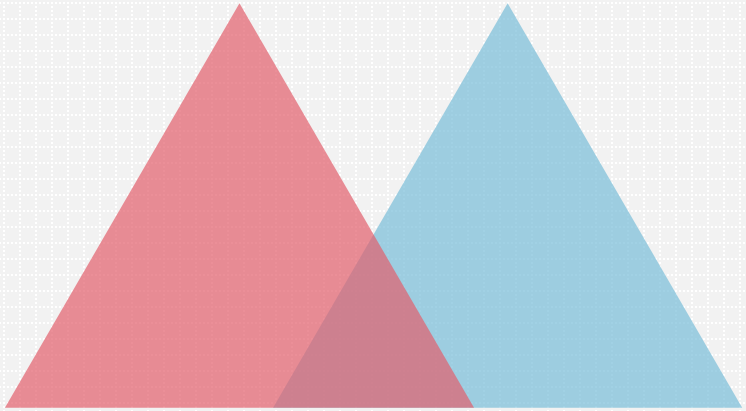
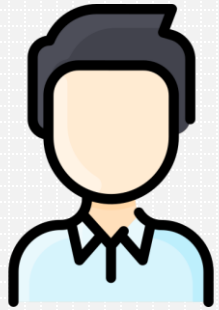
대한민국
유통 활성화를 위한
적요 표준화

INBIG 팀 이원기, 박병현, 김성아, 강동연, 서혜빈

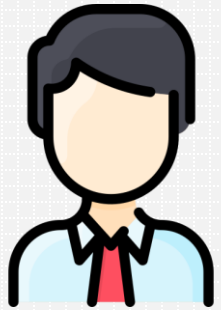


00 팀원 소개



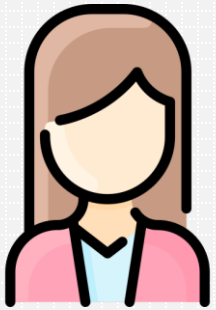
이원기

인하대학교 공간정보공학과



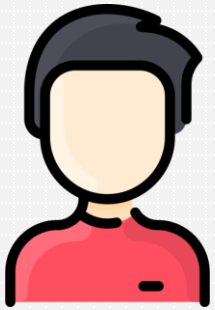
박병현

인하대학교 컴퓨터공학과



김성아

인하대학교 통계학과

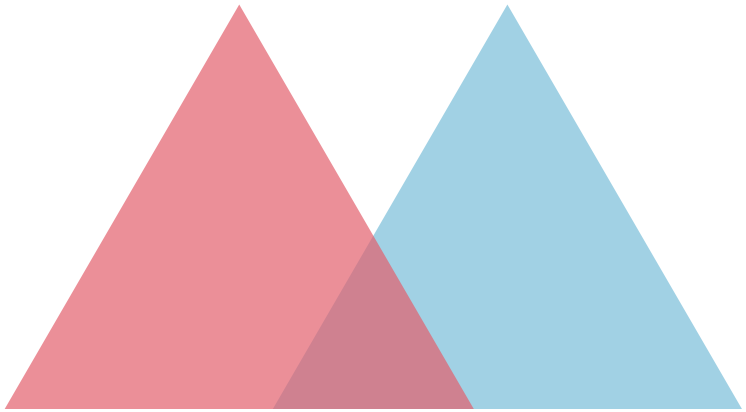


강동연

인하대학교 경영학과

서혜빈

인하대학교 정보통신공학과



INDEX

01 적요 데이터셋 생성

(1) 114개의 적요 csv 결합

```
def makeDataFrame():

    #디렉토리를 아래와 같이 설정
    dirPath = '/home/jovyan/WIDE_HOME/Competition/data/경진대회데이터/품목적요데이터'
    if os.getcwd() != dirPath:
        os.chdir(dirPath)

    # 적요 데이터프레임 담을 빈 데이터프레임 생성
    df = pd.DataFrame()

    #모든 csv 하나의 데이터프레임으로 결합
    for i in tqdm([csv for csv in os.listdir() if'적요' in csv]):
        df = df.append(pd.read_csv(i), ignore_index=True)

    return df
```

(2) 522,266개의 행을 가진 데이터프레임 생성

```
df.tail()
```

	구매/판매구분	판매구매자_업종대분류	판매구매자_업종중분류	판매구매자_업종소분류	적요
5222262	2.0	부동산업 및 임대업	부동산업	부동산 임대업	상차비
5222263	2.0	부동산업 및 임대업	부동산업	부동산 임대업	음식물 수거함 외
5222264	2.0	부동산업 및 임대업	부동산업	부동산 임대업	기계실 동 급수펌프 모터교체공사
5222265	2.0	부동산업 및 임대업	부동산업	부동산 임대업	군산신도시 아파트 정밀점검
5222266	2.0	부동산업 및 임대업	부동산업	부동산 임대업	워터펌프조합 외종 교체

