INST-FINDER and its extension with Naive Bayes classifier

(N. Otsuka, 2023-04-18, Memo CP-D/1079)

Release of INST-FINDER

You can download a new utility (Python script) to find institute codes (INST-INFDER) from the NRDC software webpage (https://nds.iaea.org/nrdc/nrdc_sft/) and NDS GitHub (https://github.com/iaea-nds/inst-finder/). See the report **IAEA-NDS-0240** for the usage of this utility.

Coding for keyword INSTITUTE does not require scientific knowledge but it is sometimes time consuming. The codes are not checked by NDS and can be a major place of erroneous coding. An attempt to automate its coding was presented by V. Devi and J. Singh in the 2022 EXFOR Workshop. Motivated by their work, I also wrote a Python script INST-FINDER proposing an institute code to be coded. Once the institute list is copied from an article website, the script shows candidates of the institute code to be coded.

Example:

```
++ iThemba Laboratory for Accelerator Based Sciences, Somerset West
7129, South Africa

   [  1] 3SAFITH(0.562) iThemba LABS, Somerset West
   [  2] 3SAFNAC(0.427) National Accelerator Centre, Faure
   [  3] 3SAFSIR(0.314) Council for Scientific and Industrial Res.,
Pretoria
   [  4] 3SAFUPR(0.267) Univ.of Pretoria, Hatfield, Pretoria
   [  5] 3SAFNLP(0.236) National Physical Research Lab., Pretoria
   [  0] 3SAFSAF(0.184) South Africa, Rep.

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

   Your choice: 3SAFITH=iThemba LABS, Somerset West
```

The script compares the string on the article website and dictionary, and calculates a similarity score according to the "Gestalt pattern matching", which calculates the score by $2 \times M/T$, where M is the total length of the matching characters and T is the total length of the two strings.

