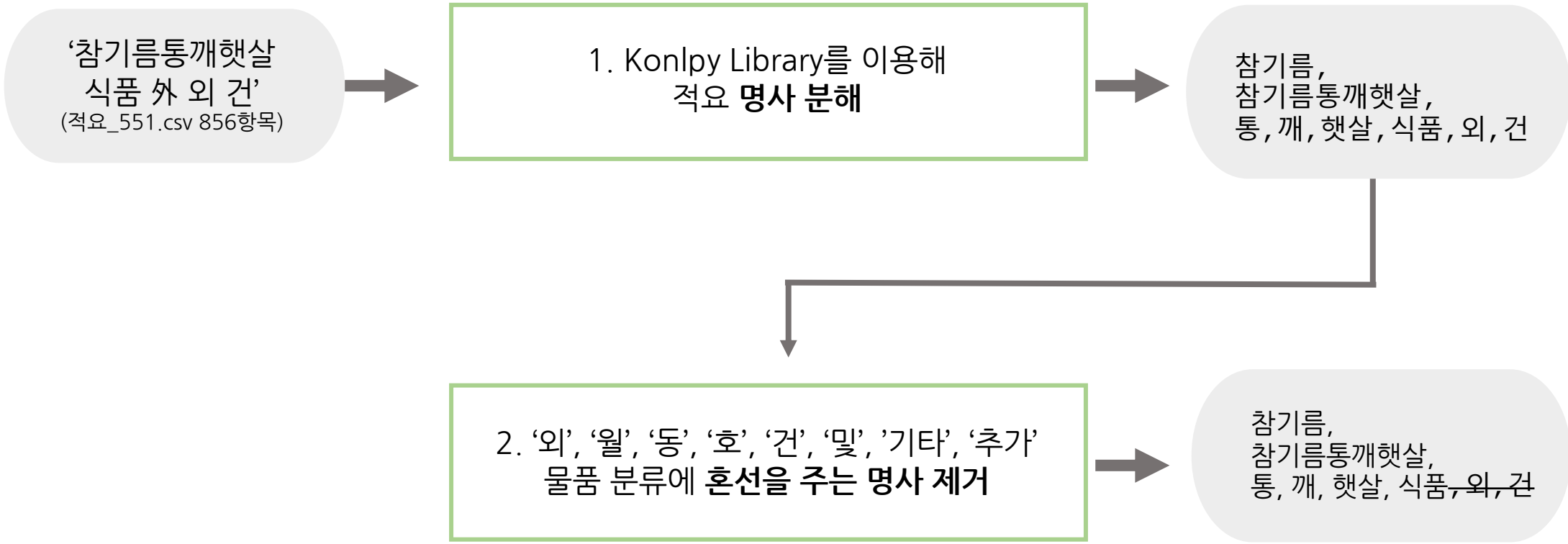
3. 적요 분류 Process 제안

0. 데이터 가공



3. 적요 분류 Process 제안

0. 데이터 가공



3. 적요 분류 Process 제안

1. 표지 단어 포함 적요 분류

Counter({'외': 8763, '건': 2001, '월': 1886, '공사': 1788, '수수료': 841, '제작': 610, '': 610, '설치': 589, '임대료': 497, '동': 480, '매출': 465, '교체': 441, '년': 415, '대금': 412, '카드': 383, '종': 378, '자재': 364, '제본': 364, '대': 362, '전기': 361, '주': 361, '보수': 343, '장비': 329, '회사': 328, '주식': 318, '요금': 311, '시': 305, '작업': 301, '박스': 292, '간': 291, '주식회사': 283, '영상': 275, '사용료': 268, '에어컨': 255, '분': 252, '인쇄': 252, '용역': 249, '유': 249, '차량': 247, '수리': 246, '케이스': 245, '점': 242, '기장': 236, '호': 234, '등': 233, '일반': 232, '관리비': 223, '자료': 221, '청소': 210, '차': 209, '통신': 206, '출력': 204, '전화': 203, '전자': 203, '상품': 203, '소방': 202, '판매': 201, '팀': 195, '포': 194, '사업': 187, '진행': 186, '스텐': 185, '마트': 182, '리': 181, '평': 178, '판': 177, '진행팀': 174, '기타': 172, 'AIR': 172, '부품': 171, '납품': 170, '소모품': 170, '시설': 169, '커피': 167, '핀': 167, '비용': 166, '서비스': 165, '설비': 165, '펌프': 164, '소': 162, '비': 162, '럭': 162, '가스': 161, '장': 160, '파이프': 159, '월분': 158, '배관': 157, '인쇄비': 156, '관': 156, '운송료': 155, '가공': 154, '세트': 153, '형': 152, '식대': 151, '고압': 150, '인증서': 150, '일': 150, '구입': 149, '티': 147, '지점': 147, '교육': 146, '추가': 145, '교환': 144, '구': 144, '外': 144, '럭시': 143, '시공': 139, '호스': 137, '개': 137, '용품': 135, '수리비': 135, '안전': 134, '설계': 134, '기계': 132, '대행': 132, '층': 130, '센터': 129, '보고서': 129, '중': 125, '컴퓨터': 124, '보조': 124, '산': 123, '평가': 123, '케이블': 121, '충전기': 121, '레': 120, '제안서': 120, '인터넷': 119, '관리': 119, '하수도': 119, '방': 118, '임대': 118, '개선': 118, '기': 118, '종합': 116, '전': 115, '미': 115, '카': 114, '광고': 114, '환경': 114, '일반전화': 113, '넬': 113, '씨': 113, 'CYLINDER': 112, '복사': 111, '다이어리': 110, '해소': 110, '기기': 109, '현금': 108, '한국': 108, '항공':