01 적요 데이터셋 생성

▼ 생성된 데이터프레임 정제

```
def refineDataFrame(df):

    #정규표현식 및 , 양 끝 공백 제거
    re_pattern = re.compile(r'[^ 가-힣a-zA-Z0-9]')
    df['적요'] = df['적요'].apply(lambda x : re.sub(pattern = re_pattern, repl = '', string = str(x)))
    df['적요'] = df['적요'].apply(lambda x : x.strip())

    #mecab 형태소분석기를 사용하여 명사 추출
    m = Mecab()
    df['적요Noun'] = df['적요'].apply(lambda x : m.nouns(x))

    #index 초기화
    df = df.reset_index()
    df.drop(['Unnamed: 0', 'index'], axis = 1, inplace = True)

    return df
```

(1) 정규표현식 및 정제
- 한글, 영어, 숫자 제외 모두 공백 처리
- 양쪽 끝 공백 제거

(2) Mecab 형태소 분석기로 적요 Column에서 명사 추출

(3) Index 초기화 및 불필요한 Column 제거

01 적요 데이터셋 생성

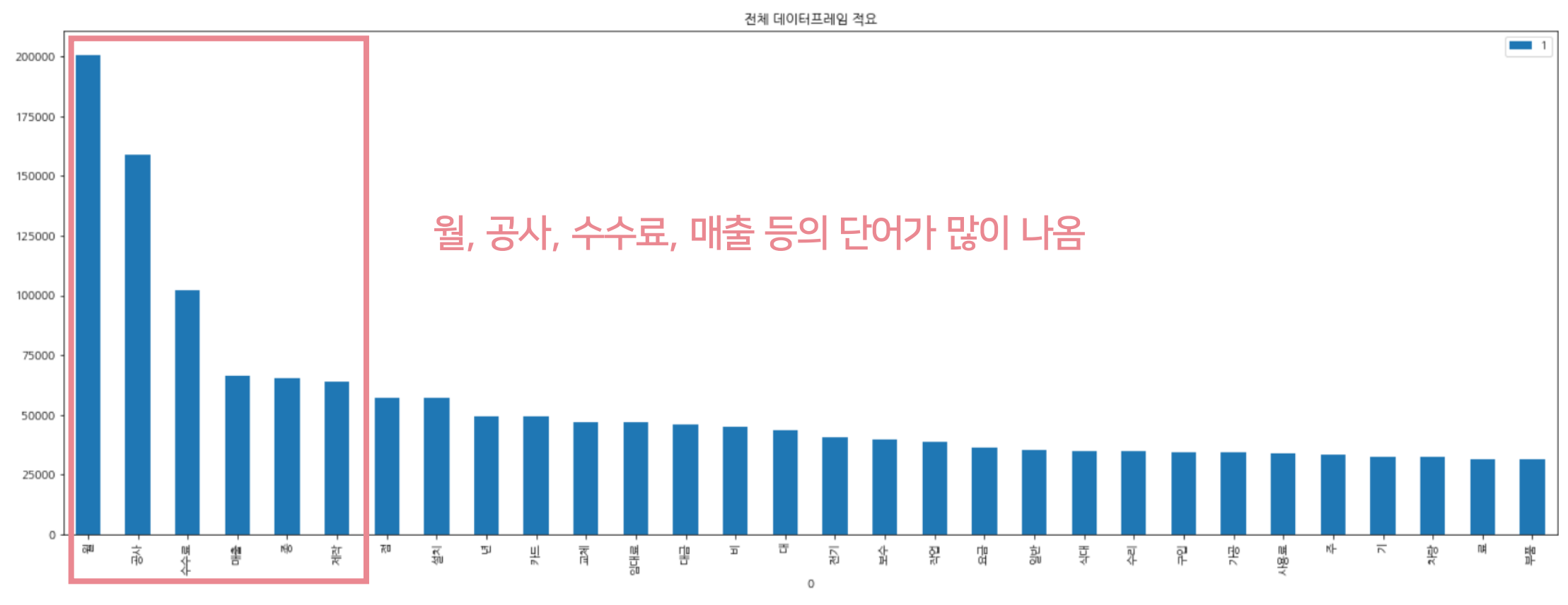
▼ 데이터프레임에서 Mecab 형태소 분석기로 **명사 추출**

적요
오프라인 월 신용카드매출
상품
수족관용품외
필그린 외부여과기 EF 외 건
수족관용품 외
바닥재 외
매출일반지식쇼핑ALLAT신용카드
KS 뉴엘 B사각 H
전자상거래수수료계좌이체
택배운송료

적요Noun
[오프라인, 월, 신용, 카드, 매출]
[상품]
[수족관, 용품]
[필, 외부, 여과기]
[수족관, 용품]
[바닥재]
[매출, 일반지식, 쇼핑, 신용, 카드]
[뉴엘, 사각]
[전자, 상거래, 수수료, 계좌, 이체]
[택배, 운송료]