Example (iThemba LABs on an article website and Dictionary 3):

```
----+----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 (*Mi*)
iThemba L	9
A	1
B	1
S	1
, Somerset West	15
Sum	**27**

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

Prediction of institute code by Naive Bayes classification

I also applied "Naive Bayes classifier" as an alternative of the Gestalt pattern matching. I vectorized each institute name (text string in Dict. 3 and article websites) to **x** which has an element for each word in the string. Then I calculated the probability of having an institute code C in Dictionary 3 (e.g., `iThemba LABS, Somerset West`) given the institute name on the article website (e.g., `iThemba Laboratory for Accelerator Based Sciences, Somerset West 7129`) by maximizing the likelihood for the training cases with the conditional probability P(**x**|C) estimated by learning. Note that Bayes theorem is applied to find C maximizing P(C|**x**).

To test this procedure, I predicted the institute codes for the 495 institute names on the articles compiled in the 1st quarter of 2022 by the following four methods (parenthesized numbers give the correct answer ratios):

1. Gestalt pattern matching (66.5%)
2. Bayes classification after training with Dictionary 3 (74.8%)
3. Same as 2 but with additional training with 542 institute names on the articles compiled in 4th quarter of 2021 (84.4%)
4. Same as 3 but with additional training with 472 institute names on the articles compiled in 3rd quarters of 2021 (87.1%)

Conclusion: The current INST-FINDER could be improved by addition of a learning process.

Possible improvement of institute code expansions

Appendix of this memo summarizes the combinations of the predicted (but incorrect) code and correct code appearing twice or more in the output from the 1st method. The correct answer ratio could be improved by revising the expansions of these codes (e.g., "Institute of Nuclear Physics" instead of "Instytut Fizyki Jadrowej" for 3POLIFJ).

Appendix: Selected combinations of predicted and correct codes

Code (predicted)	Code (correct)	Institute name on the article website
1CANUBC	1CANTMF	TRIUMF, Vancouver, British Columbia V6T 2A3, Canada
1CANUBC	1CANTMF	TRIUMF, Vancouver, British Columbia V6T2A3, Canada
1USAWMU	1USAUSA	Department of Physics, Central Michigan University, Mt Pleasant, MI 48859, USA
1USAWMU	1USAUSA	Department of Physics, Central Michigan University, Mt. Pleasant, MI 48859, USA
2FR BRC	2FR GAN	GANIL, CEA/DRF-CNRS/IN2P3, Boulevard Henri Becquerel, F-14076 Caen Cedex, France
2FR BRC	2FR GAN	GANIL, CEA/DSAM and CNRS/IN2P3, CAEN Cedex 05, France
2FR PAR	2FR FR	Irene Joliot Curie Lab, UMR8608, IN2P3-CNRS, Université Paris Sud 11, 91406 Orsay, France
2FR PAR	2FR FR	Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France
2FR PAR	2FR SAC	CEA Irfu, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
2FR PAR	2FR SAC	IRFU, CEA, Université Paris-Saclay, F-91191, 91190 Gif-sur-Yvette, France
2FR VNV	2FR CAD	CEA, DES, IRESNE, DER, SPRC, Physics Studies Laboratory, Cadarache, F-13108 Saint-Paul-lez-Durance, France
2FR VNV	2FR CAD	CEA, DES, IRESNE, DER, SPRC, Physics Studies Laboratory, Cadarache, F-13108 Saint-Paul-lès-Durance, France
2GERDRE	2GERTHD	Institut für Kernphysik, Technische Universität Darmstadt, 64289 Darmstadt, Germany
2GERDRE	2GERTHD	Institut für Kernphysik, Technische Universität Darmstadt, D-64289 Darmstadt, Germany
2GERDRE	2GERTHD	Technische Universität Darmstadt, Fachbereich Physik, Institut für Kernphysik, Darmstadt, Germany
2ITYLGS	2ITYITY	Gran Sasso Science Institute, Viale F. Crispi 7, 67100 L'Aquila, Italy
2ITYLGS	2ITYITY	Gran Sasso Science Institute, Viale F. Crispi 7, 67100 L'Aquila, Italy
2ITYLNS	2ITYBOL	Agenzia Nazionale per le Nuove Tecnologie (ENEA), Bologna, Italy
2ITYLNS	2ITYBOL	Agenzia nazionale per le nuove tecnologie (ENEA), Bologna, Italy
2ITYUPV	2ITYGVA	Università degli Studi di Genova, 16126 Genova, Italy
2ITYUPV	2ITYGVA	Università degli Studi di Genova, Via Dodecaneso 33, 16146 Genova, Italy
2JPNSUU	2JPNIPC	RIKEN Nishina Center, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
2JPNSUU	2JPNIPC	RIKEN Nishina Center, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