3. 적요 분류 Process 제안

1. 표지 단어 포함 적요 분류

Counter({'외': 8763, '건': 2001, '월': 1886, '공사': 178 수수료 341, '제작': 610, '': 610, '설치': 58 임대료 97, '동': 매출
출': 465, '교체': 441, '년': 415, '대금': 412, '카드': 3 , '자재': 364, '제본': 364, '대': 362 '주': 361
343, '장비': 329, '회사': 328, '주식': 318, '요금': 311, '시': 305, '작업': 301, '박스': 292, '간': 291, '주식회사': 283, '영상': 275, '사
용료': 268, '에어컨': 255, '분': 252, '인쇄': 252, '용역': 249, '유': 249, '차량': 247, '수리': 246, '케이스': 245, '점': 242, '기장': 236,
'호': 234, '등': 233, '일반': 232, '관리비': 223, '자료': 221, '청소': 210, '차': 209, '통신': 206, '출력': 204, '전화': 203, '전자': 203,
'상품': 203, '소방': 202, '판매': 201, '팀': 195, '포': 194, '사업': 187, '진행': 186, '스텐': 185, '마트': 182, '리': 181, '평': 178,
'판': 177, '진행팀': 174, '기타': 172, 'AIR': 172, '부품': 171, '납품': 170, '소모품': 170, '시설': 169, '커피': 167, '핀': 167, '비용': 16
6, '서비스': 165, '설비': 165, '펌프': 164, '소': 162, '비': 162, '럭': 162, '가스': 161, '장': 160, '파이프': 159, '월분': 158, '배관': 15
7, '인쇄비': 156, '관': 156, '운송료': 155, '가공': 154, '세트': 153, '형': 152, '식대': 151, '고압': 150, '인증서': 150, '일': 150, '구
입': 149, '티': 147, '지점': 147, '교육': 146, '추가': 145, '교환': 144, '구': 144, '外': 144, '럭시': 143, '시공': 139, '호스': 137, '개':
137, '용품': 135, '수리비': 135, '안전': 134, '설계': 134, '기계': 132, '대행': 132, '층': 130, '센터': 129, '보고서': 129, '중': 125, '컴
퓨터': 124, '보조': 124, '산': 123, '평가': 123, '케이블': 121, '충전기': 121, '레': 120, '제안서': 120, '인터넷': 119, '관리': 119, '하수
도': 119, '방': 118, '임대': 118, '개선': 118, '기': 118, '종합': 116, '전': 115, '미': 115, '카': 114, '광고': 114, '환경': 114, '일반전
화': 113, '넬': 113, '씨': 113, 'CYLINDER': 112, '복사': 111, '다이어리': 110, '해소': 110, '기기': 109, '현금': 108, '한국': 108, '항공':

상위 빈도 단어 중 포함 항목을 표지하는 **표지 단어** 존재

3. 적요 분류 Process 제안

1. 표지 단어 포함 적요 분류

월 배달의민족 카드매출분 (적요_551.csv 8644항목) → **표지단어** '매출' 포함 → [00]현금 분류

매출	[00]현금
전문점, 식당	[90]여행,음식,숙박및오락관련서비스
officeDEPOT, 오피스디포	[44]사무용기기액세사리및용품

표지 단어를 통한 우선 분류 -> 시스템 효율과 정확도를 향상

3. 적요 분류 Process 제안

2. 조달청 상품정보시스템 품목 검색

3. 해당하는 조달청 물품 분류로 표기

[30]건자재

건자재	건축, 토목구조물과 도로, 터널, 교량의 건설 및 시설물 유지,보수공사 등에 사용되는 각종 자재를 말하며, 철강재, 콘크리트, 목재, 조립식구조물, 내,외장마감재 등이 있음.