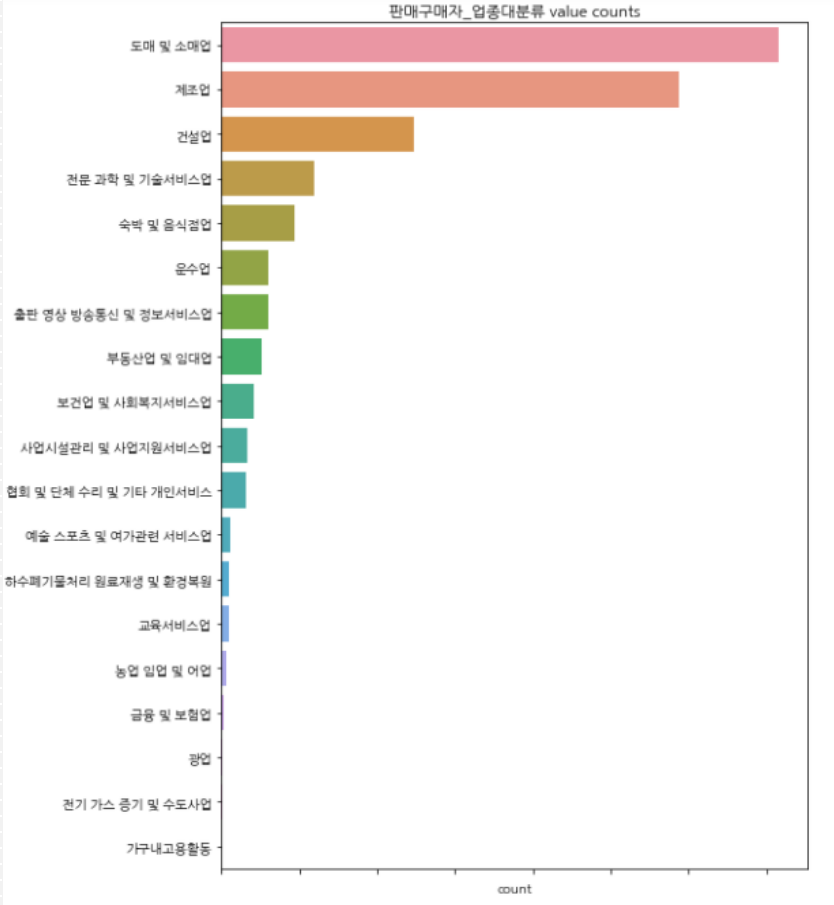
01 적요 데이터셋 생성

▼ '적요' 에서 명사 출현빈도 시각화

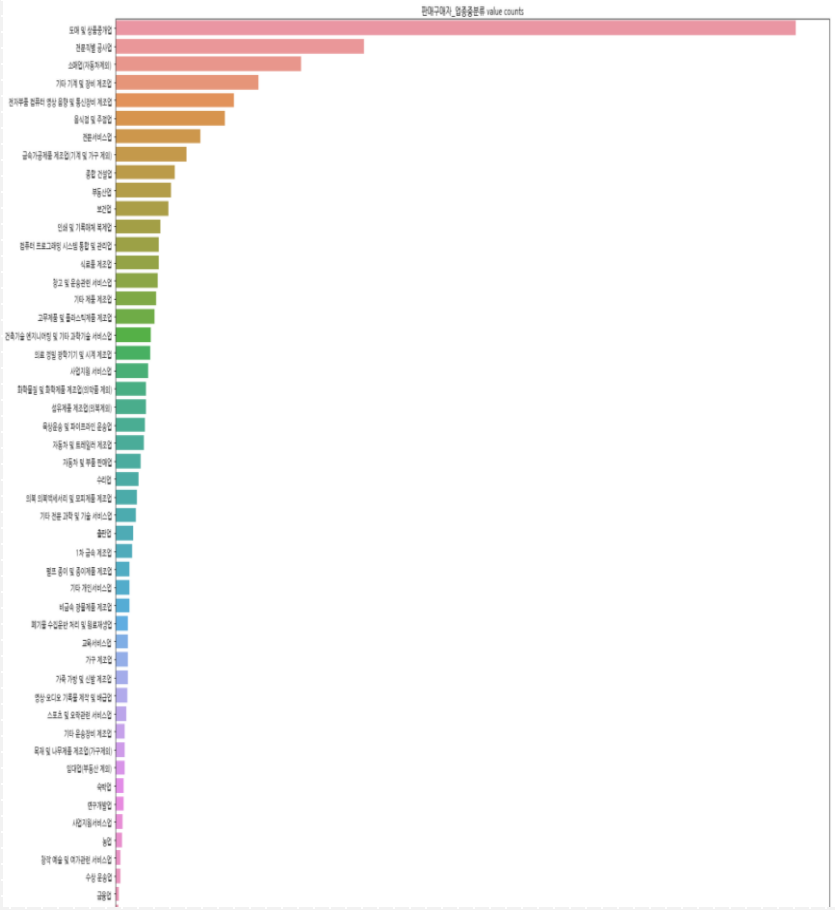


01 적요 데이터셋 생성

▼ '판매구매자_업종대분류' 별 count 시각화



▼ '판매구매자_업종중분류' 별 count 시각화



01 적요 데이터셋 생성

▼ 적요에 '커피'라는 단어가 들어가는 데이터

```
df[df['적요'].str.contains('커피')]
```

	구매/판매구분	판매구매자_업종대분류	판매구매자_업종중분류	판매구매자_업종소분류	적요	적요Noun
408	2.0	건설업	종합 건설업	토목시설물 건설업	커피외	[커피]
508	2.0	전문 과학 및 기술서비스업	전문서비스업	경영컨설팅 및 공공관계 서비스업	커피음료전문점	[커피, 음료, 전문점]
526	2.0	전문 과학 및 기술서비스업	전문서비스업	경영컨설팅 및 공공관계 서비스업	커피머신 구입	[커피, 머신, 구입]
541	2.0	전문 과학 및 기술서비스업	전문서비스업	경영컨설팅 및 공공관계 서비스업	커피재료	[커피, 재료]
555	2.0	전문 과학 및 기술서비스업	전문서비스업	경영컨설팅 및 공공관계 서비스업	원두커피	[원두, 커피]
...	...	...	...	...	...	...
5218376	2.0	제조업	식료품 제조업	도축업	커피재료외	[커피, 재료]
5219486	2.0	제조업	고무제품 및 플라스틱제품 제조업	기타 플라스틱제품 제조업	밀크커피	[밀크, 커피]
5219646	2.0	제조업	고무제품 및 플라스틱제품 제조업	기타 플라스틱제품 제조업	맥심 아이스 커피믹스 외 건	[맥심, 아이스, 커피믹스]
5219982	2.0	제조업	고무제품 및 플라스틱제품 제조업	기타 플라스틱제품 제조업	설탕커피	[설탕, 커피]
5221805	2.0	숙박 및 음식점업	음식점 및 주점업	일반 음식점업	커피	[커피]

20924 rows × 6 columns

커피머신, 커피재료, 맥심커피 등 다양한 데이터가 존재함

01 적요 데이터셋 생성

▼ 적요에 '피자'라는 단어가 들어가는 데이터

```
df[df['적요'].str.contains('피자')]
```

	구매/판매구분	판매구매자_업종대분류	판매구매자_업종중분류	판매구매자_업종소분류	적요	적요Noun
