Automatic Classification of Software Usage Purpose Statements from Scientific Articles

author : Kobro Bekalue

December 7, 2021

Contents

1 Introduction

1.1 Recap

According to the researh conducted by ()

- Analysis of research papers can give a lot of insights about software resources and their dependency.
- In a scientific research different kinds of input resources are used. One of such input is a software.
- Used resources in a research are typically mentioned in a citation. Citation practices of formal articles in a research are matured and various citation styles exist. Even if principles for formal citation of a software has already been put out, most scientists are not properly citing resources.
- Surprisingly, sometimes researchers do not mention the type of software they used entirely or mention it with a abbreviation ... not enough creadit is being given to research software developers .
- As long as software is mentioned using formal methods, like RRID, it is possible to perform citation analysis using regular expressions which can be constructed to capture the pattern of citation.
- hough regular expression based analysis can give basic insights about the software citation it has limitations because:
  - Not so many authors use formal citation of software, like RRIDs
  - Even if scientists use formal citations, they may fail to properly follow the guidelines. For example, some authors tend to ignore the RRID-part and that creates an ambiguity by it self that it is not possible to know weather the author is actually making a software citation or it is completely something else.
  - Rule based method fails to capture context information and ignores dependencies. It is not possible to be sure about the authors intention whether or not using a software citation.
- At the same time pattern based analysis, like using regX, is not suitable to extract information about software citation, for instance the particular use of a software, especially when a software mention statement lacks any form of formality where the information is concealed in a natural language description.

- Therefore it is required to automatically extract the purpose of software use in scientific literatures. This might help to answer questions like:
  - What type of software is being frequently used for what purpose in a specific area of research? This also allows to find an answer further question like what is the most common technique researchers follow when trying to solve a given research problem in a given domain)
- Previous attempts to automatically extract information using machine learning techniques, specifically supervised machine learning technique, about the software use purpose was constrained mainly because of lack of ground truth data. But this time, with the advent of SoMeSci, it is possible to do so.

1.2 Problem Satement

1.3 Objectives of this project

This project has the following objectives:

- List down the purpose of software usage .
- To extend SoMeSci with annotation of purpose of software usage.
- To select feature for the training model
- To select a classifier and train a model.
- To evaluate and optimize results

2 The role of Software in Scientific research

2.1 Introduction

Software is a collection of instructions that supervise a computer how to execute a given task (Wikipedia contributors, 2021f). The behavior of such instructions is specified by algorithms which are derived from scientific laws (Wolfram, 1984). Implementation of algorithms is carried out using programming languages, the end result being a software (Wikipedia contributors, 2021a) (Wikipedia contributors, 2021e).

Modern research is unthinkable without a use of software and scientific investigations in various areas of science are becoming increasingly reliant on software tools (Goble, 2014) (Wilson et al., 2014) (Storer, 2017).

A software is very important asset for building a scientific knowledge and more discoveries in a research are made possible than ever by a use of software tools that automate processing of huge amount of data (Jiménez et al., 2017). Typically a software is used in a research for control processes, simulation, modelling, data analysis, knowledge dissemination, etc. (Hannay et al., 2009)(Pan, Yan, & Hua, 2016).

Since software is not often considered as an academic output (Yang, Rousseau, Wang, & Huang, 2018)(Pan et al., 2016), it is usually not cited in research papers across several fields of research (Pan et al., 2016). To counteract against this culture, a task force that advocates about the role of software in a research, known as Research software Alliance(ReSA), has been established. The ReSa promotes the inclusion of software as a primary research out put, influences decision makers to value a research software and give credits to the developers. In 2019, the task force has collected literatures, at Zetoro group libraray , that evident significant roles of software in a research [22].

Scientific software is often complicated and requires specialized domain knowledge for its development (Wilson et al., 2014). Due to this, increasing number of scientists are developing a software as part of their research work or directly taking part in the development process of a research software (Jiménez et al., 2017)(Kanewala & Bieman, 2014). This fact is clearly reinforced by a survey results from 2008, 2014, and 2017 in the UK and USA. Participants of the survey were ..... around 2000 academic staff, postdocs, Tas... (Merali, 2010)(Hettrick et al., 2014)(Nangia, Katz, et al., 2017). The results indicate that:

- Almost half of scientists spend more time developing a software as part of their research work than five years ago.

- 38% of researchers spend at least 20% of their time developing a software.
- Over 90% of scientists say software is important for their research &
- Nearly 70% claim that their research directly depends on a use of a software.