(조달청의 물품분류)

3. 적요 분류 Process 제안

3. Google search 와 Word Vector를 이용한 항목 분류 Overview

기존 적요의 문제점

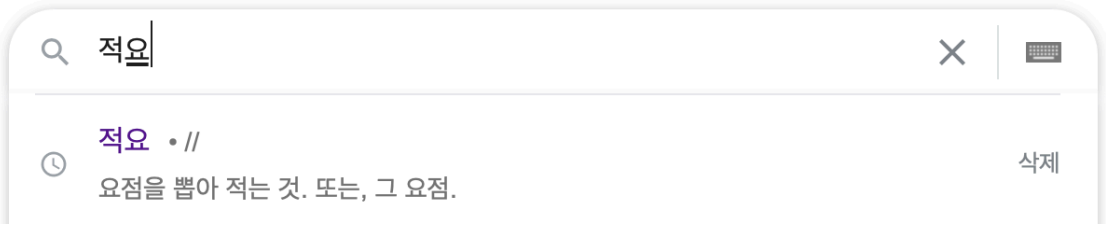
수기로 작성하는 적요 데이터의 특성상 작성자의 **주관적 판단**에 달려있어 표준화가 어려움

3. 적요 분류 Process 제안

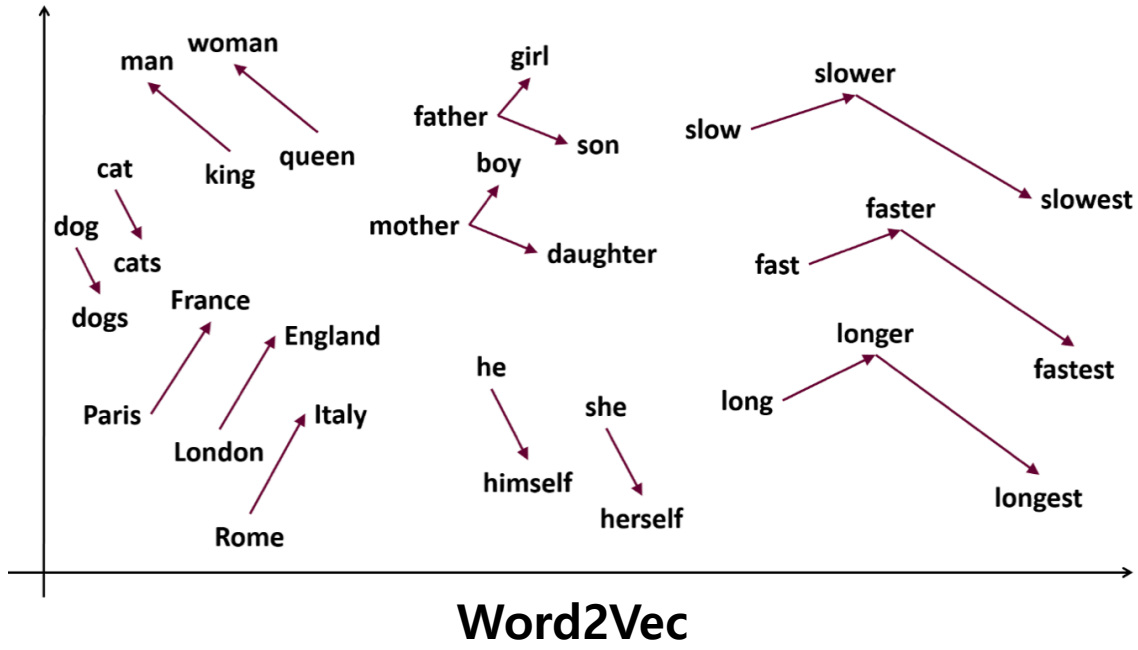
3. Google search 와 Word Vector를 이용한 항목 분류 Overview

기존 적요의 문제점

수기로 작성하는 적요 데이터의 특성상 작성자의 **주관적 판단**에 달려있어 표준화가 어려움



Google search를 통한 관련 빅데이터 수집



3. 적요 분류 Process 제안

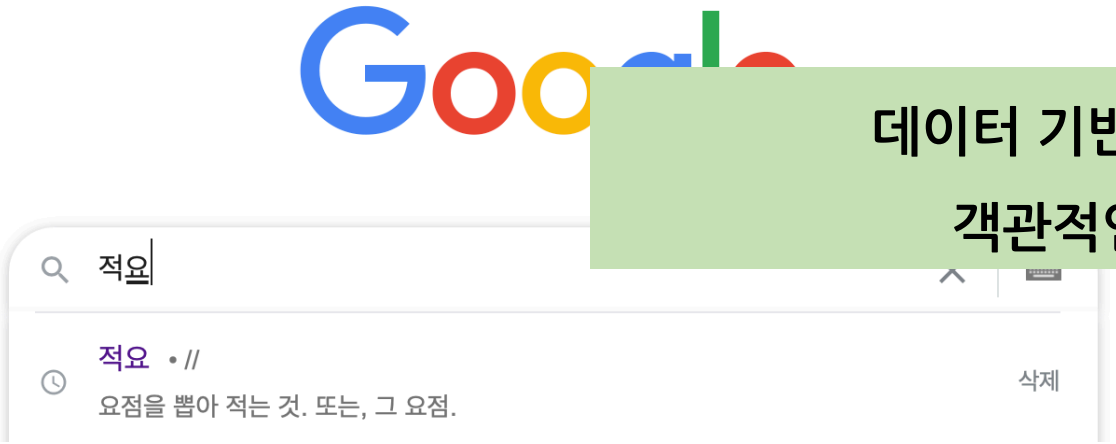
3. Google search 와 Word Vector를 이용한 항목 분류 Overview

