기업 협업 프로젝트 WISEAPP × CODE STATES 수행 기간 : `21.8.19 ~ 9.8

소비자 결제 내역 데이터 기반 유사 단어별 grouping 및 금액별 분류 알고리즘 제안

CODESTATES
AI 2기 김병섭

Contents

- 1. 프로젝트 소개
  - *기업 및 서비스*
  - *요청 사항*
- 2. 프로젝트 진행 사항
  - *문제 사항*
  - *검토 방안*
- 3. 진행 결과
  - *진행 과정 설명*
  - *결과 확인*
- 4. 마무리

1. 프로젝트 소개

업체 및 서비스

- 기업명 : 아이디어웨어
- 서비스 : 결제 내역 데이터 분석 (분류/통계)
- 분석 분야 : 앱 시장, 소매 시장, 온라인 상품

※ 빅데이터를 통한 비즈니스 인사이트 도출

※ 프로젝트 데이터 : DB 내 21-07-01일자 전체

< 아이디어웨어(와이즈앱) 제공 서비스 예시 >

1. 프로젝트 소개

요청 사항

- 현 분류 rule 엔진**(수동 분류)**의 효율성 증가

※ rule 엔진 = 등록된 키워드 매칭 분류 (그 외 = 미분류)

→ 문제 = 높은 정확도, 긴 소요 시간 (자동화?)

- **미분류** 중 유의미한**(금액)** 브랜드별 클러스터링*

※ 미분류 중 건수/금액이 높은 경우, 사람이 확인하여 분류 키워드 추가

→ 클러스터링하여 높은 금액을 쉽게 확인하는 방법?

* 미분류 데이터이므로 클러스터링 결과 내 오차 허용

< 아이디어웨어(와이즈앱) 분류 엔진 예시 >

2. 프로젝트 진행 사항

문제 사항

- 데이터 (33개 컬럼) 중 **모델 학습 데이터 부족**

  ※ 카드 메시지별 벡터화 후 학습 검토

  → 카드 메시지가 짧아 NLP 적용 어려움 (말뭉치)

  = Word2Vec, GloVe, FastText 등 적용 시 학습 불가 or 기존 대비 성능 유사/저하 (클러스터링: DBSCAN)

- 불완전 단어, 비브랜드, 단순 입/출금, 비용 납부 등과 브랜드별 상이한 표기법, 브랜드명+지점 등 **데이터 전처리 방안 부재**

	결제금액	카드메시지2	브랜드명
0	15000	이미숙6월캡	미분류
1	3000	GS25중곡덕산	GS25
2	200000	경주페이	미분류
3	32880	하이웨이마트송탄지*	하이웨이마트
4	12000	월드약국	약국(종합)
...	...	...	...
2421099	3500	상주농협하나로	농협하나로마트
2421100	54000	이니시스-정	미분류
2421101	40000	잉크벨양구점	미분류
2421102	1000000	MAENWILAI	미분류
2421103	104930	(주)조이젠	미분류

2421104 rows × 3 columns

	결제금액	카드메시지2	브랜드명
0	15000	이미숙6월캡	미분류
2	200000	경주페이	미분류
7	271750	한전(정은영)	미분류
8	105000	대포항회수산	미분류
10	22100	신한카드	미분류
...	...	...	...
2421096	173215	하나카드금융	미분류
2421100	54000	이니시스-정	미분류
2421101	40000	잉크벨양구점	미분류
2421102	1000000	MAENWILAI	미분류
2421103	104930	(주)조이젠	미분류

