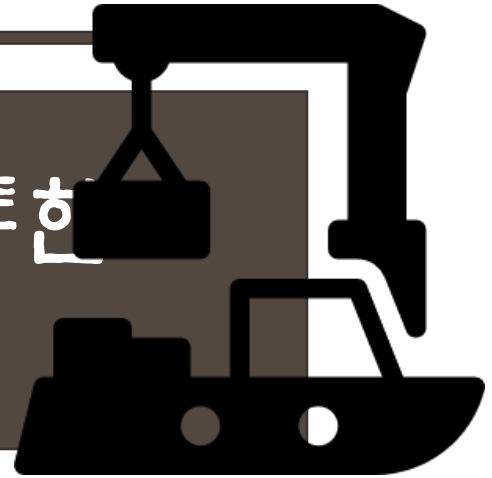
제8회 빅데이터 경진대회

항만물류 데이터 정형화를 통한 비즈니스 절차 개선



팀원 :김정현,박건우,양명철,위혜린

PROJECT STEPS

01 분석 배경 및 목적

02 분석 과정

03 분석 결과

04 데이터 활용 방안

연구 배경

품명의 분류 기준과 집계 과정에서 문제가 발생 , 통계 오류의 주된 원인이 발생.
➔상품 분류 체계의 통일을 기하여 일관성 있는 품목 분류

울산항: 대한민국 액체 화물 처리 1위 항만

2018년 액체화물 처리량

구분		2018년
총 물동량(천톤)		202,862
액체 화물	소계	166,594
	원유(역청류),석유	71,580
	석유정제품	51,813
	석유가스	8,152
	케미칼	35,049

국내 액체화물의 34.3% , 도입 원유의 47.7%를 처리.

1 품목 분류의 중요성

- 수출입물품에 대한 통관 및 승인요건 뿐만 아니라 물품의 원산지 및 FTA 양허대상 여부 등을 결정하는 핵심적인 요소
- 품목별 상세 분류가 함께 동반 돼야 HS코드의 명확한 분류 기준을 정립, 수입 항만 통계 수치의 정확도를 높일 수 있음.