기존 적요의 문제점

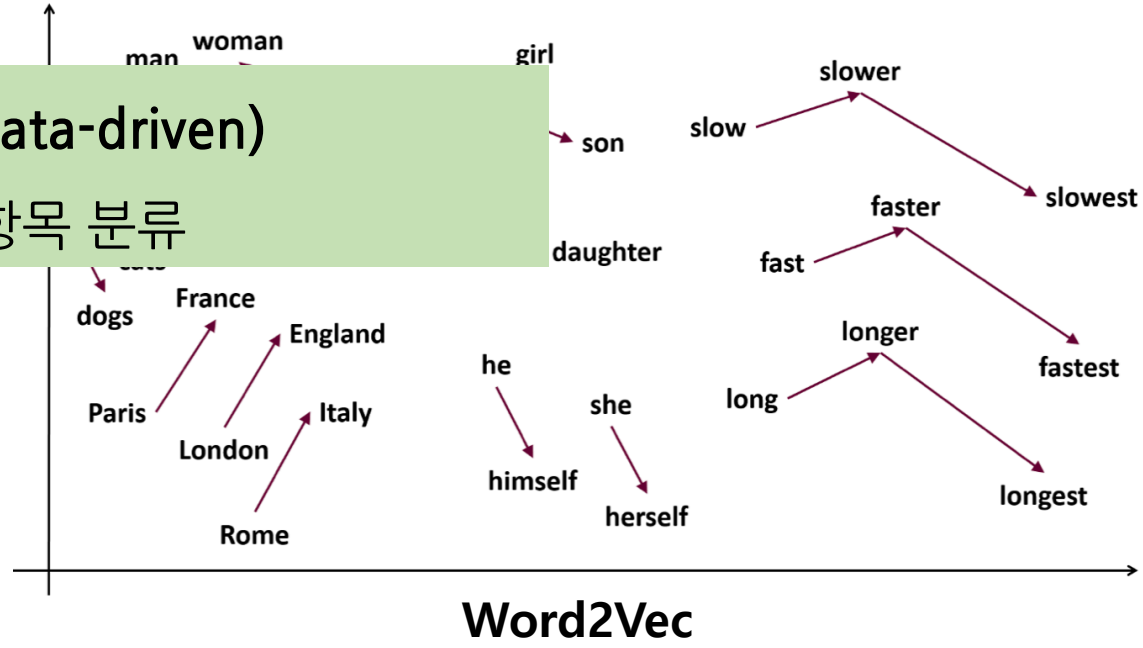
수기로 작성하는 적요 데이터의 특성상 작성자의 **주관적 판단**에 달려있어 표준화가 어려움

데이터 기반(Data-driven)

객관적인 항목 분류



Google search를 통한 관련 빅데이터 수집



3. 적요 분류 Process 제안

3. Google search 와 Word Vector를 이용한 항목 분류 Overview

Google Search를 통해
적요 관련
빅데이터 크롤링

↓

관련도가 높은 키워드 추출

↘ word2vec

키워드 벡터

조달청 물품 분류

- [71]광업,석유및가스서비스
- [73]공산품제조서비스
- [77]환경관련서비스
- [80]경영관련서비스
- [82]편집디자인그래픽및예술관련서비스