1996	2.0	숙박 및 음식점업	음식점 및 주점업	기타 음식점업	광고비파파존스피자	[광고비, 파파, 존스, 피자]
1998	2.0	숙박 및 음식점업	음식점 및 주점업	기타 음식점업	피자박스	[피자, 박스]
2000	1.0	숙박 및 음식점업	음식점 및 주점업	기타 음식점업	피자 음료외	[피자, 음료]
4675	1.0	숙박 및 음식점업	음식점 및 주점업	기타 음식점업	피자빵 품목외	[피자, 빵, 품목]
13139	2.0	숙박 및 음식점업	음식점 및 주점업	일반 교습 학원	도미노피자	[도미노피자]
...	...	...	...	...	...	...
5214569	2.0	제조업	전자부품 컴퓨터 영상 음향 및 통신장비 제조업	컴퓨터 제조업	도미노피자 등서초점	[도미노피자, 등서, 초점]
5214717	2.0	제조업	전자부품 컴퓨터 영상 음향 및 통신장비 제조업	컴퓨터 제조업	도미노피자 논현점	[도미노피자, 현점]
5214778	2.0	제조업	전자부품 컴퓨터 영상 음향 및 통신장비 제조업	컴퓨터 제조업	도미노피자 명동점	[도미노피자, 동점]
5214797	2.0	제조업	전자부품 컴퓨터 영상 음향 및 통신장비 제조업	컴퓨터 제조업	미스터피자충무로역점	[미스터피자, 충무로, 역점]
5215347	2.0	제조업	전자부품 컴퓨터 영상 음향 및 통신장비 제조업	컴퓨터 제조업	번가피자신서귀점	[번가, 피자, 신서, 귀점]

2749 rows × 6 columns

광고, 박스, 피자빵, 피자집 체인점 등 다양한 데이터가 존재함

전체적인 분석 Flow

Embedding을 위한 학습 데이터 구축	적요 데이터 Embedding	적요 군집화	군집 Labeling
→ 명사 5음절 이상인 적요 데이터 활용	→ 각 행들의 적요 Embedding	→ 1. 구매/판매구분코드, 2. 적요데이터 (200차원) 3. 업종 중분류 위 세가지 변수를 활용한 군집화	→ 각 군집별 특성을 고려한 Labeling
→ 네이버 사전 Crawling - 회계, 재무, 음식레시피 등			
→ 공정거래위원회 정보제공 시스템 - 프랜차이즈 음식점 및 편의점 등의 데이터 구축	→ FastText 모델을 통한 학습	→ K-means Clustering	→ t-SNE를 통한 군집별 시각화

WordEmbedding 이란?