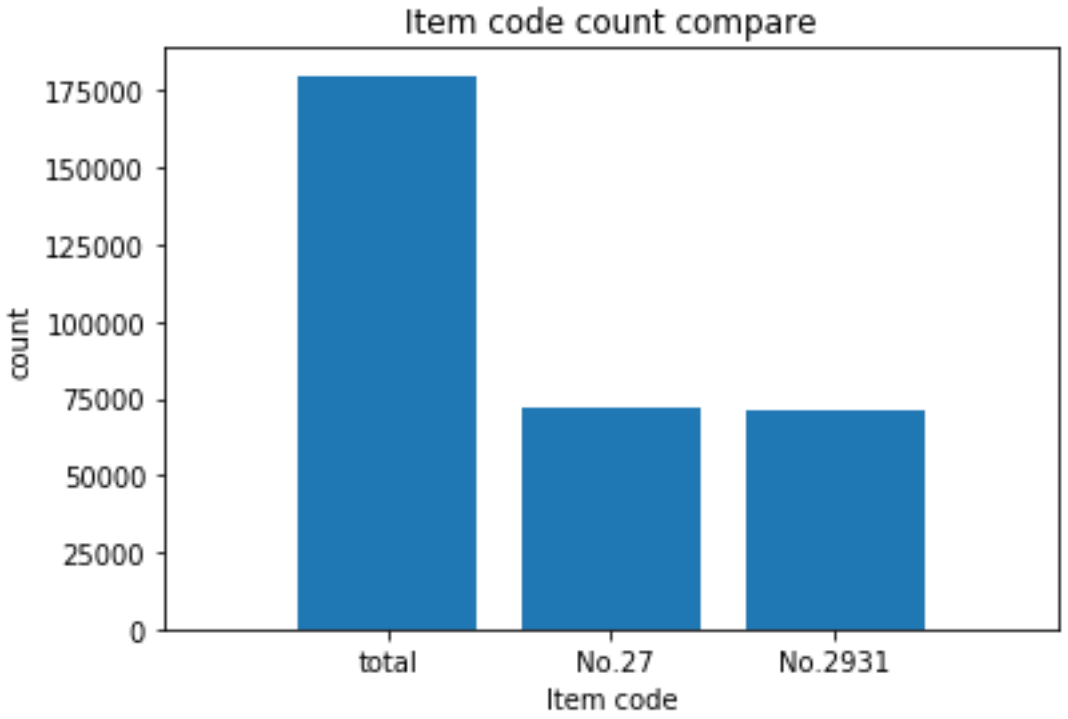
2 HS코드별 분류 체계 이용시

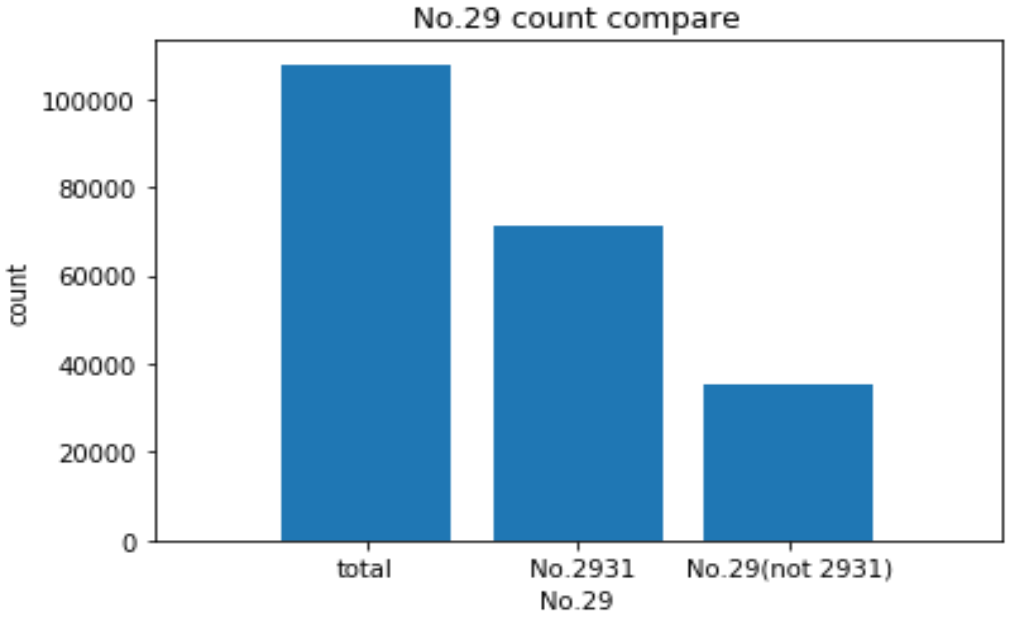
- 자체사정을 고려하여 자율적으로 더 세분화하여 사용할 수 있음.

사용되는 데이터 SET+ 전처리

코드	품명
27****	제 27류 광물성연료, 광물류와 이들의 증류물, 역청물질 및 광물성 왁스
2931**	그 밖의 유기· 무기화합물

- 2007년~2018년 데이터
- 품목코드와 품목명 분류





"품목 코드 29**** 비교 후 2931** 데이터 도출"

사용되는 데이터 SET+ 전처리

No. 271000

(41684, 13)

왜 품목 코드를 나누었는가?