↙ word2vec

물품 분류명 벡터

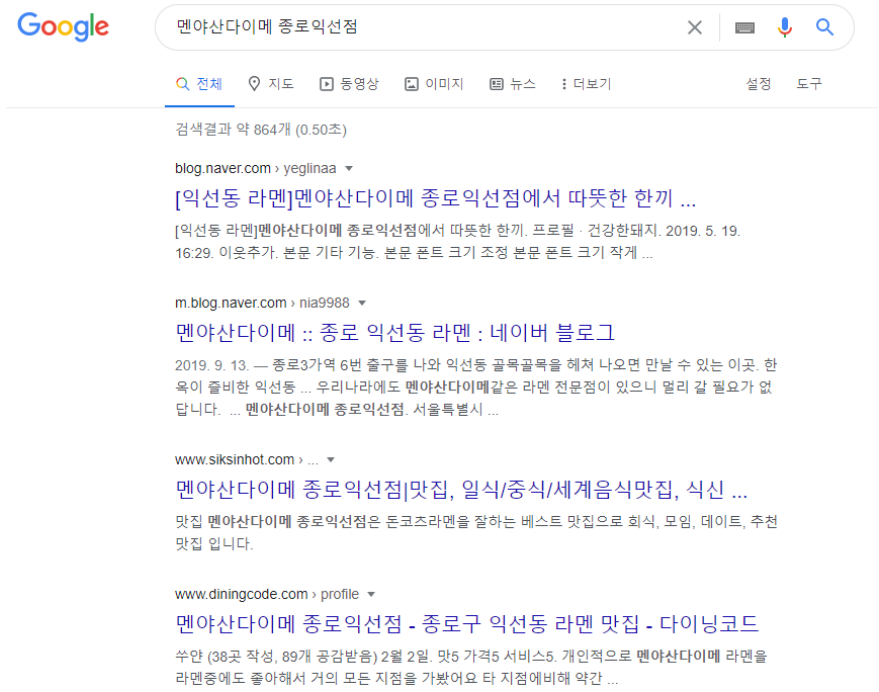
두 벡터 간의 유사도(cosine similarity) 계산

Max[{중요 키워드 Word Vector}⊗{항목 단어 Word Vector}]가
최대가 되게 하는 항목으로 분류

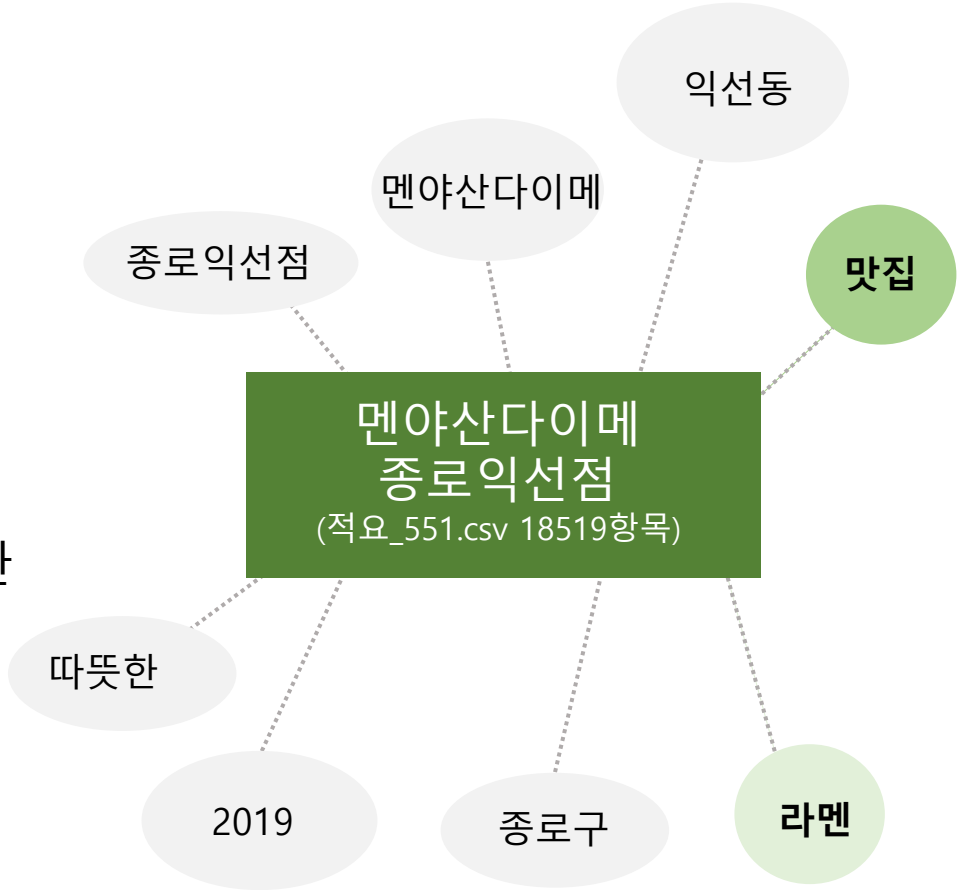
3. 적요 분류 Process 제안

3.1 Google Search 통한 적요 중요 키워드 선정

1. Google Search를 통해 관련 빅데이터 크롤링
(Python library BeautifulSoup)

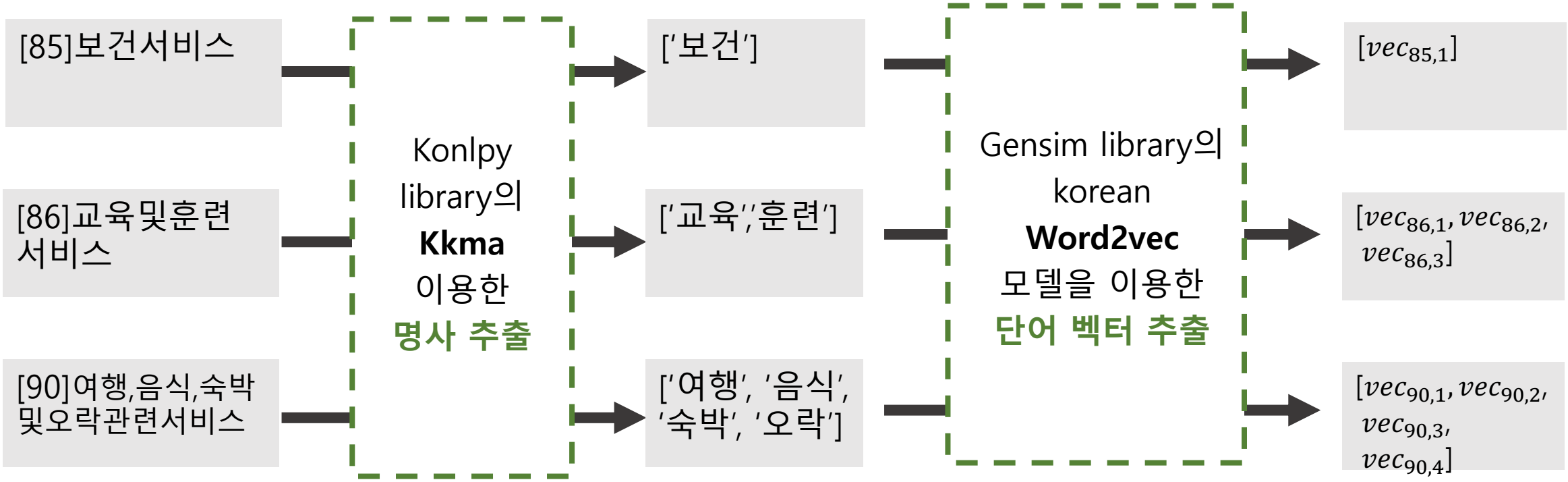


2.
날짜/url/지명/조사 제외한
최다빈도 단어 두 개를
중요 키워드 로 설정



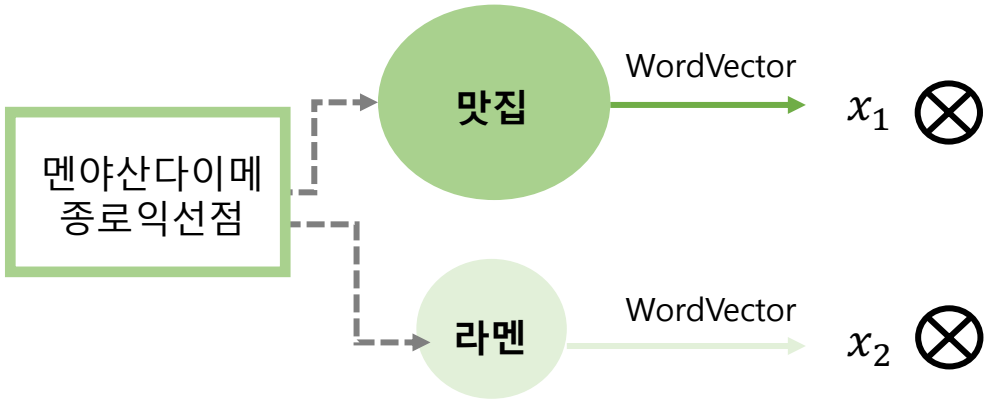
3. 적요 분류 Process 제안

3.2 분류 항목의 벡터화



3. 적요 분류 Process 제안

3.3 Word Vector 코사인 유사도를 이용한 항목 분류



x_1 ⊗

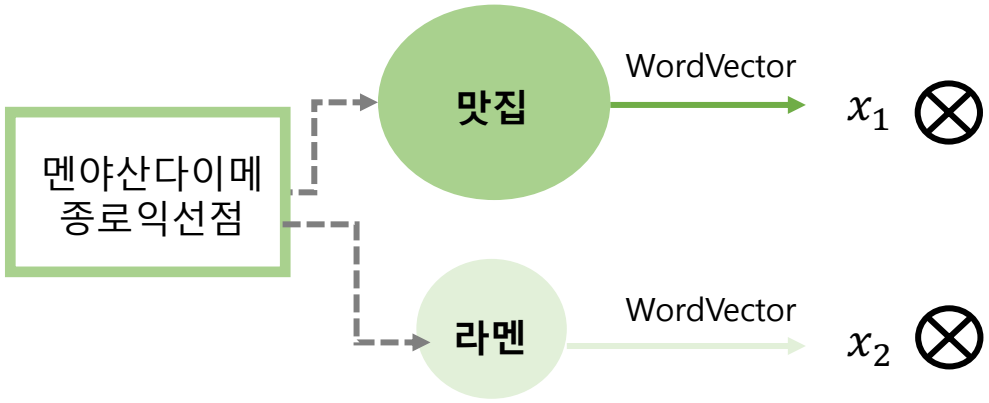
x_2 ⊗