WordEmbedding

→ 특정 단어가 문장 내의 주변 단어들을 이용해
각 단어들의 의미를 예측해 vector화 시키는 과정

→ 비정형화된 Text를 숫자로 바꿈으로써
사람의 언어를 컴퓨터의 언어로 번역하는 것

FastText

→ Facebook 에서 개발한 Embedding 기법

FastText 사용 이유

(1) 학습에 사용되지 않은 단어(Out Of Vocabulary, OOV)
에 대해서도 임베딩 가능

(2) 다른 임베딩 기법보다 빠른 학습 속도

▼ 대표적인 Embeding 모델

Word2Vec

Glove

FastText

03 Word Embedding을 위한 학습데이터 구축

(1) 적요 5음절 이상 데이터 활용

한 번에 학습하는 단어의 양

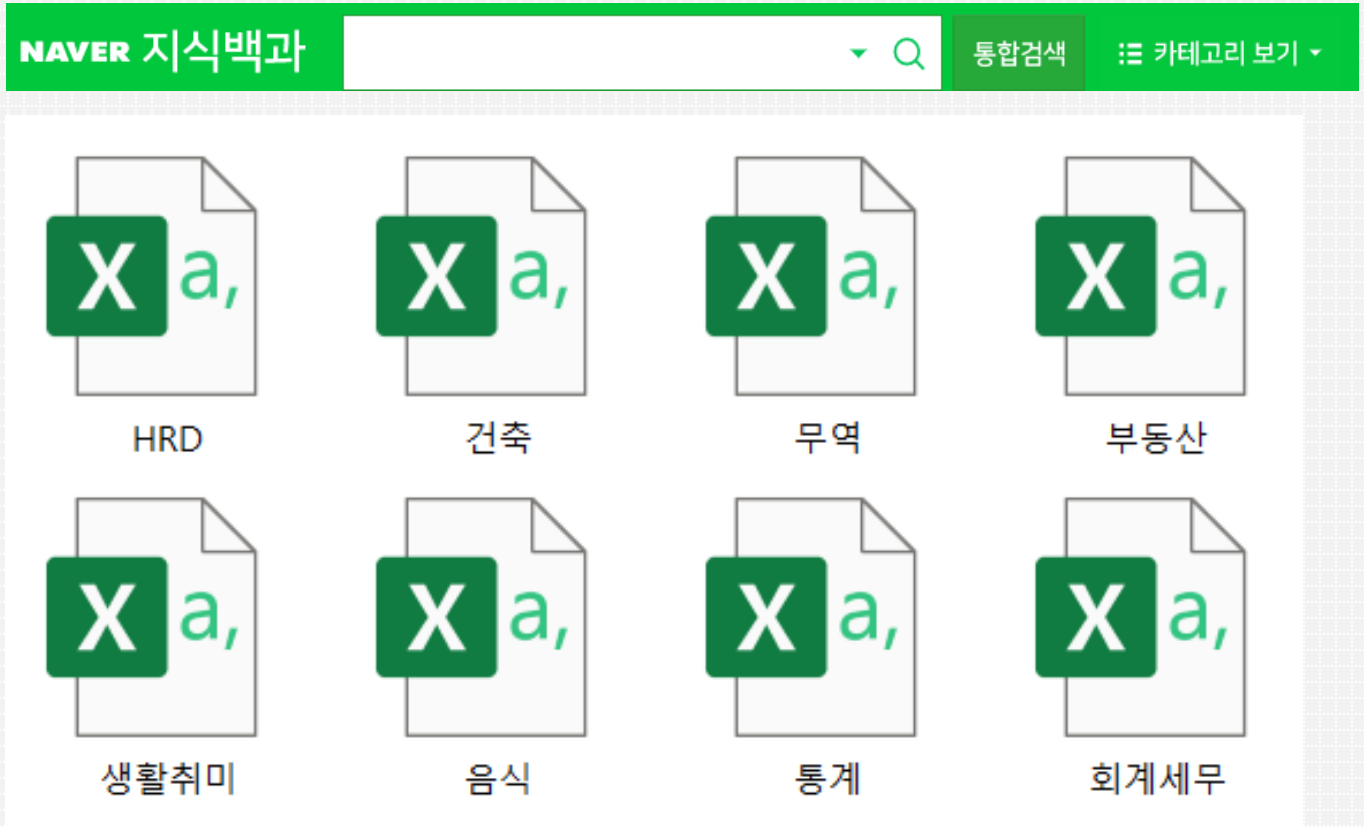
FastText의 Parameter 값인 ' Window ' 값 5로 설정

```
wordOver5.tail(10)
```