정확도를 위해서 큰 데이터 SET에서 군집화를 하지 않고 품목코드로 1차 분류 후 군집화

LUBRICATINGBASEOIL150N
NEODOL25-7IMONAME:ALCOHOL(C12-C1(
GASOLINE
STOCK4733LUBRICATINGOIL(PIBSUPPLYFO
LUBEBASEOIL500SOLVENTNEUTRAL(2)
SBCHVIBASEOIL3
GASOIL0.5PCTSULPHUR
LOWSULPHURFUELOIL

데이터 분석을 위해,
대소문자 구분 통일 + 띄어쓰기 삭제

패스트텍스트(FastText)

1 모르는 단어(Out of Vocabulary)에 대한 대응

각 단어는 글자들의 n-gram으로 나타냄. N을 몇으로 결정하는지에 따라서 단어들이 얼마나 분리되는 지 결정됨.
예를 들어 n을 3으로 잡은 트라이그램(tri-gram)의 경우, apple은 app,ppl,ple로 분리하고 이들 또한 임베딩을 함.

2 단어 집합 내 빈도 수가 적었던 단어(Rare Word)에 대한 대응

모든 훈련 코퍼스에 오타 (Typo) 나 맞춤법이 틀린 단어가 없으면 이상적이겠지만, 실제 많은 비정형 데이터에는 오타가 섞여 있음. 그리고 오타가 섞인 단어는 당연히 등장 빈도수가 매우 적으므로 일종의 희귀 단어가 됨.

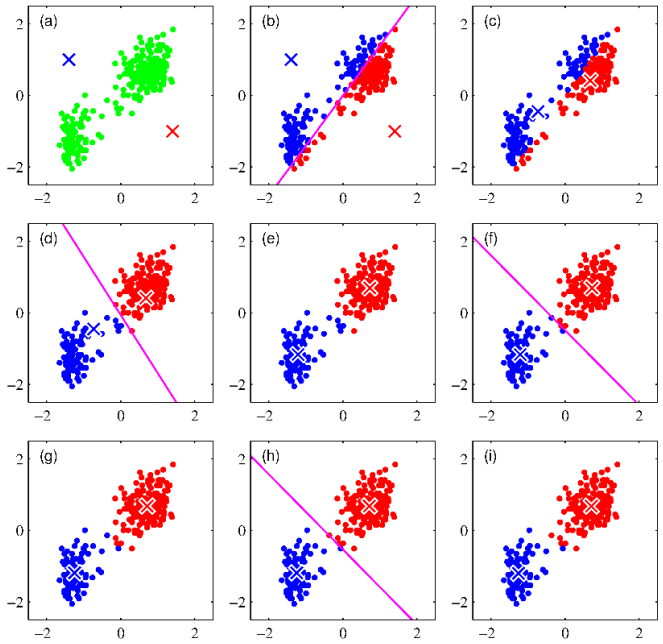
3 사전 훈련된 임베딩(Pre-trained FastText Embedding) 사용

사전에 훈련된 워드 임베딩을 갖고 와서 사용하거나, 갖고온 것에 추가 학습을 하는 방식으로 사용.
이는 word2Vec로 학습을 하든 글로브로 학습을 하든 마찬가지임. 현재 페이스북의 패스트텍스트는 294개 언어에 대하여 위키피디아로 학습한 사전 훈련된 벡터들을 제공함.

K-평균 군집화(K-means Clustering)

1 알고리즘 개요

대표적인 분리형 군집화 알고리즘. 각 군집은 하나의 중심(centroid)을 갖음. 각 개체는 가장 가까운 중심에 할당
각 개체는 가장 가까운 중심에 할당되며, 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성함.
사용자가 사전에 군집 수(k)를 정해야 알고리즘을 실행할 수 있음. 즉, k가 하이퍼파라메터(hyperparameter)임.



