Virtual Event 15-18 June 2020

2020 Asia-Pacific Statistics Week

Leaving no one and nowhere behind

Application of the text mining technique to improve the dataset integration in foreign trade price indexes

Session Methodological Approach to Integrated Analysis: Use of Sound Methodologies

Presenter:

Reza Hadizadeh

Leader group of PPI, Statistical Center of Iran

#apstatsweek2020

DECADE OF ACTION

UNITED NATIONS ESCAP

Economic and Social Commission for Asia and the Pacific

UKaid from the British people

Introduction

- Price Indices has been considered as one of the oldest indices for monitoring economic changes and fluctuations.
- These values imply the price variations on merchandise and services in a predetermined period. Normally, there are four main indices in economics.
- Such as Customer Price Index(CPI), Producer Price Index(PPI), Export Price Index(XPI) and Import Price Index(MPI).
- XPI and MPI are used to different purposes:

○ XPI depict the price trend of exported merchandise outside the country borders.

○ MPI focuses on a similar trend for imported merchandise from in a specific period.

Unit Value index Method

- This method that most countries have been using it, is the most common way of calculation. This method uses the customs registers for both imported and exported merchandise based of values of trades and other supplementary data in customs organization.

- This index is the ratio of the value of a unit in considered period on the reference period.

$$p_u = \left(\frac{\sum_{m=1}^{M} p_m^t q_m^t}{\sum_{m=1}^{M} q_m^t} \right) / \left(\frac{\sum_{n=1}^{N} p_n^0 q_n^0}{\sum_{n=1}^{N} q_n^0} \right)$$

DATA SET

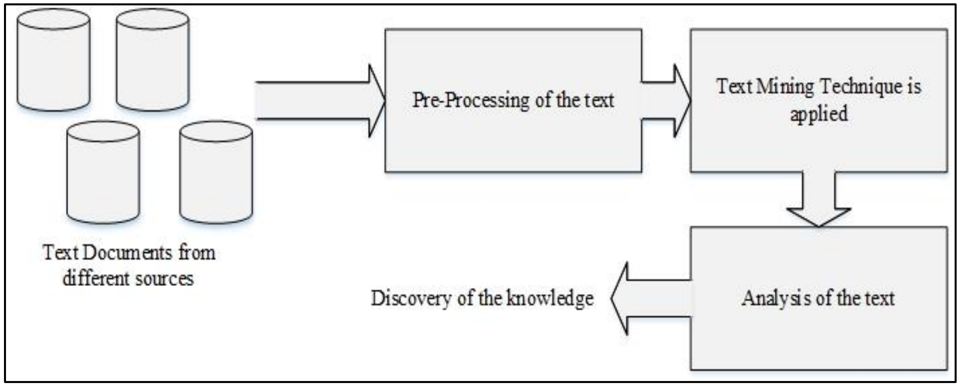
- In Iran Customs organization register there is a series of information based on the HS coding system that is called on attribute set.
- This attribute set include tariff code, value in US dollar and Iran's currency (Rial), weight, country of origin, transportation type, data of arrival and average of exchange rate.
- there is a big challenge that is related to the HS coding system that has been group classified. this is in contrast with the definition of price index calculation.
- For example, HS 71110000 (Base metals, silver or gold, clad with platinum, ….) includes gold and silver with all side products.
- If the price index displays a growth in comparison to last period. it is not exactly possible to emphasize that which one resulted to this increase, gold or silver?

To solve and overcome this problem, text-mining technique is offered in this study.

Text Mining

- Text mining is the process of seeking or extracting the useful information from the textual data. It is an exciting research area as it tries to discover knowledge from unstructured texts. It is also known as Text Data Mining (TDM) and knowledge Discovery in Textual Databases (KDT).

- Text mining process is same as data mining, except, the data mining tools are designed to handle structured data whereas text mining can able to handle unstructured or semi-structured data sets such as emails HTML files and full text documents etc.



Text Documents from different sources

Pre-Processing of the text

Text Mining Technique is applied

Analysis of the text

Discovery of the knowledge

- The Harmonized System is an international nomenclature for the classification of products. It allows participating countries to classify traded merchandise on a common basis for customs purposes. At the international level, the Harmonized System (HS) for classifying merchandise is a six-digit code system.