	구매/판매구분	판매구매자_업종대분류	판매구매자_업종중분류	판매구매자_업종소분류	적요	적요Noun
5222206	2.0	부동산업 및 임대업	부동산업	부동산 임대업	인버터 파워부 불량 교체	[인버터, 파워, 부, 불량, 교체]
5222224	2.0	부동산업 및 임대업	부동산업	부동산 임대업	소방시설종합정밀점검	[소방, 시설, 종합, 정밀, 점검]
5222231	2.0	부동산업 및 임대업	부동산업	부동산 임대업	층표시기커플링루버 교체	[층, 표시기, 커플링, 루버, 교체]
5222233	2.0	부동산업 및 임대업	부동산업	부동산 임대업	설비교체작업급수펌프	[설비, 교체, 작업, 급수, 펌프]
5222234	2.0	부동산업 및 임대업	부동산업	부동산 임대업	승강기벽면 페인트 작업 현수막	[승강기, 벽면, 페인트, 작업, 현수막]
5222237	2.0	부동산업 및 임대업	부동산업	부동산 임대업	어린이놀이시설 검사수수료	[어린이, 놀이, 시설, 검사, 수수료]
5222245	1.0	부동산업 및 임대업	부동산업	부동산 임대업	가스정압시설 부지 임대료	[가스, 정압, 시설, 부지, 임대료]
5222264	2.0	부동산업 및 임대업	부동산업	부동산 임대업	기계실 동 급수펌프 모터교체공사	[기계실, 동, 급수, 펌프, 모터, 교체, 공사]
5222265	2.0	부동산업 및 임대업	부동산업	부동산 임대업	군산신도시 아파트 정밀점검	[군산, 도시, 아파트, 정밀, 점검]
5222266	2.0	부동산업 및 임대업	부동산업	부동산 임대업	워터펌프조합 외종 교체	[워터, 펌프, 조합, 외종, 교체]

03 Word Embedding을 위한 학습데이터 구축

(2) 네이버 사전



→ 파이썬에서 Crawling

→ Crawling 데이터 정제 및 전처리

→ Mecab 형태소 분석기로 명사 추출

03 Word Embedding을 위한 학습데이터 구축

(3) 공정거래위원회 가맹사업거래

번호	상호	영업표지	대표자	등록번호	업종
1	마라루	마라루	권화숙	20200225	중식
2	(주)스시히로미	스시노칸도	이지훈, 이형락	20150882	일식
3	마시기통차	마시기통차	안진근	20130100277	한식
4	제이에스푸드(JS푸드)	꽃차돌	서정익	20180916	한식
5	스시카츠	스시카츠	강명구	20190177	일식
6	WyndhamHotelAsiaPacificCo.Limited	데이즈	Eng San QUEK	20150059	숙박
7	더블유엠에스테틱시지점	더블유엠에스테틱두피센터	전계선	20180426	이미용
8	바보아재막창	바보아재막창	설동철	20180613	한식
9	(주)루아	루아 저세상 볼쭈꾸미	김현규	20201406	기타 외식
10	제이에스푸드(JS푸드)	강대포식당	서정익	20190723	주점

03 Word Embedding을 위한 학습데이터 구축

▼ FastText에 사용할 626,113개의 학습데이터셋 구축

0	[건강, 생각, 요즘, 다이어트, 관심, 샐러드, 열량, 영양, 열량식, 애용, 샐...
1	[입맛, 때, 음식, 입맛, 때, 꽈리고추, 볶음, 제격, 매운맛, 여러분, 입맛,...
2	[버본, 경마, 고장, 캔터, 키, 풍경, 속, 추억, 무대, 캔터, 키, 경마, ...
3	[주재료, 스위스, 와인, 화이트, 와인, 컵, 멘탈, 치즈, 튀, 에르, 치즈, ...
4	[어복쟁반, 놋, 쟁반, 갖가지, 고기, 편육, 채소, 사람, 육수, 추위, 일종,...
...	...
626108	[승강기, 벽면, 페인트, 작업, 현수막]
626109	[어린이, 놀이, 시설, 검사, 수수료]
626110	[기계실, 동, 급수, 펌프, 모터, 교체, 공사]
626111	[군산, 도시, 아파트, 정밀, 점검]
626112	[워터, 펌프, 조합, 외종, 교체]