2 수렴해가는 과정

$$X = C_1 \cup C_2 \ldots \cup C_K, \quad C_i \cap C_j = \phi$$

$$argmin_C \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - c_i\|^2$$

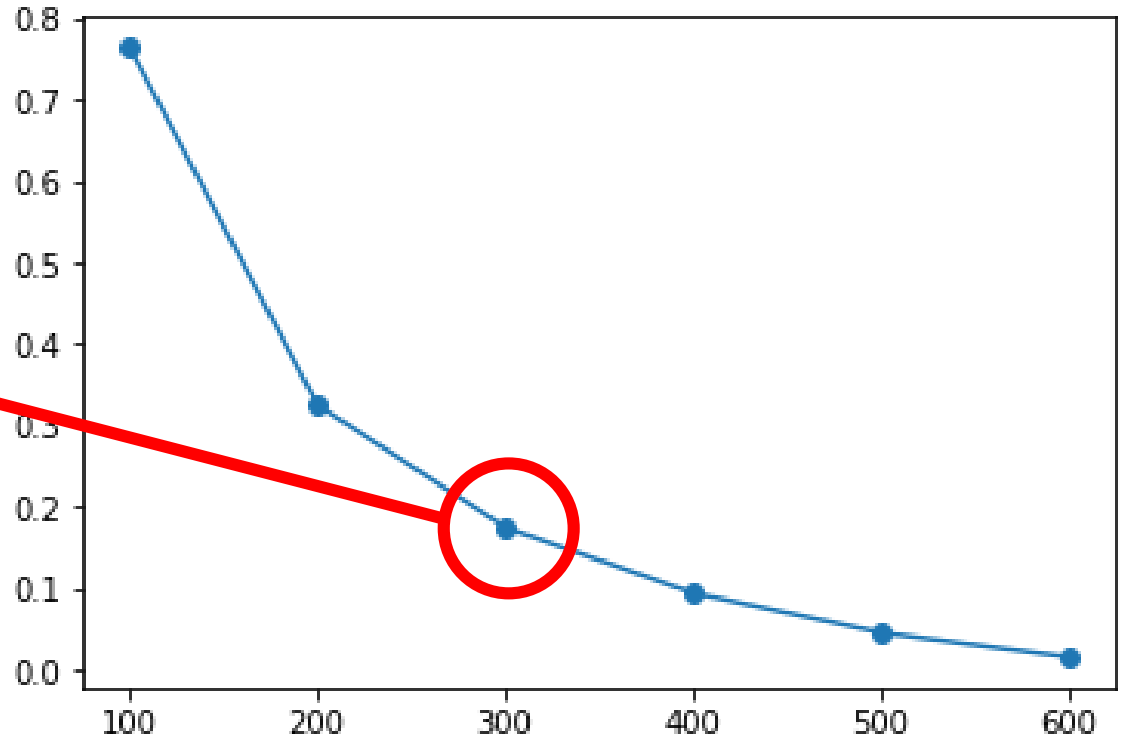
- 우선 2개의 클러스터를 구성한다고 결정한 상태.
- 임의의 2점을 평균 값으로 하여 E(expectation)와 M(Maximization)단계를 반복.
- 그림 순서대로 평균 값을 계산한 뒤, 각각의 샘플에 대해 어느 클러스터에 속할 지 할
- 이렇게 할당된 데이터를 기준으로 다시 평균 값을 계산.
- 수렴 조건이 만족할 때까지 반복함.

편집 거리 알고리즘 + Clustering

```
AUTOMOTIVEDIESELFUEL10PPMSULPHUR
AUTOMOTIVEDIESELFUEL10PPMSULPHUR
AUTOMOTIVEDIESELFUEL10PPMSULPHUR
AUTOMOTIVEDIESELFUEL10PPMSULPHUR
AUTOMOTIVEDIESELFUEL10PPMSULPHUR
AUTOMOTIVEDIESELFUEL10PPMSULPHUR
GASOIL0.05PCTSULPHUR
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LOWSULPHURWAXYRESIDUE
LUBRICATINGBASEOIL(BASEOILPREMIUM150N)
TOLUENE(COALTARBASE)
TOLUENE(COALTARBASE)
TOLUENE(COALTARBASE)
TOLUENE(COALTARBASE)
DOWANOLTHPMGLYCOLETHER
ULTRALOWSULFURDIESEL
ULTRALOWSULFURDIESEL
ULTRALOWSULFURDIESEL
ULTRALOWSULFURDIESEL
NORMALPARAFFIN(C10-13)
```