1252380 rows × 3 columns

< (좌) 초기 데이터프레임, (우) 미분류 필터링 결과 >

2. 프로젝트 진행 사항

검토 방안

- 현업 방식과 유사한 알고리즘 구현으로 접근

※ 데이터 상태, 규모 / 문제 해결에 적합한 것으로 판단

→ 데이터 전처리 사항 (카드 메시지 표시 방법) 수립

- 전처리한 단어에 **편집거리 알고리즘*** 사용, grouping

※ 예상 결과 : 그룹 수, 기준 단어, 유사 단어 목록, 그룹별 총 금액

→ 금액별 정렬, 업체에서는 높은 금액 그룹 확인 가능

* 서로 다른 단어를 똑같이 만들기 위해 필요한 횟수 (수정, 삭제, 추가 등)
* 유사도 = 1 - 편집거리 결과 / (첫번째 단어 길이 + 두번째 단어 길이)

```
# 편집거리 결과 글자별 예시

# 글자 수 동일한 경우 (같은 종류 : 뒤가 같은 경우)
s1 = '재밌는빵집'
s2 = '신선한빵집'
print(levenshtein(s1, s2)) #.7

# 한쪽이 더 긴 경우 (s1,s2 순서 바꾸어도 결과 동일)
s1 = '좋은빵집'
s2 = '맛있고싼빵집'
print(levenshtein(s1, s2)) #.6

# 한쪽이 더 긴 경우 (앞이 같은 경우)
s1 = '현대카드'
s2 = '현대자동차'
print(levenshtein(s1, s2)) #.667

# 한쪽이 더 긴 경우 (중간이 같은 경우)
s1 = '나는현대철물'
s2 = '우리현대식당'
print(levenshtein(s1, s2)) #.667

# 글자 수 동일한 경우 (다른 종류)
s1 = '현대카드'
s2 = '착한빵집'
print(levenshtein(s1, s2)) #.5

s1 = '재밌는부동산'
s2 = '친절한부동산'
print(levenshtein(s1, s2)) #.75

# 한쪽이 더 긴 경우 (s1,s2 순서 바꾸어도 결과 동일)
s1 = '좋은부동산'
s2 = '매우착한부동산'
print(levenshtein(s1, s2)) #.667

# 한쪽이 더 긴 경우 (앞이 같은 경우)
s1 = '부동산좋아'
s2 = '부동산아저씨다'
print(levenshtein(s1, s2)) #.667

# 한쪽이 더 긴 경우 (중간이 같은 경우)
s1 = '진짜부동산가요'
s2 = '우리부동산갑니다'
print(levenshtein(s1, s2)) #.667
```

```
0.7
0.6
0.667
0.667
0.5
0.41700000000000004
0.75
0.667
0.667
0.667
```

< 편집거리 알고리즘 적용 예시 >

3. 진행 결과

진행 과정 설명

1) 편집거리
(하나 이하 유사할 시
0으로 일괄 적용)

2) 유사도별 분류
(0.667 이상)
→ 글자 2개 이상 동일

* 전 과정 약 25분 소요

1 데이터 불러오기 → 2 미분류 데이터 확인 → 3 금액 기준 정렬 → 4 1차 전처리 → 5 2차 전처리 → 6 3차 전처리 → 7 유사도 비교, 그룹화 → 8 유사 그룹별 확인 → 최종 결과

- 1) 전체 데이터 중 미분류 부분 확인
  2) 결제 금액 내림차순 정렬
- 1) 데이터 분포 확인 (hist, kde)
  2) 천 원 이상 결제 비율 확인
- 1) 주식/유한/재단
  2) 영어
  3) 특수 문자, 숫자
  4) 띄어쓰기 등
- 1) '내용없음'
  2) 빈 칸
  3) 비브랜드
  4) 메시지로 묶기
  5) 결제 횟수
  ※ 평균, 3분위 중 큰 값
- 1) 사람 이름
  2) 입/출금, 송금
  3) 일반 명사, 각종 비용/세금
  4) 지점 이름
  5) 결제 횟수
- 1) 그룹 수
  2) 유사 단어 수
  3) 금액 총합