	[85]보건 서비스		[84]교육및훈련 서비스		[90]여행,음식,숙박및 오락관련서비스		…
$vec_{85,1}$	-0.119	$vec_{84,1}$	-0.035	$vec_{90,1}$	0.091	
				$vec_{90,2}$	**0.586**	
		$vec_{84,2}$	**-0.012**	$vec_{90,3}$	0.037	
				$vec_{90,4}$	0.093	
$vec_{85,1}$	**-0.052**	$vec_{84,1}$	-0.026	$vec_{90,1}$	0.060	
				$vec_{90,2}$	0.313	
		$vec_{84,2}$	-0.083	$vec_{90,3}$	0.122	
				$vec_{90,4}$	0.120	
Max	-0.052		-0.012		**0.586**	

Thanks to @Kyubyong Park, Github

3. 적요 분류 Process 제안

3.3 Word Vector 코사인 유사도를 이용한 항목 분류



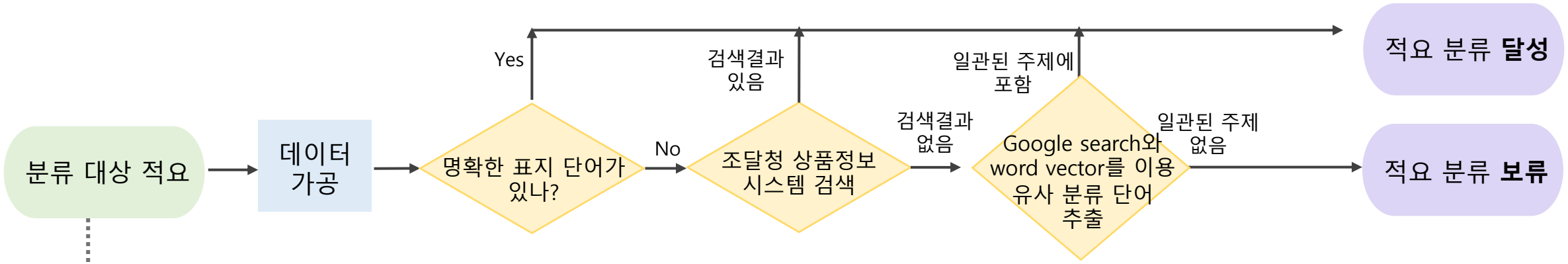
	[85]보건 서비스		[84]교육및훈련 서비스		[90]여행,음식,숙박및 오락관련서비스		...
x_1	$vec_{85,1}$	-0.119	$vec_{84,1}$	-0.035	$vec_{90,1}$	0.091	