Elbow : K → 300

Result : 같은 군집 내 동질성이 떨어짐



편집 거리 알고리즘(Levenshtein distance)

유사도 판단 기준 : 두 개의 문자열이 같아지기 위해서 몇 번의 추가(Add), 편집(Edit), 삭제(Delete)가 이루어

져야 하는 지 그 최소 개수를 구해줌.

❑ 한 string을 s1을 s2로 변환하는 최소 횟수를 두 string간의 거리로 정의함

1. Delete : '점심을먹자 → 점심먹자 ' (을 삭제)
2. Insert: '점심먹자 → 점심을먹자' (을 삽입)
3. Substitution: '점심먹자→ 점심먹장' (자→ 장으로 편집)

❑ 동적 프로그래밍 : 특정한 문제를 잘게 쪼개어 작은 부분부터 천천히 해결해 나가는 것

	ø	a	b	c	d	e	f
ø	0	1	2	3	4	5	6
a	1						
z	2						
c	3						
e	4						
d	5						

거리(Distance): 현재 위에서 숫자들이 의미하는 바
공집합(아무 것도 없는 상태)와 a의 거리는 1이고 공집합과 ab의 길이는 2.
이것들이 차례대로 반영되어 공집합과 abcdef의 거리는 6.
이 기본적인 데이터를 토대로 모든 알고리즘이 진행됨.

주성분분석(Principal Component Analysis)

데이터의 분산(Variance)을 최대한 보존하면서 서로 직교하는 새 기저(축)을 찾아, 고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간으로 변환하는 기법

❑ PCA의 목적

데이터의 벡터 공간 차원은 엄청나게 크지만 실제로 필요한 true data는 작은 차원 공간으로 표현해도 충분한 경우에 사용

❑ Feature selection 방법

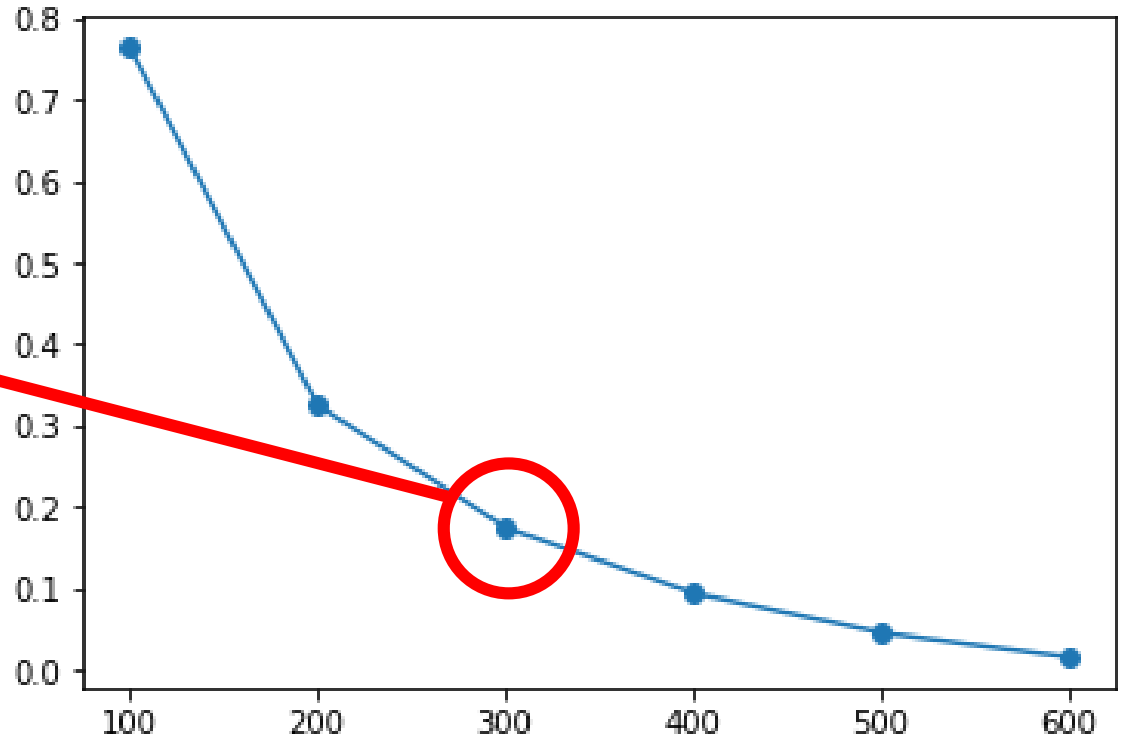
D개의 차원을 가지는 원소들 중 m개만 뽑아서 씀.