2.2 General roles

Software is playing various crucial roles in a research and making a shift in a research culture. For example, software tools are making most of research to be increasingly data driven i.e. insights from an in-depth analysis of large volume of data-sets form the basis of a research conclusion (Goble, 2014)(Jay, Haines, & Katz, 2020). Some of the most general roles of a software in a research are:

- Software helps to explore und understand a research problem (Hannay et al., 2009).
- Results from a scientific software is presented as an evidence to support a research result (Kanewala & Bieman, 2014).
- A software dictates the quality of a research outcome(Hannay et al., 2009) [23]. Outcome of a research becomes unreliable or even useless if there is an error in the software (Soergel, 2014). For example, several scientists retracted their scientific publications up on a retrospective discovery of a bug in their software (Wilson et al., 2014)(Merali, 2010)(Miller, 2006). A more palpable failure of a research ambition due to an error in the control-system software, for instance, is the failure of Ariane rocket in 1996 (Wikipedia contributors, 2021b).
- A software also helps to document a research process and to validate results of a given research (Jay et al., 2020). Executable cells in a Jupyter notebook is one real world example where a software can be used to validate a research result.
- Software allows experiments to be made beyond constrains of the physical world. This is because experiments that run on a computer are not limited by processes that occur in nature but only by the laws imbedded in the computer code (Wolfram, 1984).

2.3 Domain Specific Examples

A software is being extensively used for a research in various areas of science such as physics, chemistry, space science, life science and so on.

The physics research facility, the Large Hydron Collider at CERN, for instance uses a software with more than 5 million lines of code which is used for

processing of terabytes of data generated from experiments (Storer, 2017). In a nuclear research, a software is being developed increasingly to be used for experiments (Yan & Yatabe, 2017). For example, testing a modification in a nuclear weapon can not be field tested, but instead a software that simulate the impact of modification is usually used (Kanewala & Bieman, 2014). This is because of regulations like nuclear test ban treaties and the potential disaster, to the environment and life, associated with nuclear weapons [20].

In chemistry research, a software can be used to model and simulate chemical processes that are challenging, too complex or expensive to conduct in reality. Karplus and Levitt used computer simulations for their joint-research "the development of multi-scale models for complex chemical systems" and won a Nobel prize in 2013 for their work (Storer, 2017)(André, 2014). In a climate and environmental studies, software is used to make predictions about climate changes. For example a historical temperature data can be integrated to make predictions about future temperature variations (Storer, 2017).

In a space science, space probes heavily rely on software. In this case a software navigates space crafts to other planets, processes and transmits scientific data back to Earth fur more processing, helps researchers interpret results, etc(Lutz, 2011).

2.4 The Role in Research Breakthroughs

A use of software also allowed to produces better scientific discoveries and several research breakthroughs has been made possible(Goble, 2014).

One of the research breakthroughs is creation of the very first visual representation of a black hole using an open source software NumFOCUS. To observe a black hole that is 55 million light years away, it would have required to build a huge telescope of size of planet earth. But instead of building one giant telescope, hundreds of scientists spent decades of years creating a global network of telescopes, known as Event Horizon Telescope (EHT) (Wikipedia contributors, 2021c), synchronized precisely using atomic clocks. The EHT gathered a huge amount of data for years. However there was a lot of noise in the collected data because :

- The EHT was a network of non-similar telescopes.
- The radio signals were coming through attenuated due to atmospheric effect like water vapor, clouds, turbulence ... etc.

Therefore the scientists had to use various algorithms and data analysis pipelines. The resulting image from various data processing was compared to ensure the integrity of the result. This huge scientific breakthrough in a

space research, can be attributed to mainly the use of powerful data processing software.

Other scientific breakthroughs that can be attributed to software use in a research include:

- The detection and visualization of gravitational waves for the first time, using a LIGO software (Wikipedia contributors, 2021d)[31].
- Software accelerates drug discovery [32].

3 Literature Review on Software Usage Purpose in Research

3.0.1 Introduction

Scientists use various kinds of software, during their research, for different purposes. Some times software is used for execution of some trivial tasks like word processing and in other cases they use a software to perform critical tasks that can ultimately determine their research end result.

3.0.2 Software usage purposes in a research

In a modern research, where a research is increasingly relying on processing of huge amount of data, the most common purpose of software usage purpose is to perform data analysis.

Data analysis is a broad term which can refer to inspecting, cleaning, transforming, modelling data, etc. with a particular goal of discovering a meaningful information from the data which can be used to make conclusions or decisions [13] .