- The HS comprises approximately 5,300 article/product descriptions that appear as headings and subheadings, arranged in 99 chapters, grouped in 21 sections.

- **The six digits can be broken down into three parts:**

- The first two digits (HS-2) identify the chapter the merchandise are classified in, e.g. 09 = Coffee, Tea, Maté and Spices.
- The next two digits (HS-4) identify groupings within that chapter, e.g. 09.02 = Tea, whether or not flavored.
- The next two digits (HS-6) are even more specific, e.g. 09.02.10 Green tea (not fermented)... Up to the HS-6 digit level, all countries classify products in the same.

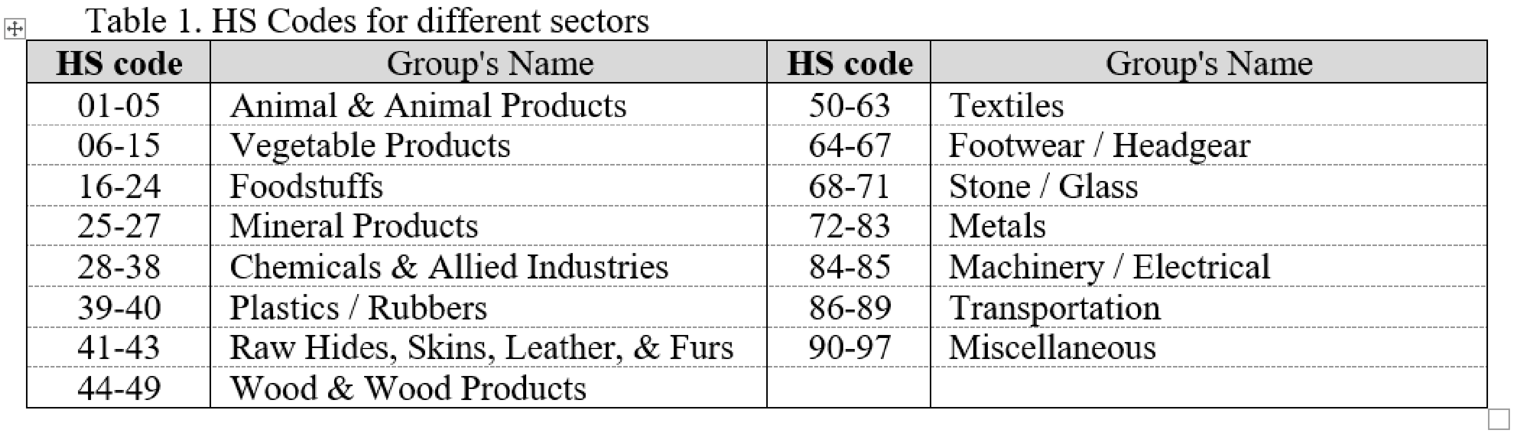
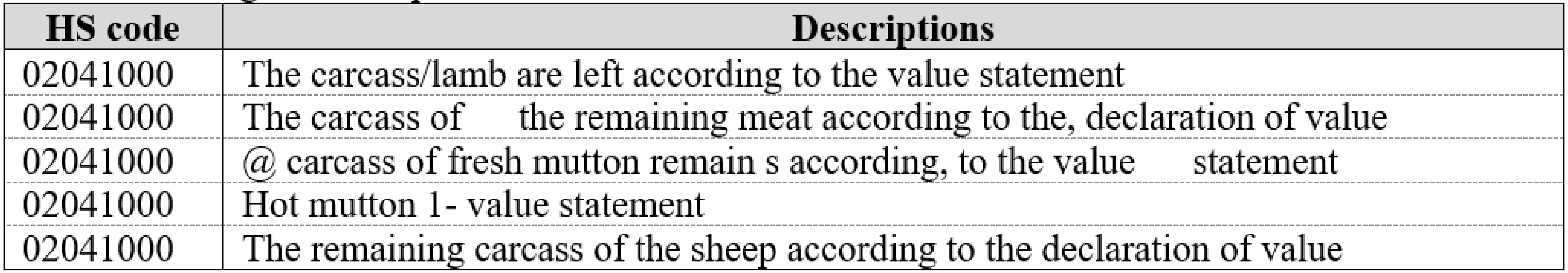