3CPRIHP	3CPRBJG	. Institute of Heavy Ion Physics, Peking University, Beijing 100871, China;
3CPRIHP	3CPRBJG	. School of Physics and State Key Laboratory of Nuclear Physics and Technology, Peking University, Beijing 100871, China
3CPRIHP	3CPRBJG	State Key Laboratory of Nuclear Physics and Technology, School of Physics, Peking University, Beijing 100871, China
3CPRSST	3CPRHST	Department of Engineering and Applied Physics, University of Science and Technology of China, Hefei 230026, China
3CPRSST	3CPRHST	Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China
3HUNWRC	3HUNHUN	Konkoly Observatory Research Centre for Astronomy and Earth Sciences, Hungarian Academy of Sciences, Budapest, Hungary
3HUNWRC	3HUNHUN	Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Hungarian Academy of Sciences, 1121 Budapest, Hungary
3HUNWRC	3HUNHUN	Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, MTA Centre for Excellence, 1121 Budapest, Hungary
3INDBDA	3INDIND	Department of Physics, Central University of Kerala, Kasaragod, 671316, India

3INDBDA	3INDIND	Department of Physics, University of Kerala, Thiruvananthapuram, 695034, India
3INDBDA	3INDIND	Department of Physics, University of Petroleum and Energy Studies, Dehradun-248 007, Uttarakhand, India
3INDIIB	3INDTAT	India-Based Neutrino Observatory, Tata Institute of Fundamental Research, Mumbai, 400005, India
3INDIIB	3INDTAT	Pelletron Linac Facility, Tata Institute of Fundamental Research, Mumbai, 400005, India
3INDJCB	3INDIND	Department of Physics, AKI's Poona College, Camp, Pune 411001, India
3INDJCB	3INDIND	Department of Physics, Shri Varshney College, Aligarh-202 001, Uttar Pradesh, India
3INDTAT	3INDIND	Homi Bhabha National Institute, Anushaktinagar, Mumbai, 400094, India
3INDTAT	3INDIND	Homi Bhabha National Institute, Anushaktinagar, Mumbai-400094, India
3IRNTEH	3IRNIRN	Department of Nuclear Physics, Faculty of Physics, University of Kashan, Kashan, Iran
3IRNTEH	3IRNIRN	Department of Physics, School of Science, Yazd University, Yazd, Iran
3IRNTEH	3IRNIRN	Department of Physics, University of Sistan and Baluchestan, Zahedan, Iran
3IRNTEH	3IRNIRN	Departments of Physics, Faculty of Science, University of Kashan, Kashan, Iran
3KORDAU	3KORKOR	Department of Physics, Sungkyunkwan University, Suwon 16419, Korea
3KORYON	3KORKOR	Department of Physics, Ewha Womans University, Seoul 03760, Korea
3MEXIPN	3MEXUMX	Instituto de Ciencias Nucleares, UNAM, Apartado 70-543, 04510 Mexico City, Mexico
3MEXIPN	3MEXUMX	Instituto de Física, Universidad Nacional Autónoma de México, Mexico City, Mexico
3POLITJ	3POLIFJ	Institute of Nuclear Physics Polish Academy of Sciences, PL-31342 Krakow, Poland
3POLITJ	3POLIFJ	Institute of Nuclear Physics Polish Academy of Sciences, PL-31342 Kraków
3POLITJ	3POLIFJ	Institute of Nuclear Physics, Polish Academy of Sciences, PL 31-342 Cracow, Poland
3POLSLS	3POLIFJ	Institute of Nuclear Physics, Cracow 23, Poland
3POLSLS	3POLIFJ	Institute of Nuclear Physics, PAS, Kraków, Poland
3POLUJK	3POLWWA	Heavy Ion Laboratory University of Warsaw, PL-20-093, Warsaw, Poland
3POLUJK	3POLWWA	Heavy Ion Laboratory, University of Warsaw, PL 02-093 Warsaw, Poland
3POLUJK	3POLWWA	Heavy Ion Laboratory, University of Warsaw, Warsaw, Poland
3RUMBUU	3RUMBUC	IFIN-HH, Bucarest, Romania
3RUMBUU	3RUMBUC	IFIN-HH, Bucharest, Romania
3RUMCIP	3RUMBUC	Extreme Light Infrastructure - Nuclear Physics, IFIN-HH, 077125 Bucharest-Măgurele, Romania
3RUMCIP	3RUMBUC	Horia Hulubei National Institute of Physics and Nuclear Engineering, Magurele, Romania
4RUSFVE	4RUSFEI	Institute for Physics and Power Engineering, Bondarenko square 1, Obninsk 249033, Russian Federation
4RUSFVE	4RUSFEI	Institute of Physics and Power Engineering (IPPE), Obninsk, Russia
4RUSMIF	4RUSKUR	National Research Center «Kurchatov Institute», 1, Akademika Kurchatova pl., Moscow, 123182, Russian Federation
4RUSMIF	4RUSKUR	National Research Center “Kurchatov Institute”, 123182, Moscow, Russia
4UZ NUU	4UZ UZ	Branch of National Research Nuclear University MEPhI, 100214, Tashkent, Uzbekistan
4UZ NUU	4UZ UZ	Gulistan State University, 120100, Gulistan, Uzbekistan
4UZ NUU	4UZ UZB	Institute of Nuclear Physics, 100214, Tashkent, Uzbekistan
4UZ NUU	4UZ UZB	Institute of Nuclear Physics, 702132, Tashkent, Uzbekistan