When it comes to the application of data analysis in actual research works, various kinds of data analysis techniques exit. Some of the data analysis techniques can be more general where as others are more domain specific. Some common examples of software usage purposes:

- Data Analysis, Mathematical Analysis, Statistical Analysis, Numerical Analysis, Text Analysis
- Domain specific Analysis e.g. Densitometric Analysis, Voxel-based Analysis
- Data Processing , e.g. Image processing
- Data Collection
- Modelling
- Simulation
- Programming

4 Classification of Software Usage Purposes

4.1 Citation example

some claim (Goble, 2014) that according to (Hannay et al., 2009) it was true and accoording to wiki (Wikipedia contributors, 2021f) it was ... other scholars also have pointed out that ... (Wilson et al., 2014)

References

André, J.-M. (2014). The nobel prize in chemistry 2013. *Chemistry International*, *36*(2), 2–7.

Goble, C. (2014). Better software, better research. *IEEE Internet Computing*, *18*(5), 4–8.

Hannay, J. E., MacLeod, C., Singer, J., Langtangen, H. P., Pfahl, D., & Wilson, G. (2009). How do scientists develop and use scientific software? In *2009 icse workshop on software engineering for computational science and engineering* (pp. 1–8).

Hettrick, S., Antonioletti, M., Carr, L., Chue Hong, N., Crouch, S., De Roure, D., ... others (2014). Uk research software survey 2014.

Jay, C., Haines, R., & Katz, D. S. (2020). Software must be recognised as an important output of scholarly research. *arXiv preprint arXiv:2011.07571*.

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., ... others (2017). Four simple recommendations to encourage best practices in research software. *F1000Research*, *6*.

Kanewala, U., & Bieman, J. M. (2014). Testing scientific software: A systematic literature review. *Information and software technology*, *56*(10), 1219–1232.

Lutz, R. (2011). Software engineering for space exploration. *Computer*, *44*(10), 41–46.

Merali, Z. (2010). Computational science:... error. *Nature*, *467*(7317), 775–777.

Miller, G. (2006). *A scientist's nightmare: software problem leads to five retractions.* American Association for the Advancement of Science.

Nangia, U., Katz, D. S., et al. (2017). Track 1 paper: surveying the us national postdoctoral association regarding software use and training in research. In *Workshop on sustainable software for science: Practice and experiences (wssspe 5.1).*

Pan, X., Yan, E., & Hua, W. (2016). Disciplinary differences of software use and impact in scientific literature. *Scientometrics*, *109*(3), 1593–1610.

Soergel, D. A. (2014). Rampant software errors may undermine scientific results. *F1000Research*, *3*.

Storer, T. (2017). Bridging the chasm: A survey of software engineering practice in scientific programming. *ACM Computing Surveys (CSUR)*, *50*(4), 1–32.

Wikipedia contributors. (2021a). *Algorithm — Wikipedia, the free encyclopedia.* Retrieved from `https://en.wikipedia.org/w/index`

.php?title=Algorithm&oldid=1055624679 ([Online; accessed 7-December-2021])

Wikipedia contributors. (2021b). *Ariane 5 — Wikipedia, the free encyclopedia.* Retrieved from https://en.wikipedia.org/w/index.php?title=Ariane_5&oldid=1054482061 ([Online; accessed 7-December-2021])

Wikipedia contributors. (2021c). *Event horizon telescope — Wikipedia, the free encyclopedia.* Retrieved from https://en.wikipedia.org/w/index.php?title=Event_Horizon_Telescope&oldid=1052167868 ([Online; accessed 7-December-2021])

Wikipedia contributors. (2021d). *Ligo — Wikipedia, the free encyclopedia.* Retrieved from https://en.wikipedia.org/w/index.php?title=LIGO&oldid=1047100294 ([Online; accessed 7-December-2021])

Wikipedia contributors. (2021e). *Programming language — Wikipedia, the free encyclopedia.* Retrieved from https://en.wikipedia.org/w/index.php?title=Programming_language&oldid=1055665216 ([Online; accessed 7-December-2021])

Wikipedia contributors. (2021f). *Software — Wikipedia, the free encyclopedia.* Retrieved from https://en.wikipedia.org/w/index.php?title=Software&oldid=1056292826 ([Online; accessed 7-December-2021])

Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., ... others (2014). Best practices for scientific computing. *PLoS biology*, *12*(1), e1001745.

Wolfram, S. (1984). Computer software in science and mathematics. *Scientific American*, *251*(3), 188–203.

Yan, H., & Yatabe, S. (2017). Case studies of nuclear research software development. *CNL Nuclear Review*, *8*(1), 35–51.

Yang, B., Rousseau, R., Wang, X., & Huang, S. (2018). How important is scientific software in bioinformatics research? a comparative study between international and chinese research communities. *Journal of the Association for Information Science and Technology*, *69*(9), 1122–1133.