4500차원을 공분산을 구하여 28차원으로 줄임.

패스트텍스트(FastText) + clustering

GS150NBASEOIL
SK500NBASEOIL
GS150NBASEOIL
00NBASEOIL
600NBASEOIL
GS150NBASEOIL
GS600NBASEOIL
GS150NBASEOIL
SK500NBASEOIL
SK500NBASEOIL
SK500NBASEOIL
SK500NBASEOIL
SK500NBASEOIL
LUBRICATINGOIL
SK500NBASEOIL
BS150BASEOIL
LUBRICATINGOIL
PALMSLUDGEOIL
SK500NBASEOIL
HSB150NBASEOIL

비슷한 단어끼리 Group화



패스트텍스트(FastText) + clustering + 최빈값

PARAXYLENE	1092	0.963813
AUTOMOTIVEDIESELFUEL-10PPMSULPHUR	179	0.978142
DIPROPYLENEGLYCOL,REGULARGRADE17,2	1	0.142857
MOLTENSULPHURINBULK	172	0.988506
PROPYLENEGLYCOLMONOBUTYLETHER(PN	2	0.133333
LUBRICATINGBASEOIL(BASEOILULTRA-S8(2	39	0.39
ASPHALT60-80PENETRATION	232	0.935484
NAPHTHA	1545	0.992931
LUBRICATINGBASEOIL150N	4	0.190476
NEODOL25-7IMONAME:ALCOHOL(C12-C16	3	0.115385
GASOLINE	927	0.985122
STOCK4733LUBRICATINGOIL(PIBSUPPLYFO	1	0.066667
LUBEBASEOIL500SOLVENTNEUTRAL(2)	11	0.366667
SBCHVIBASEOIL3	16	0.188235
GASOIL0.5PCTSULPHUR	171	0.52454
LOWSULPHURFUELOIL	805	0.996287
BASEOILPREMIUM60N	146	0.561538
PALMFATTYACIDDISTILLATE	75	0.303644
GASOIL10PPM	278	0.76584
SPINDLEOIL60NII	51	0.380597
YUBASE4LUBRICATINGBASEOIL	148	0.319654
JETA-1	1589	0.868781