Table 1. HS Codes for different sectors

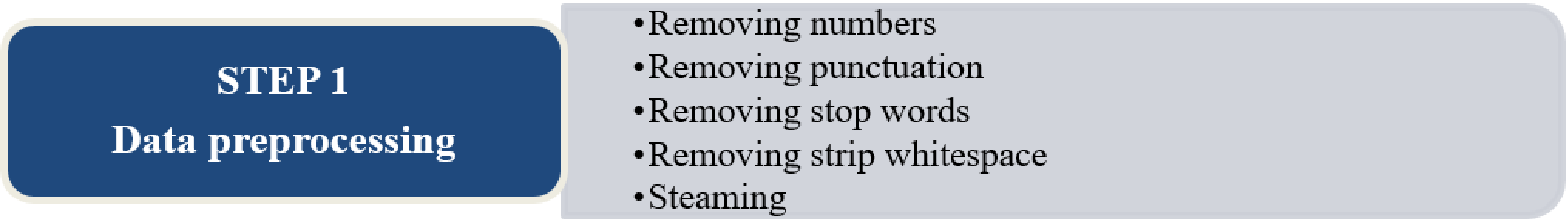
HS code	Group's Name	HS code	Group's Name
01-05	Animal & Animal Products	50-63	Textiles
06-15	Vegetable Products	64-67	Footwear / Headgear
16-24	Foodstuffs	68-71	Stone / Glass
25-27	Mineral Products	72-83	Metals
28-38	Chemicals & Allied Industries	84-85	Machinery / Electrical
39-40	Plastics / Rubbers	86-89	Transportation
41-43	Raw Hides, Skins, Leather, & Furs	90-97	Miscellaneous
44-49	Wood & Wood Products		

Table 2. Sub categories for specific HS code



HS code	Descriptions
02041000	The carcass/lamb are left according to the value statement
02041000	The carcass of the remaining meat according to the, declaration of value
02041000	@ carcass of fresh mutton remain s according, to the value statement
02041000	Hot mutton 1- value statement
02041000	The remaining carcass of the sheep according to the declaration of value

❖ Taking register data as a data source, there are some remarkable challenges of different type of merchandise as well as redundant characters resulting in low efficient linkage. So, if the linkage is apply based only 8-digit, the calculated indices will mislead and be biased. In this paper we proposed text mining technique for solving this problem.



STEP 1

Data preprocessing

- Removing numbers
- Removing punctuation
- Removing stop words
- Removing strip whitespace
- Steaming

Commonly each code’s description contains a series of characters includes but not limited to numbers, symbols, low importance signs and redundant spaces. Thus, in order to prepare high quality analysis on theses codes’ descriptions it would be necessary to remove these characters (punctuation, numbers, stop words and whitespaces). Last but not least step is Steaming that is done in different ways that are demanding and interested readers can find details in many sources. Steaming is one of the prominent steps in text mining techniques.

STEP 2
Documents similarity

- Jaccard similarity
- Cosine similarity

❖ Many methods have been introduced for similarity findings in two different texts. One of the famous methods is the metric distance that exists between two texts. Technically, the R programming uses Jaccard similarity and Cosine similarity methods to find similarities between two texts .

❖ It is worth mentioning that unqualified data matching based on merchandise' descriptions is one of the immense challenges for precise calculation pf imported and exported price indices. Technically, linkage with the only merchandise' descriptions (tariff code, the value in US dollar and Iran's currency (Rial), weight, country of origin, transportation type, data of arrival and average of the exchange rate) can result to miss some important data while in price indices calculation the important parts are the specific merchandise' attributes and stability of considered merchandise during the specific period.

❖ In connection with relative price formula, after preprocessing phase, similarity verdict of merchandise' descriptions for current and last month is necessary for the superior quality of prices. In the other words, each merchandise's description of 8-digit tariff code in data set (1), current month, should be linked with similar description in dataset (2), last month. The relative price criteria will be the similarity degree with at least 70% similarity threshold.

STEP 3 Library making

- Labeling
- Identify Code

❖ Following the similarity determination, a label was appointed to each of tariff codes based on the keyboard's in description and the number of repetitions in the database. These labels however can be a 4-digit number which can be attached to previous 8-digit tariff codes. The new codes had 12-digit that would be more beneficial for the next linkage.

