IAEA-NDS-0240

INTERNATIONAL ATOMIC ENERGY AGENCY

NUCLEAR DATA SERVICES

DOCUMENTATION SERIES OF THE IAEA NUCLEAR DATA SECTION

INST-FINDER
(Utility for EXFOR INSTITUTE coding)

Naohiko Otuka
IAEA Nuclear Data Section, Vienna, Austria

April 2023

IAEA Nuclear Data Section, Vienna International Centre, A-1400 Vienna, Austria

Note:

The IAEA-NDS-reports should not be considered as formal publications. When a nuclear data library is sent out by the IAEA Nuclear Data Section, it will be accompanied by an IAEA-NDS-report which should give the data user all necessary documentation on contents, format and origin of the data library.

IAEA-NDS-reports are updated whenever there is additional information of relevance to the users of the data library.

For citations care should be taken that credit is given to the author of the data library and/or to the data centre which issued the data library. The editor of the IAEA-NDS-report is usually not the author of the data library.

Neither the originator of the data libraries nor the IAEA assume any liability for their correctness or for any damages resulting from their use.

96/11

Citation guideline:

When quoting INST-FINDER in a publication this should be done in the following way:

N. Otuka, “INST-FINDER: Utility for EXFOR INSTITUTE coding”, report IAEA-NDS-0240, International Atomic Energy Agency, 2023.

IAEA-NDS-0240

INST-FINDER
(Utility for EXFOR INSTITUTE coding)

Naohiko Otuka
IAEA Nuclear Data Section, Vienna, Austria

Abstract

INST-FINDER is a utility code (Python script) to support EXFOR compilers by constructing a code string for the keyword INSTITUTE. A text string copied from an electronic content such as web page is compared with the expansions of the institute codes defined in EXFOR/CINDA Dictionary 3 (Institutes). Their similarity is evaluated according to Gestalt pattern matching, and the institute codes having high scores are shown as the candidates of the institute code to be coded.

April 2023

Introduction

It is time consuming to choose the right institute code from Dictionary 3 when the authors of the source article are from many institutes. INST-FINDER is a code (Python script) proposing an institute code to be coded. Once the institute list is copied from an electronic content (e.g., article website), the script shows candidates of the institute code to be coded under the keyword INSTITUTE.

The script compares the text string on website and dictionary and calculate a similarity score according to the "Gestalt pattern matching", which calculates a similarity score by

$$2\sum_{i=1}^{n} M_i \Big/ (T_1 + T_2)$$

where M_i is the length of the matching characters and T is the total length of the two strings.

Example

The similarity score between "*iThemba Laboratory for Accelerator Based Sciences, Somerset West 7129*" copied form the article (T_1=69) and "*iThemba LABs*" in Dictionary 3 (T_2=27):

```
----+----1----+----2----+----3----+----4----+----5----+----6----+----7
iThemba Laboratory for Accelerator Based Sciences, Somerset West 7129
iThemba LABS, Somerset West
```

Characters in both strings	Length of match (M_i)
`iThemba L`	9
`A`	1
`B`	1
`S`	1
`, Somerset West`	15
Sum of M_i	27

The total lengths of the two strings are 69 and 27. Therefore, the score is 2×27/(69+27)=0.5625.

The code detects the country of the copied text string and compare the text string with the expansions of the institute codes belonging to the same country. After evaluation of the similarity scores, the institute codes are sorted so that the institute code having the highest similarity score as the top candidate and proposed to the compiler.

You can download the Python script from the NRDC Software webpage (https://nds.iaea.org/nrdc/nrdc_sft/).

Files

Place the following six files in the same directory.

1. **inst-finder.py**: The Python script file

2. **dict_alia.txt**: Dictionary defining aliases other than country names (e.g., "INFN" is defined as an alias of the "Istituto Nazionale di Fisica Nucleare, Sezione di". It has fixed length format. If the alias is the full stop symbol, the words are removed without any replacement.

   Example

```
----+----1----+----2----+----3----+----4----+----5----+----6----+-
Istituto Nazionale di Fisica Nucleare, Sezione di      INFN
Cracow                                                 Krakow
```

3. **dict_cnab.txt**: Dictionary defining aliases of country names (e.g., "USA" is defined as an alias of the "United States of America" which is the expansion of 1USAUSA.). It has fixed length format.

   Example

```
----+----1----+----2----+----3----+----4----+----5----+----6----+-
South Korea                    Republic of Korea
The Netherlands                Netherlands
UK                             United Kingdom
USA                            United States of America
```

4. **dict_pref.txt**: Dictionary defining a part of words *separated by a comma* to be removed (e.g., "*Faculty of Science,* University of Zagreb" becomes "University of Zagreb".)

   Example

```
----+----1----+----2----+----3----+----4----+----5----+----6----+-
Dipart
Depart
Facul
```

5. **dict_zzzc.txt**: Dictionary defining countries of international organization locations (e.g., "2AUS" is defined as the location of "3ZZZIAE"). It has fixed length format.

   Example

```
----+----1----+----2----+----3----+----4----+----5----+----6----+-
2ZZZGEL 2BLG EC Joint Research Centre (EC-JRC), Geel
3ZZZIAE 2AUS IAEA, Vienna
4ZZZDUB 4RUS Joint Institute for Nuclear Research (JINR), Dubna
```

6. **EXFOR/CINDA Trans dictionary** (Distributed from the NDS open area, e.g., TRANS.9127.).