군집화 list를 뽑아내어, 그 중 최빈값을 정규값(대표값)으로 사용

Clustering(각 군집별 List)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	PARAXYLENE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE	PARAXYLE
1	AUTOMOTIVEDIE	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT	AUTOMOT
2	DIPROPYLENEGLY	STOCK473	STOCK473	LIQUIDPAI	CIFHUANC	ADDITIVES	BASEOILU	LTRAS8(25ON)	ORIGINR	EPUBLICOF	KOREACO	NTRACTNO	BASE06225205P/	ONO4800191938/0001(		
3	MOLTENSULPHUR	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS	MOLTENS
4	POLY(2-8)ALKYLEN	BASEOILU	OTHERPET	PROPYLEN	PROPYLEN	DENATURI	MARPOLA	JAYFLEXDI	LIQUIDPAI	MARPOLX	GASTURBI	GASTURBI	12UNITSO	2UNITSOF	ALIMET(METHIONINI	
5	LUBRICATINGBAS	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI
6	ASPHALT60-80PEN	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6	ASPHALT6
7	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA	NAPHTHA
8	LUBRICATINGBAS	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI	LUBRICATI
9	NORMALRUSSIAN	NORMALR	YUBASE4-(	SUPER-31(	MIXEDARC	ETHYLENE	SLURRYOI	TERTIARY(	TERTIARY(	NAPHTHEI	TRANSFOF	DIMETHYL	TRANSFOF	NAPHTHEI	NAPHTENI	NEODOL2
10	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE	GASOLINE
11	STOCK4733LUBRI(	STOCK473	STOCK473	STOCK473	STOCK473	STOCK473	YUBASE4Y	STOCK473	STOCK473	YUBASE4L	LUBRICAN	YUBASE4L	HSB(DAES	ISOPARHS	BASEOILU	LTRA-S2(60
12	BASEOIL150BS(DI	BASEOIL1!	LUBEBASE	BASEOILJC	SUPER500	GROUPIIB	LUBEBASE	LUBEBASE	LUBEBASE	LUBEBASE	LUBEBASE	LUBEBASE	LUBEBASE	NAPHTHEI	BASEOILU	LUBEBASE
13	GS150NBASEOIL	SK500NBA	GS150NBA	600NBASE	600NBASE	GS150NBA	GS600NBA	GS150NBA	SK500NBA	SK500NBA	SK500NBA	SK500NBA	LUBRICATI	SK500NBA	BS150BAS	LUBRICATI
14	GASOIL1.5PCTSUL	GASOIL1.5	GASOIL1.5	GASOIL0.5	GASOIL0.5	GASOIL0.5	GASOIL1.5	GASOIL1.5	GASOIL0.5	GASOIL1.5	GASOIL0.5	GASOIL1.5	GASOIL1.5	GASOIL0.5	GASOIL0.5	GASOIL0.5
15	LOWSULPHURFUE	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI	LOWSULPI
16	BASEOILPREMIUN	BASEOILPI	BASEOILSI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI	BASEOILPI
17	FULLYREFRIGERAT	FULLYREFI	AH90SKBR	PALMFAT1	FUELOIL(N	KOCOSOL:	EXXSOL9SI	PALMFAT1	FULLYREFI	FULLYREFI	PALMFAT1	FULLYREFI	NOXIOUSL	FULLYREFI	PALMFAT1	OTHERHE/
18	GASOIL10PPM	GASOIL10	GASOIL50	GASOIL10	GASOIL10	GASOIL10	GASOIL10	GASOIL10	GASOIL10	GASOIL10	GASOIL0.(	GASOIL0.(	GASOIL10	GASOIL10	GASOIL10	GASOIL10
19	NEUTRALOIL220R	NEUTRALC	EXTENDER	PARAFENI	NEUTRALC	NEUTRALC	AVIATION	NEUTRALC	NEUTRALC	NEUTRALC	NEUTRALC	AVIATION	AVIATION	AVIATION	FUELOIL3!	VOLATILE(
20	YUBASE4ALUBRIC	YUBASE3L	YUBASE6J	YUBASE6L	YUBASE2L	YUBASE4L	YUBASENC	YUBASE8L	YUBASE4L	SK500NLU	YUBASE4L	YUBASE2L	YUBASE4L	YUBASE6J	YUBASE6L	YUBASE4L
21	JETA1	JETA1	JETA-1	JETA-1	JETA-1	JETA-1	JETA-1	JETA-1	JETA-1	JETA-1	JETA1	JETA-1	JETA-1	JETA-1	JETA-1	JETA-1

각 군집 안에 어떤 품명 list가 들어있는 지 알아냄

How can we use?