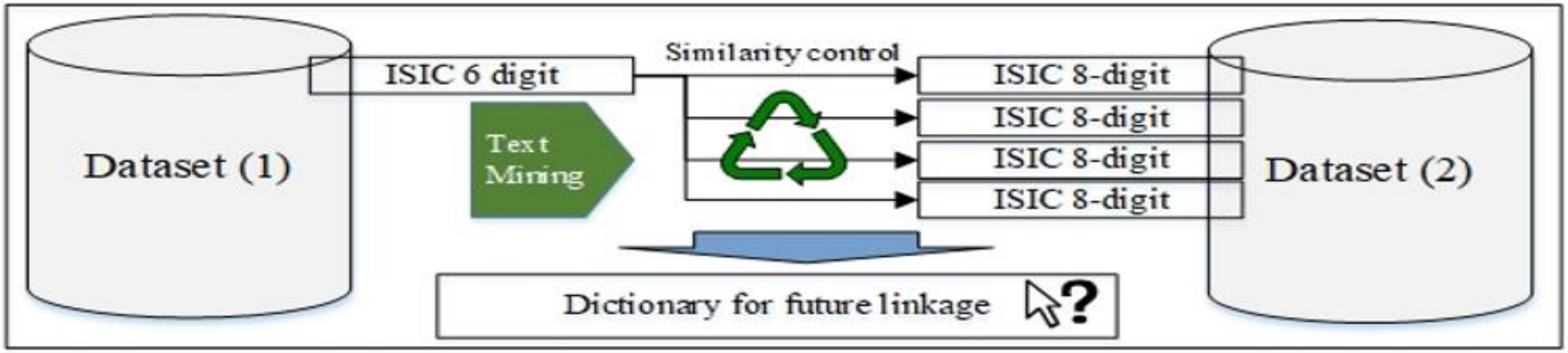
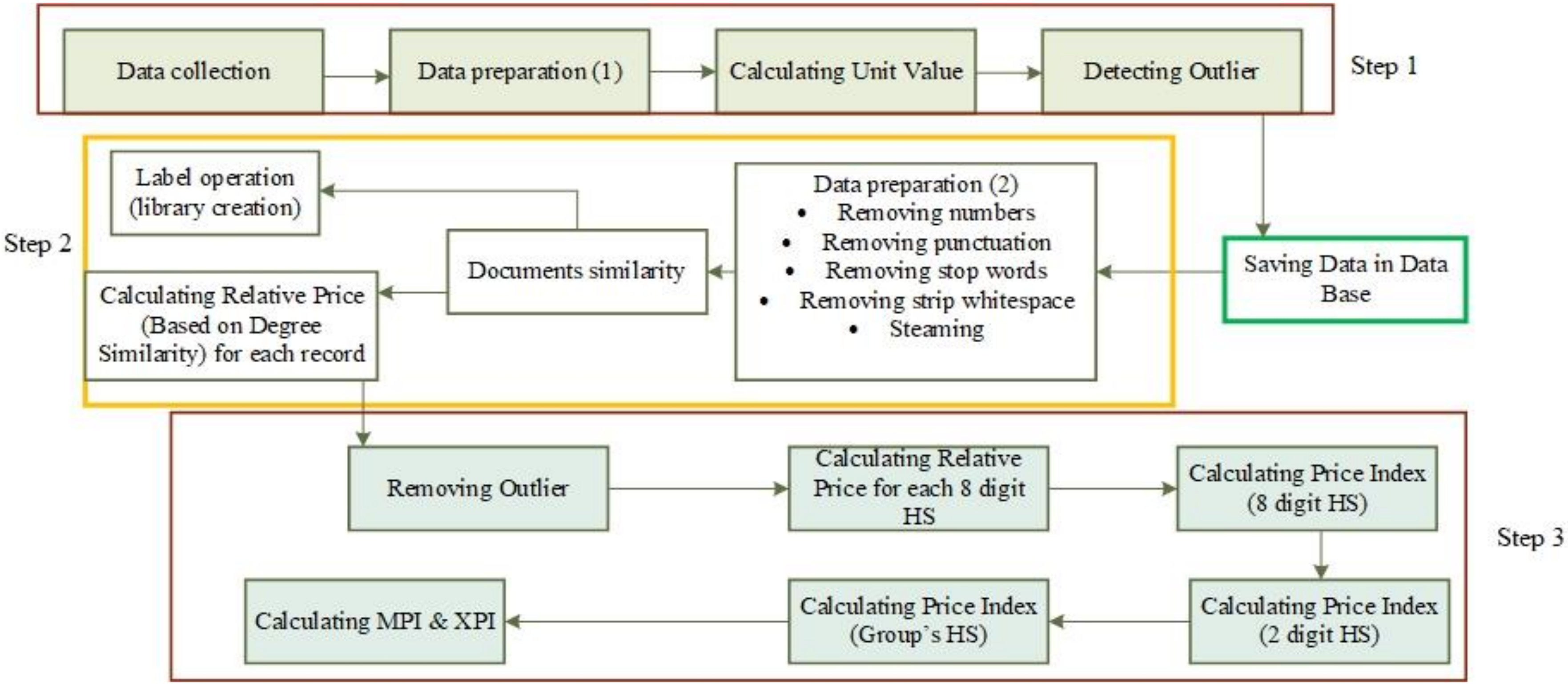