					$vec_{90,2}$	**0.586**	
			$vec_{84,2}$	**-0.012**	$vec_{90,3}$	0.037	
					$vec_{90,4}$	0.093	
x_2	$vec_{85,1}$	**-0.052**	$vec_{84,1}$	-0.026	$vec_{90,1}$	0.060	
					$vec_{90,2}$	0.313	
			$vec_{84,2}$	-0.083	$vec_{90,3}$	0.122	
Max	-0.052		-0.012		**0.586**		

[90]으로 분류

Thanks to @Kyubyong Park, Github

4. 정리

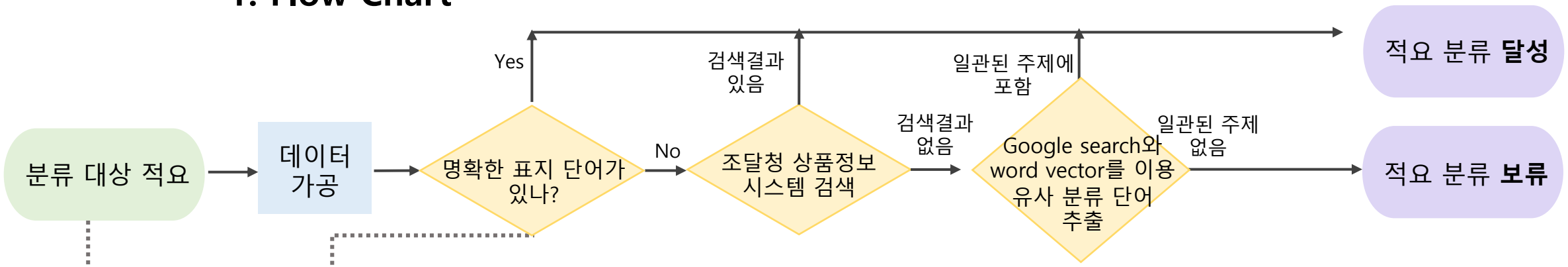
1. Flow Chart



ex)월 신용카드 매출,
해표식용유 L ,
행복란HappyEgg 왕란판란,
라즈G 외 건

4. 정리

1. Flow Chart



월 신용카드 매출 -> 월/신용카드/매출
→ [00]현금 분류

ex)월 신용카드 매출,
해표식용유 L ,
행복란HappyEgg 왕란판란,
라즈G 외 건

4. 정리

1. Flow Chart

분류 대상 적요 → 데이터 가공 → 명확한 표지 단어가 있나?