3. 진행 결과

결과 확인

1 데이터 불러오기 → 2 미분류 데이터 확인 → 3 금액 기준 정렬

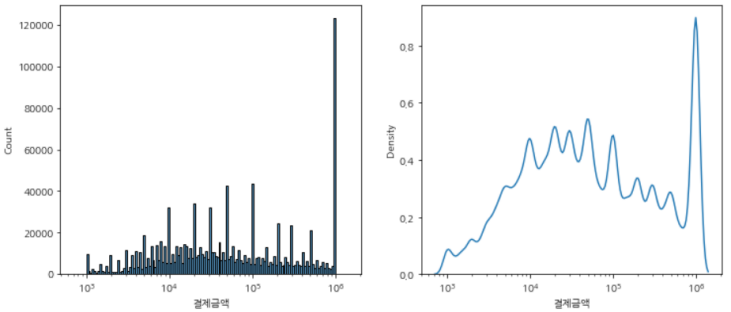
체크일자	데이터타입		업체메시지NID	결제방법	사용일시	카드승인타입	카드사명	결제금액	결제금액	카드메시지	카드메시지	사용가능여부	브랜드NID	브랜드명
2021-07-01	1	42b89b	a	₩N	5	2021-07-01 16:10:11	1	NH농협	15000	15000	이미숙6월캡	이미숙6월캡	₩N	-1	미분류
2021-07-01	1	8dc66	b	₩N	1	2021-07-01 22:15:03	1	신한	3000	3000	GS25중곡덕산	GS25중곡덕산	₩N	10354	GS25
2021-07-01	1	2ac7c	7	₩N	5	2021-07-01 15:17:43	1	NH농협	200000	200000	경주페이	경주페이	₩N	-1	미분류
2021-07-01	1	d3d777	e	₩N	1	2021-07-01 13:46:53	1	우리	32880	32880	하이웨이마트송탄지*	하이웨이마트송탄지*	₩N	18710	하이웨이마트
2021-07-01	1	e943a6	7	₩N	1	2021-07-01 18:03:59	1	우리	12000	12000	월드약국	월드약국	₩N	18509	약국(종합)
2021-07-01	1	0b13c5	0	₩N	1	2021-07-01	1	삼성	3500	3500	상주농협하나로	상주농협하나로	₩N	16396	농협하나로마트

< 제공받은 데이터 >

	결제금액	카드메시지2	브랜드명
0	8870	(주)이마트천안서	이마트
1	34180	리더스마트	미분류
2	6300	투썸플레이스	투썸플레이스
3	226000	우다움용산점	미분류
4	3700	(주)티머니 법인택시	티머니
...	...	...	...
1423730	36900	(주)애플마트	미분류
1423731	9500	Netflix_INIC	NETFLIX
1423732	36540	노브랜드 춘천점	노브랜드
1423733	10000	(주)오지소프트	미분류
1423734	14680	주식회사부방유	이마트

< 데이터 컬럼 준비 >

* 천 원 이상 사용한 데이터



< 데이터 분포 확인(histo, kde) >

3. 진행 결과

4 1차 전처리 → 5 2차 전처리 → 6 3차 전처리

결과 확인

* 해당 키워드 / 전처리 후 전체

	결제금액	카드메시지2	카드메시지_수정
0	1000500	현대리싸이클	현대리싸이클
1	1000000	대광통신 정	대광통신정
2	1000000	김현희	김현희
3	1000000	K5(4353)매입	케이매입
4	1000000	김순길	김순길
...	...	...	...
1181536	1000	3팝피씨	팝피씨
1181537	1000	김윤서	김윤서
1181538	1000	여수광양항만공	여수광양항만공
1181539	1000	SMS 06월요금	에스엠에스월요금
1181540	1000	SMS 06월요금	에스엠에스월요금

1181540 rows × 3 columns

< 주식/유한회사, 영어, 문자, 숫자 등 >

	카드메시지_수정	횟수
10	현대캐피탈	2965
11	교통버스건	2942
12	임대료	2597
15	우리카드결제	2340
16	회	2265
...	...	...
68326	은평구청구내식	3
68327	중문수두리보	3
68328	교동면옥구미봉	3
68329	정부세종청사총	3
68330	홍리마라탕	3

68133 rows × 2 columns

< 메시지별 묶기, 빈칸, 내용없음 등 >

	카드메시지_수정	횟수	비율(%)
0	현대캐피탈	2965	0.833
1	교통버스건	2942	0.826
2	경기지역화폐	2199	0.618
3	교통지하철건	1978	0.556
4	착한페이	1778	0.499
...	...	...	...
8892	송현충전소	8	0.002
8893	대성고기마트	8	0.002
8894	플로리안	8	0.002
8895	우정하이퍼	8	0.002
8896	재아름다운	8	0.002

8897 rows × 3 columns

< 사람 이름, 일반 명사, 입/출금 등 >

3. 진행 결과

7 유사도 비교, 그룹화 → 8 유사 그룹별 확인 → 최종 결과

결과 확인

1468 번째 cal_similarity, cal_similarity2 실행
비교 기준 단어(s1) : 사단법인두란노서원
-------남은 데이터: 1001 -------

1469 번째 cal_similarity, cal_similarity2 실행
비교 기준 단어(s1) : 메르시엠이알씨아이
1470 번째 cal_similarity, cal_similarity2 실행
비교 기준 단어(s1) : 크라상점