Fig. 2: Linkage process with text mining in price indices

In this new process, step 2 added to previous process.


Data collection
Data preparation (1)
Calculating Unit Value
Detecting Outlier
Step 1
Saving Data in Data Base
Data preparation (2)
• Removing numbers
• Removing punctuation
• Removing stop words
• Removing strip whitespace
• Steaming
Documents similarity
Label operation (library creation)
Calculating Relative Price (Based on Degree Similarity) for each record
Step 2
Removing Outlier
Calculating Relative Price for each 8 digit HS
Calculating Price Index (8 digit HS)
Step 3
Calculating Price Index (2 digit HS)
Calculating Price Index (Group's HS)
Calculating MPI & XPI

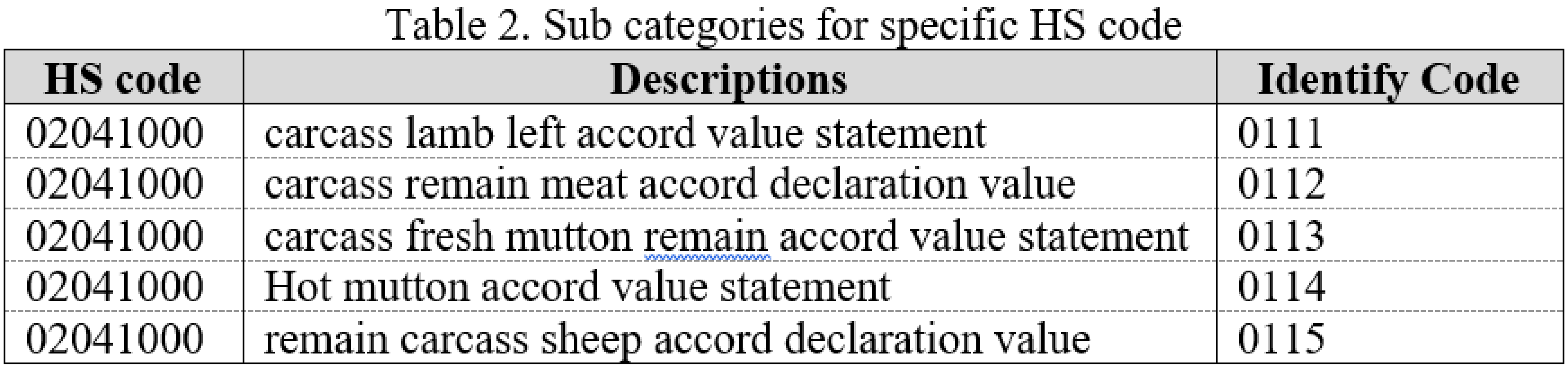
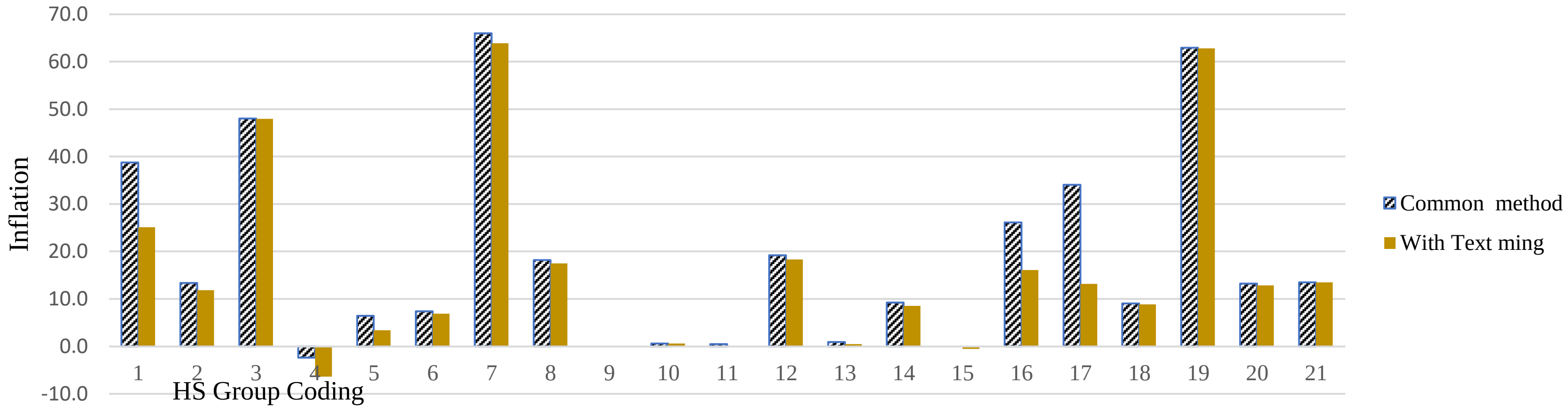