Yes → 적요 분류 **달성**

No → 조달청 상품정보 시스템 검색

검색결과 있음 → 적요 분류 **달성**

검색결과 없음 → Google search와 word vector를 이용 유사 분류 단어 추출

일관된 주제에 포함 → 적요 분류 **달성**

일관된 주제 없음 → 적요 분류 **보류**

월 신용카드 매출 -> 월/신용카드/매출
→ [00]현금 분류

해표식용유 L-> 해표/식용유/(L)
→ [50] 식음료및담배제품

ex)월 신용카드 매출,
해표식용유 L ,
행복란HappyEgg 왕란판란,
라즈G 외 건

4. 정리

1. Flow Chart

분류 대상 적요 → 데이터 가공 → 명확한 표지 단어가 있나?

- Yes → 적요 분류 **달성**
- No → 조달청 상품정보 시스템 검색
  - 검색결과 있음 → 적요 분류 **달성**
  - 검색결과 없음 → Google search와 word vector를 이용 유사 분류 단어 추출
    - 일관된 주제에 포함 → 적요 분류 **달성**
    - 일관된 주제 없음 → 적요 분류 **보류**

월 신용카드 매출 -> 월/신용카드/매출
→ [00]현금 분류

해표식용유 L-> 해표/식용유/(L)
→ [50] 식음료및담배제품

ex)월 신용카드 매출,
해표식용유 L ,
행복란HappyEgg 왕란판란,
라즈G 외 건

4. 정리

1. Flow Chart

분류 대상 적요 → 데이터 가공 → 명확한 표지 단어가 있나?

Yes → 적요 분류 **달성**

No → 조달청 상품정보 시스템 검색

검색결과 있음 → 적요 분류 **달성**

검색결과 없음 → Google search와 word vector를 이용 유사 분류 단어 추출

일관된 주제에 포함 → 적요 분류 **달성**

일관된 주제 없음 → 적요 분류 **보류**

월 신용카드 매출 -> 월/신용카드/매출
→ [00]현금 분류

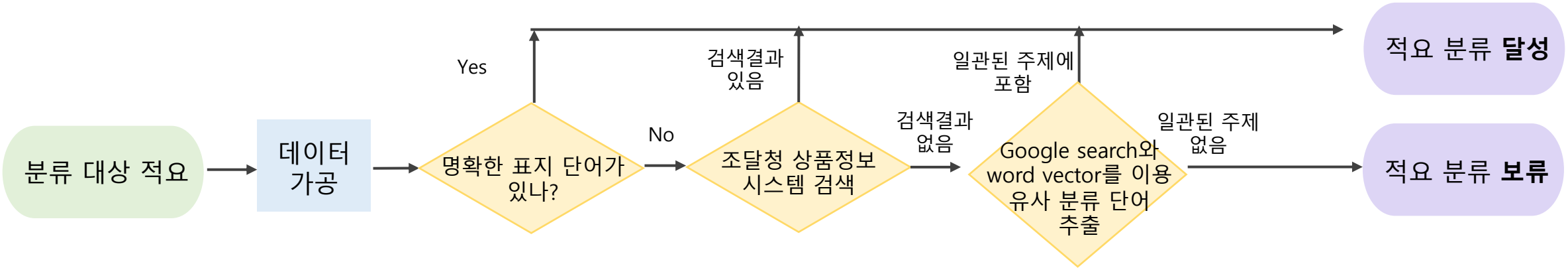
해표식용유 L-> 해표/식용유/(L)
→ [50] 식음료및담배제품

라즈G 외 건-> 라즈G/(외)/건
→ 시스템 안정성을 위한 분류 보류

ex)월 신용카드 매출,
해표식용유 L ,
행복란HappyEgg 왕란판란,
라즈G 외 건

4. 정리

2. Process 의의



- 주관적인 판단을 배제하고, google search를 이용한 **빅데이터 기반으로 객관성을 지닌** 적요 표준화를 시도.
- **조달청의 물품분류체계**를 이용해 추후 **대한민국 유통 데이터** 전산화 및 활용 활성화에 적합
- 더존비즈온이 보유하고있는 방대한 적요 데이터를 기반으로 한국어 모델 발전 가능

  -> 추후 **차별화 되는 적요 표준화 시스템 개발**할 수 있을 것으로 전망.

QnA

4. 정리

Reference

- 한국조달연구원, 사업안내
  (http://www.kip.re.kr/mall/business/listSphere.asp)
- KyuByong Park, GitHub, Pre-trained word vectors of 30+ languages
  (https://github.com/Kyubyong/wordvectors)

Thank you