-------그룹 결과-------
그룹 개수 : 909
남은 데이터

	카드메시지_수정	횟수	비율(%)	s1과_유사_비율
0	델문도	11	0.003	0.2
1	모빙유씨피	11	0.003	0.2
2	하은정	11	0.003	0.2
3	자연그리고사람	11	0.003	0.2
4	경기두레소비자	11	0.003	0.2

< 유사도 기준 grouping 진행 결과 >

그룹 번호

	s1_이름	s1_유사_단어	유사_단어_개수	총금액(원)	총금액(원)/전체_금액(%)
0	현대캐피탈	[현대캐피탈주, 신한현대캐피탈, 한국캐피탈, 산은캐피탈, 케이비캐피탈, 애큐온캐피탈, 디지비캐피탈, 한국캐피탈주, 오케이캐피탈, 현대커머...	29	2,990,775,381	1.245
1	경기지역화폐	[천안지역화폐, 밀양지역화폐, 경남지역상품, 오픈지역사랑, 경기관광개발, 경기두레소비, 경기고속도로, 경기마트개봉, 경기할인마트, 대구지...	10	857,873,009	0.357
2	착한페이	[청주페이, 경주페이, 천안페이, 강릉페이, 착한탕국, 삼성페이, 착한밥상, 착한식판, 착한식당, 착한밥집, 착한가게, 제로페이, 착한돼...	35	584,248,129	0.243
3	케이비	[케이비충, 케이알, 케이엘, 케이에, 아이비, 케이카, 케이원, 제이비, 케이스, 케이지, 케이헴, 케이, 케이비제이, 케이비오픈, 케...	41	464,044,824	0.193
4	대구행복충전	[대경교통충전, 대영가스충전소, 대명가스충전소, 대양가스충전소, 대전가스충전소, 포항사랑충전, 대공원충전소, 대구시설공단, 대구축산농협,...	18	439,760,578	0.183
5	우리마트	[우리마트, 우리마트, 우리마트, 우리마트, 우리마트, 우리마트총, 우리들마트, 우리마트로, 우리마트감, 우리마트만, 우리마트진, 우리마...	755	420,756,780	0.175
6	엔에이치	[엔에이치엔, 비에이치, 엠에이치, 디에이치, 에이치, 엔에이, 엔에이치농협, 제이에이치, 이앤에이치, 대주에이치, 에이치앤, 에이치에,...	44	384,178,242	0.160
7	여민전	[여민락]	1	305,921,000	0.127
8	에치와이	[에치와이충, 에이치와이, 에스와이, 에치와, 에이치케이, 에스제이, 에스에이, 에스케이, 에스지이, 에이치와이제이, 에이치앤디이, 에스...	27	304,652,014	0.127
9	에스엠에스알림	[에스엠아이스크림, 에스에스마트, 에스에스유통, 에스에스마트, 에스에스충, 에스케이에스케이, 씨에스씨에스마트, 에스엠서비스토탈, 에스에스...	29	289,743,029	0.121
10	에이티엠	[에이엠피엠, 에스티엠넷, 에이피알, 에이플러, 제이티넷, 에스지엠, 에이스식, 에이마트, 에이디티, 에스티유, 에이블루, 에스엠, 제이...	31	286,775,798	0.119
11	오픈지역사랑상품	[오픈지역사랑상, 충주사랑상품, 춘천사랑상품]	3	281,238,879	0.117
12	엔에이치카드인터넷	[엔에이치비씨인터넷, 선불카드인터넷, 엔에이치카드에이알에스, 엔에이치비씨카드전화, 엘에이치할인마트, 디에이치디자인, 더티에이치이드림마트,...	10	272,745,125	0.114

< 편집거리 적용, 최종 grouping 결과 데이터프레임 >

* 기준 금액 또는 비율로 필터링 가능 (현재는 0.1% 이상 필터링)

4. 마무리

Good things

- 실제 기업 데이터 사용
- 현업 담당자와의 미팅
- 다양한 데이터 전처리
- 팀원 간 피드백

To be desired

- NLP용 데이터 형태
- 학습 모델 미사용 (알고리즘)
- 수작업 한계성 (처리 조건)

Any feedback or questions?

You can find me at @Sammy308 (github)