626113 rows × 1 columns

03 Word Embedding을 위한 학습데이터 구축

▼ FastText 모델을 통해 , 각 행의 적요데이터를 200차원으로 변환

Parameter
- Size = 200
- Window = 5

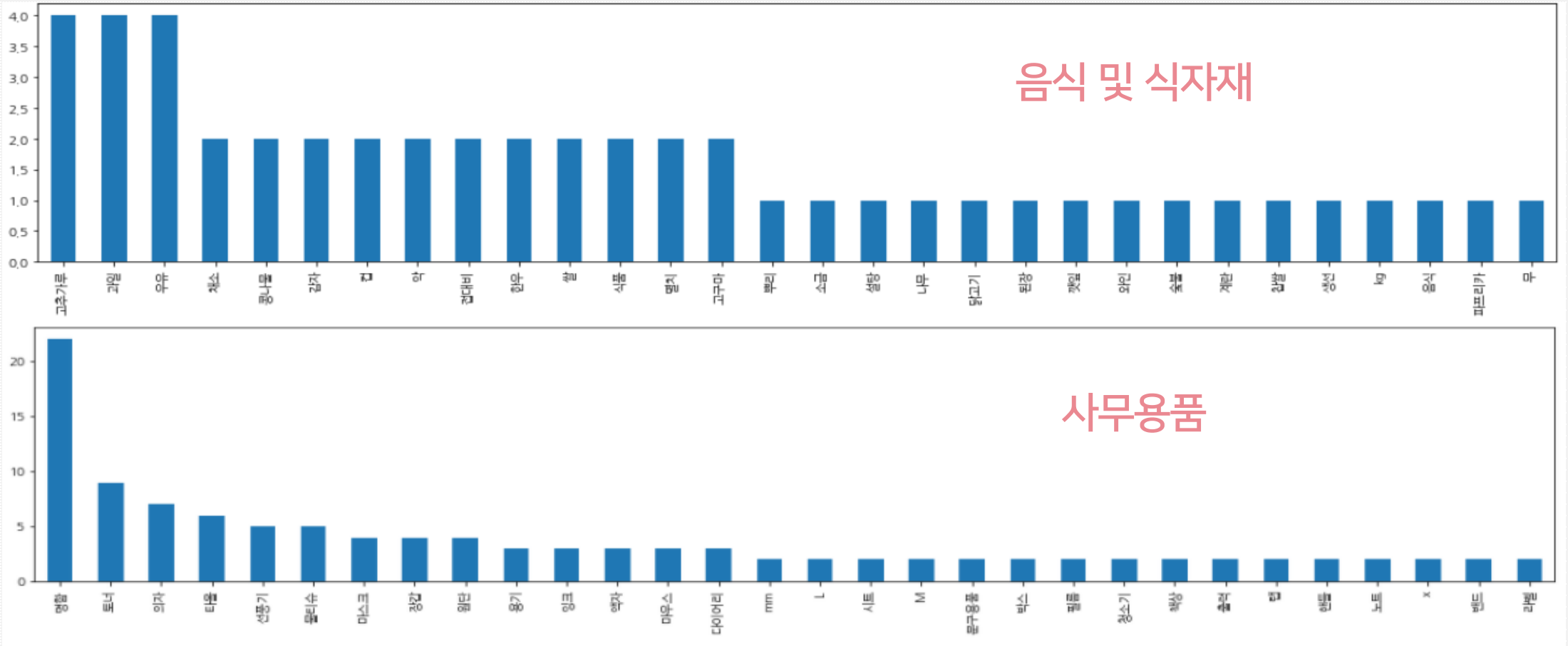
	0	1	2	3	4	5	6	7	8	9	...	193	194	195	196	197	198	199	구매/판매구분	판매구매자_업종중분류	적요
0	0.084362	-0.075441	-0.071083	0.037512	-0.088706	-0.158561	-0.075895	0.032675	0.037972	-0.048111	...	-0.149829	-0.040984	0.068560	0.046596	0.166463	0.004435	0.007909	1.0	건축기술 엔지니어링 및 기타 과학기술 서비스업	유공블럭
1	0.047755	-0.018198	0.011152	0.042948	-0.057102	-0.027782	0.003616	0.003856	-0.007079	-0.010749	...	0.008846	-0.006856	0.061941	0.027266	0.015732	0.041743	0.044895	1.0	건축기술 엔지니어링 및 기타 과학기술 서비스업	세대통합박스 외
2	-0.001081	-0.017972	-0.003601	0.002613	-0.014677	-0.001292	0.002152	-0.005143	-0.002944	-0.014857	...	-0.008173	-0.003485	0.007013	-0.001464	0.000686	-0.000132	0.015577	1.0	건축기술 엔지니어링 및 기타 과학기술 서비스업	G 상하판외
3	0.097306	-0.210276	0.124873	0.233049	-0.183333	-0.195097	0.015446	0.102707	0.216890	-0.298953	...	-0.204705	-0.004620	0.325190	-0.465051	-0.127514	0.079008	-0.021594	2.0	건축기술 엔지니어링 및 기타 과학기술 서비스업	모니터인치
4	-0.048982	-0.325185	0.582175	0.157065	-0.951694	-0.347834	1.329734	-0.994042	-0.506865	-1.266335	...	-0.460482	0.436430	0.684176	-0.814207	-0.012128	-0.194489	1.197356	1.0	소매업(자동차제외)	카드매출
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
299996	0.000907	-0.065849	-0.026493	-0.017683	-0.045024	0.043475	-0.030202	0.015759	-0.043594	-0.048908	...	-0.034353	-0.026894	0.014929	-0.019772	-0.011763	-0.009473	0.061216	1.0	인쇄 및 기록매체 복제업	대한예방의학회 제자 가을학술대회 심포지엄 일자의료
299997	0.030924	-0.046883	-0.041859	-0.006957	-0.036744	0.018819	-0.028140	-0.019224	-0.048544	-0.018129	...	-0.018849	-0.029109	0.012887	-0.024310	0.022991	-0.018401	0.074010	1.0	인쇄 및 기록매체 복제업	제회 적국체육대회 해단식 조청장 및 봉투
299998	0.012382	-0.015673	-0.062696	-0.027873	-0.055307	0.022745	-0.041068	0.017477	-0.082348	-0.067756	...	-0.071736	-0.032739	0.020383	-0.017582	-0.013427	-0.021503	0.117291	1.0	인쇄 및 기록매체 복제업	학년도 제주년 종교개혁제
299999	0.029835	-0.034879	0.019496	-0.001116	-0.030963	0.008634	0.008280	0.016658	0.001341	-0.011447	...	-0.037210	-0.024346	0.004664	-0.016114	0.003535	-0.004641	0.010754	1.0	인쇄 및 기록매체 복제업	자문위원 연수 명찰
300000	-0.010538	-0.462754	0.292087	-0.184036	0.270367	0.183247	-0.031315	0.114517	-0.030514	-0.491021	...	-0.088051	-0.257052	0.131146	-0.108359	0.111960	-0.073467	0.118487	1.0	인쇄 및 기록매체 복제업	광고 인쇄