- 최빈값과 각 군집에 어떤 list가 있는 지 알아냈을 때 활용 방안

- 검색을 하고 싶을 때,

  검색 값과 정확하게 일치하는 list 뿐 아니라 오타, 자세히 입력된 값 등에 대한 list 확인 가능.

- 표준값이 아닌 다른 값이 입력된다면, 그것이 속할 군집의 최빈값으로 대체하여 자동 입력 가능.

In [121]:
```
a = input()
```
```
METHANOL
```

In [122]:
```
print(a)
```
```
METHANOL
```

In [123]:
```
correct = 0

for i in range (len(data)):
    if a == data["Name"][i]:
        print("올바른 입력입니다.")
        print(a + " 을(를) 입력하셨습니다.\n")

        print("관련된 적하항은")
        print(Landing_Ports.loc[i])

        print("\n관련된 양하항은" )
        print(discharging.loc[i])
        correct = 1


if correct == 0:
    print("잘못된 입력입니다.\n")

    edit_list =[]
    for i in range(len(data)):
        edit_list.append(editdistance.eval(a,data["Name"][i] ))

    print("올바른 입력은 ")
    print(data["Name"][edit_list.index(min(edit_list))])
```
```
올바른 입력입니다.
METHANOL 을(를) 입력하셨습니다.

관련된 적하항은
Unnamed: 0        86
0               IRABD
2               IRBKM
4               SAJUB
6               KRUSN
8               OMSOH
10              IRBAH
Name: 86, dtype: object

관련된 양하항은
Unnamed: 0        86
0               IRABD
2               IRBKM
4               SAJUB
6               KRUSN
8               OMSOH
Name: 86, dtype: object
```

In [125]:
```
b = input()
```

```
METANAL
```

In [126]:
```
print(b)
```

```
METANAL
```

In [128]:
```
correct = 0

for i in range (len(data)):
    if b == data["Name"][i]:
        print("올바른 입력입니다.")
        print(b + " 을(를) 입력하셨습니다.\n")

        print("관련된 적하항은")
        print(Landing_Ports.loc[i])

        print("\n관련된 양하항은" )
        print(discharging.loc[i])
        correct = 1


if correct == 0:
    print("잘못된 입력입니다.\n")

    edit_list =[]
    for i in range(len(data)):
        edit_list.append(editdistance.eval(a,data["Name"][i] ))

    print("올바른 입력은 ")
    print(data["Name"][edit_list.index(min(edit_list))])
```

```
잘못된 입력입니다.

올바른 입력은
METHANOL
```

데이터 활용방안

- 개선 방안 및 활용

1. 품목명 정형화를 통해, 품목코드에 대한 세부 정보를 쉽게 알 수 있음.
2. 데이터의 오류를 줄여 항만공사/선원 등 다시 소통하는 시간, 비용 감소
3. 오타를 입력하더라도, 품목명 규격 등록 처리시 정형화 된 룰을 쉽게 적용할 수 있음.
4. 각 군집에 대한 최빈 품목코드를 알 수 있음
5. 물품 명으로 코드 값을 쉽게 찾아낼 수 있음

데이터 활용방안

-장점 및 활용 대상

1. 어느 품목이 어떤 항구에 있는지 정보가 궁금할 경우 사용
2. 데이터 입력 오류 시 warning message를 전달 받고 수정 가능
3. 통계적 오류 줄이기 : 문자가 모두 수치화 되어 있다면, 모든 군집에 대한 유사도를 비교 후 같은 군집의 품명이지만 다른 코드의 경우, 그 군집의 최빈 품목 코드값으로 자동 변경 가능
4. 문서 작업의 시간 효율성 극대화

감사합니다

항만물류 데이터 정형화를 통한
비즈니스 절차 개선