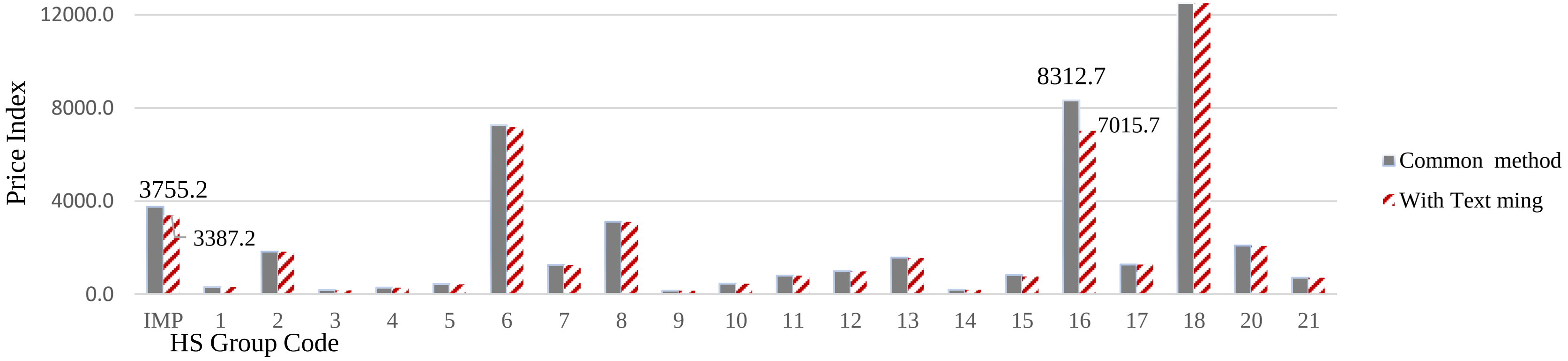
Table 2. Sub categories for specific HS code

HS code	Descriptions	Identify Code
02041000	carcass lamb left accord value statement	0111
02041000	carcass remain meat accord declaration value	0112
02041000	carcass fresh mutton remain accord value statement	0113
02041000	Hot mutton accord value statement	0114
02041000	remain carcass sheep accord declaration value	0115

A practical sample of applying text mining technique in phase 2 of indices calculation process was shown in table 2.

Computational improvement is evident in the following two charts, the index and inflation.





Discussion

Average price report has been under attention of statistical users as merchandise price indices. This managerial report has been prepared based on 8-digit HS codes. In many cases that the price fluctuation for each tariff code of merchandise group is not in a harmony, the average price in not precise and may be biased. In these cases, however text mining can be considered as a last resort for solving the problem. Text mining can cluster similar merchandise with a high level of similarities resulted in high quality average prices for each merchandise group.

THE END