Installation

1. Install Python3 (e.g., https://www.python.org/downloads/) .
2. Install a Python modules “pyperclip”.

   (e.g., “`> pip install pyperclip`”)
3. Place the four files and a trans dictionary on the same folder.

Execution

1. Run the script with the trans dictionary name as the argument.

   (e.g., “`> python3 inst-finder.py trans.9127`”).
2. Type the entry number (e.g., “`D1010`”).
3. Copy (without paste) the text strings including institute names. (This process can be repeated. The copied lines can include a line other than an institute name such as an email address.)
4. Click “Ctrl + C” to terminate copy.
5. Choose a proposed institute by tying its ID. Hit the enter key to see more candidates. Type “0” to choose the country code for an institute undefined in Dict. 3.
6. When the program cannot find a country code in the text line, the program asks to type the country name (e.g., “Austria”). Just hit the enter key if the line is not for an institute name (e.g., email address).
7. When all text lines are analyzed, the program asks to confirm your selections. If you agree with your selections, type “y” then the institute code string is printed and saved in “inst-find.log”.

Example

Getting the institute codes for A. Budzanowski+,J,NP/A,161,610,1971 (EXFOR D1010) by copying the following text lines from the article website (https://doi.org/10.1016/0375-9474(71)90389-7):


```
[otsukan@NB635819] Python3 inst-finder.py trans.9127
-----------------------------------------------------------------
             INST-FINDER: EXFOR Institute Code Finder
-----------------------------------------------------------------
Type entry number.
D1010
Copy institute text from web page (Ctrl+c when complete):

a
Institute of Nuclear Physics, Cracow 23, Poland
b
Institute of Physics Jagellonian University, CracowPoland
c
Institute of Physics Silesian University, Katowice, Poland
Received 7 July 1970, Revised 7 November 1970, Available online 21 October
2002.
^C
Input completed


++ Institute of Nuclear Physics, Cracow 23, Poland

   [  1] 3POLSLS(0.462) Univ. of Silesia, Katowice

   [  2] 3POLWRO(0.413) Univ.of Wroclaw, Wroclaw
   [  3] 3POLKPS(0.373) Wyzsza Szkola Pedagogiczna, Katowice
   [  4] 3POLIBJ(0.338) Inst. Badan Jadr., Swierk and Warszawa
   [  5] 3POLKPI(0.329) Wyzsza Szkola Pedagogiczna, Kielce
   [  0] 3POLPOL(0.044) Poland

   Hit return to see more candidates. Type 0 to choose the country code ->
(Hit entre key)
++ Institute of Nuclear Physics, Cracow 23, Poland

   [  6] 3POLIFJ(0.322) Niewodniczanski Instytut Fizyki Jadrowej, Krakow
   [  7] 3POLUJK(0.314) Jagiellonian University, Krakow
   [  8] 3POLLOU(0.310) Univ. of Lodz, Lodz
   [  9] 3POLPWA(0.290) Politechnika Warszawska
   [ 10] 3POLITJ(0.278) AGH University of Science and Technology
   [  0] 3POLPOL(0.044) Poland

   Hit return to see more candidates. Type 0 to choose the country code -> 6

   Your choice: 3POLIFJ=Niewodniczanski Instytut Fizyki Jadrowej, Krakow
```

++ Institute of Physics Jagellonian University, CracowPoland

If this is an institute, type its country name, otherwise hit the entre key.
Poland
[1] 3POLUJK(0.595) Jagiellonian University, Krakow

[2] 3POLWWA(0.444) Warszawa, University
[3] 3POLSKU(0.410) Curie-Sklodowska University, Lublin
[4] 3POLIBJ(0.346) Inst. Badan Jadr., Swierk and Warszawa
[5] 3POLIFJ(0.330) Niewodniczanski Instytut Fizyki Jadrowej, Krakow
[0] 3POLPOL(0.163) Poland

Hit return to see more candidates. Type 0 to choose the country code -> **1**

Your choice: 3POLUJK=Jagiellonian University, Krakow

++ Institute of Physics Silesian University, Katowice, Poland

[1] 3POLSLS(0.605) Univ. of Silesia, Katowice

[2] 3POLUJK(0.519) Jagiellonian University, Krakow
[3] 3POLKPS(0.419) Wyzsza Szkola Pedagogiczna, Katowice
[4] 3POLSKU(0.376) Curie-Sklodowska University, Lublin
[5] 3POLWWA(0.371) Warszawa, University
[0] 3POLPOL(0.143) Poland

Hit return to see more candidates. Type 0 to choose the country code -> **1**

Your choice: 3POLSLS=Univ. of Silesia, Katowice

++ Received 7 July 1970, Revised 7 November 1970, Available online 21 October 2002.

If this is an institute, type its country name, otherwise hit the entre key.
(Hit entre key)

----------Summary of your selections----------

3POLIFJ Institute of Nuclear Physics, Cracow 23, Poland

3POLUJK Institute of Physics Jagellonian University, CracowPoland

3POLSLS Institute of Physics Silesian University, Katowice, Poland

---

Are these correct? (y/n) **y**

INSTITUTE (3POLIFJ,3POLUJK,3POLSLS)

Good bye!

Nuclear Data Section
International Atomic Energy Agency
P.O. Box 100
A-1400 Vienna
Austria

e-mail: nds.contact-point@iaea.org
fax: (43-1)26007
telephone: (43-1)2600-21710
web: http://www-nds.iaea.org/