APES WEEK 2019
ASIA-PACIFIC ECONOMIC STATISTICS WEEK
Integrating economic statistics in monitoring the 2030 Agenda

Trade by Enterprise Characteristics: Managing Structured Data in a Big Data Environment

RPMES REGIONAL PROGRAMME IN ASIA-PACIFIC ECONOMIC STATISTICS

UNITED NATIONS ESCAP
Economic and Social Commission for Asia and the Pacific

UKaid from the British people

Context

The aim:

"To match import and export unit data with the Malaysia Statistical Business Register to produce estimates of imports and exports by business"

The problem:

Due to poor quality of unique ID of businesses in customs records, the matching process need to be used exporters and importers companies' name. There are 16 million trade records need to be matched to possible 3.6 million businesses on the MSBR using names.

The Solution

Ideal scenario.....

Trade Database MSBR

Business Registration Number (BRN)
- Unique ID

MATCH

Trade Database MSBR

TEXT

MATCH

Traders' name  Businesses' name

Trade Database

Text 1
- Data cleansing
- Data standardization

MSBR

Text 2
- Data cleansing
- Data standardization

Text Matching

Data cleansing:

Removed any invalid or unwanted characters

NO	INVALID CHARACTER
1	&#39
2	&
3	'''
4	^(C *\/ *O) *
5	&quot
6	á
7	*:;?
8	+C\s*\/\s*O .*
9	\+
10	+
11	\[[^\[]*\]
12	,
13	&amp
14	[-]$
15	#

Data Standardization:

Postfix & Generic

NO	ORIGINAL WORD	STANDARDIZE WORD	P/G
1	M SIA	MALAYSIA	P
2	M&#39SIAN	MALAYSIA	P
3	MSIA	MALAYSIA	P
4	M'SIA	MALAYSIA	P
5	MSIAN	MALAYSIA	P
6	PACFC	PACIFIC	P
7	PACICIC	PACIFIC	P
8	PASIFIC	PACIFIC	P
9	ENG	ENGINEERING	G
10	ENGG	ENGINEERING	G
11	ENGINEERIING	ENGINEERING	G
12	ENGINEERINGS	ENGINEERING	G
13	INEERING	ENGINEERING	G
14	RESOU	RESOURCE	G
15	RESOUCE	RESOURCE	G
16	RESOURCES	RESOURCE	G

Data Matching:

Levenshtein Distance Algorithm

Levenshtein distance is a string metric of calculating the difference of two words. The distance is said as the minimum number of single-character edits (i.e. insertions, deletions, or substitutions) required to change one word into another.

Results

100% = exactly match

>88% and < 100 % = partly match

< 88% = not match

No. of trade and MSBR records

TYPE OF RECORDS	2014	2015	2016	2017	2018
TRADE RECORDS	13.2mil	14.0mil	14.6mil	15.5mil	16.2mil
MSBR RECORDS	3.1mil	3.2mil	3.3mil	3.4mil	3.6mil

The matching rate

EXPORTS: IN TERMS OF NUMBER OF RECORDS					
	2014	2015	2016	2017	2018
EXACT MATCH (100%)	64.4%	63.0%	68.1%	71.7%	72.0%
PARTLY MATCH (88% - 99%)	5.2%	5.0%	5.6%	5.2%	5.2%
NOT MATCH (<88%)	30.1%	25.9%	23.5%	22.9%	22.4%
HARDCOPY FORM	0.4%	6.1%	2.7%	0.2%	0.4%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%

EXPORTS: IN TERMS OF TRADE VALUE (IN RM)					
	2014	2015	2016	2017	2018
EXACT MATCH (100%)	59.9%	59.3%	64.9%	70.0%	67.4%
PARTLY MATCH (88% - 99%)	6.1%	6.0%	7.4%	6.5%	7.2%
NOT MATCH (<88%)	30.4%	29.8%	25.6%	22.7%	24.9%
HARDCOPY FORM	3.6%	4.9%	2.2%	0.7%	0.5%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%

Quality Assurance & Quality Check (QAQC):

Manual checking process where human perspective is needed since automated scripts may not pick up the visual issues.

Sample of standardize companies' name using QAQC system

NO	ORIGINAL COMPANIES' NAME	STANDARDIZE COMPANIES' NAME
1	ADEL ELECTRONIC COMPONENT MNFG SDN	ADEL ELECTRONICS COMPONENT MANUFACTURING SDN BHD
2	AGILENT TECHNOLOGY LDA (MALAYSIA) SDN BHD TAX AGENT	AGILENT TECHNOLOGIES LDA MALAYSIA SDN BHD
3	AIRBUS HELICOPTRS MALAYSIA	AIRBUS HELICOPTERS MALAYSIA SDN BHD

Matching rate before and after QAQC

EXPORTS: IN TERMS OF NUMBER OF RECORDS					
	2014	2015	2016	2017	2018
BEFORE QAQC	69.6%	68.0%	73.7%	76.9%	77.2%
AFTER QAQC	-	-	84.6%	-	-

EXPORTS: IN TERMS OF TRADE VALUE (IN RM)					
	2014	2015	2016	2017	2018
BEFORE QAQC	66.0%	65.3%	72.3%	76.5%	74.6%
AFTER QAQC	-	-	83.0%	-	-

Limitations

1. Only matched **70%** of the records
2. Need human interference to increase the matching rate another **10%**
3. Need to **update dictionary** when a new set of data come in.
4. Not suitable for all situations
5. Inferior in all measures to matching with Unique ID.

Group Discussion on Fuzzy Matching

- What examples have you seen?
- What are the limitations?
- Where is it suitable?
- Where would it not be suitable?

- Questions for the author?