300001 rows × 203 columns

벡터화 (Embedding)

04 Clustering – K-means

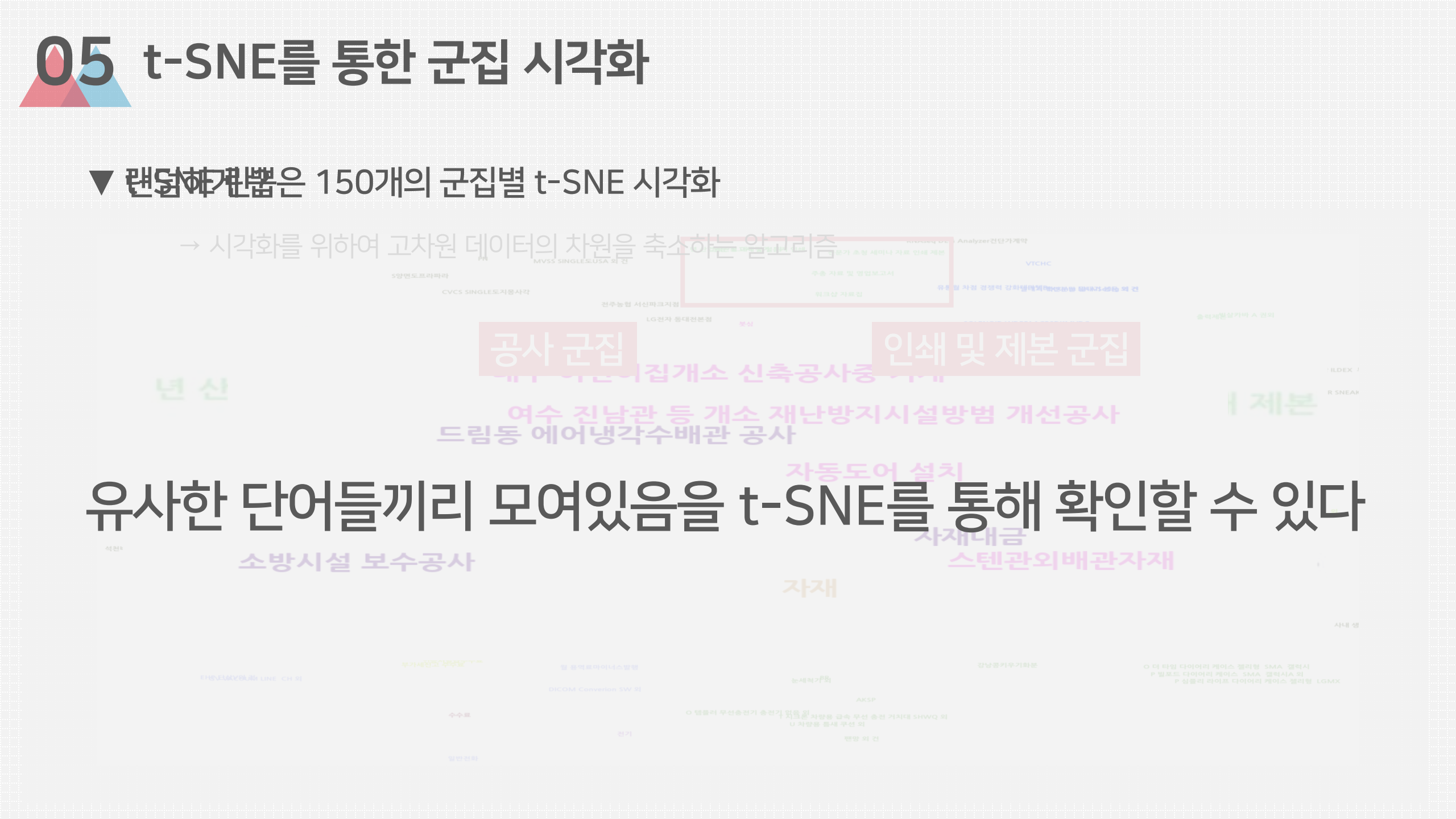
▼ 군집화가 잘 된 군집 내의 적요 Value_counts()



05 t-SNE를 통한 군집 시각화

▼ 랜덤하게 뽑은 150개의 군집별 t-SNE 시각화

→ 시각화를 위하여 고차원 데이터의 차원을 축소하는 알고리즘



유사한 단어들끼리 모여있음을 t-SNE를 통해 확인할 수 있다

THANK YOU