Scaffolding Beginning Research Students Using Open Source Tools

Dr. Dhana Rao, Marshall University

Dhana Rao is an Assistant Professor in Microbiology at Marshall University, West Virginia. She obtained her PhD in 2006 from the University of New South Wales, Australia. Her research interest are in metagenomics and bioinformatics.

Dr. Rajeev K Agrawal, North Carolina A&T University (Tech)

Dr. Rajeev Agrawal is an assistant professor in the department of computer systems technology at North Carolina A&T State University. He has published 30 referred journal and conference papers and 4 book chapters. His current research focuses on Anomaly Detection in Computer Networks, Bigdata Analytics, and Content-based Image Retrieval. He has also worked at HP Company in transportation, Medicaid Management Information System (MMIS) domains.

Dr. Venkat N Gudivada, Marshall University

Venkat N Gudivada is a Professor of Computer Science in the College of Information Technology and Engineering at Marshall University. He received his Ph.D. in Computer Science from the Center for Advanced Computer Studies, University of Louisiana at Lafayette. His current research interests are in high performance computing, software visualization, and personalized eLearning.

Scaffolding Beginning Research Students Using Open Source Tools

Abstract

During the last few years, there has been an explosive growth in the number of academic research conferences, and open access as well as subscription-based online journals. Furthermore, there is an increased thrust in engaging undergraduate students in research across universities and colleges. Given that undergraduate students have limited time for research and less developed knowledge base and technical expertise in their domains, this poses special challenges. In this paper, we illustrate how several high quality open source tools can be used to overcome some of these challenges. We identify various tasks that comprise the *research workflow* pipeline and discuss solutions for a subset of the tasks.

1 Introduction

The number of open access and online scholarly journals featuring disciplinary research has increased tremendously in recent years. Added to this is an increasing number of annual research conferences which range in scope from regional to international. For example, there are literally hundreds of international conferences being held every year in computer science discipline. Keeping track of advances even in a specialized area of an academic discipline is a challenging task for professors at research-intensive universities, let alone faculty at teaching-intensive universities and beginning student researchers. Assessing current advances in the field, developing a unified understanding of a subdiscipline, determining interesting problems to work on is a tedious, labor-intensive, and intellectually daunting task especially for beginning student researchers.

In a related development, there has been increased emphasis and efforts to promote undergraduate research across universities and colleges in the country. For example, the Council on Undergraduate Research (CUR) is a national organization established to support and promote high-quality undergraduate students collaborative research with faculty.[1] The National Conferences on Undergraduate Research (NCUR) has been an annual conference since 1987. NCUR's primary goal is to promote undergraduate research in all fields of study.

National Science Foundation (NSF) funds undergraduate research through Research Experiences for Undergraduates (REU) program. Providing research experiences for undergraduate students and increasing the number of students interested in graduate programs are the goals of the REU program. Students work on REU projects during summer months for 8 to 10 weeks. REU programs entail several benefits to students including increased awareness of their discipline and technical expertise, better career opportunities, gains in confidence levels, and elevated likelihood of pursuing graduate degrees and research careers.[6] REU programs also contribute to faculty professional development.[5]

As impressive as the benefits are, REU programs pose several challenges.[27,30] Students have limited time for research, less developed knowledge base and technical expertise in the domain,

varying skill levels among students, and academic maturity that is still in the development phase. Finding interesting and challenging problems whose solutions are accessible to undergraduate students and are amenable for completion in 10 to 12 weeks is even a greater challenge. Solutions include using interdisciplinary research topics such as bioinformatics[7] and medical informatics,[34] multi-institution projects using web-based collaborative research environments,[18] industry-driven projects,[13] teams comprising both graduate and undergraduate students,[11] mechanisms for gender equity,[24] emerging technologies,[25] virtual labs,[47] and issues specific to underrepresented minorities.[29]

2 Motivation

Irrespective of the research domain, the following activities are common to all researchers: (1) locating relevant research literature, (2) critical analysis of the existing work, gaining a unified view of various facets of the domain, (3) identifying and formulating challenging research problems, (4) designing and executing a solution, (5) analyzing and interpreting results, (6) writing research papers, (7) and presenting results in a research forum. We refer to this sequence of tasks as *research workflow*. The above tasks also arise in independent study and writing-intensive courses, research methods courses,[31] and capstone projects.[26] New graduate students also experience difficulties in productive engagement with research as they lack experience with the *research workflow* tasks.

Given this backdrop, the overarching goal of this paper is to illustrate how various open source tools can be used to help with various activities in the research workflow except tasks 4, 5, and 7. More specifically, we discuss how we and our students have used them in our own research. These open source tools are based on advances in machine learning, information retrieval, and search engine technologies.

The remainder of the paper is organized as follows. Tools for the first three tasks in the *research workflow* are discussed in section 3. Section 4 presents tools for analyzing and interpreting results, and writing research papers. Our reflections on this work and future research are indicated in section 5.

3 Literature Search, Review, and Assessment

When students embark on research, often they have no clear cut ideas about which area or topics they want to investigate. At best they will have one or two key phrases (e.g., comparative genomic analyses) to begin their exploration. This is also true with seasoned researchers who want to foray into emerging areas (e.g., bigdata visual analytics). A first step in this scenario is to get acquainted with the research topic area as quickly as possible by getting a feel for the "lay of the land." Free tools in this category include Google Scholar (GS)[19] and Ultimate Research Assistant.[4,20] Given a search phrase, GS provides a list of publications. Clicking on a publication will show complete bibliographic information, abstract, citation count, other articles citing this article, and a link to full text of the article (if the article is freely available). Though this is very valuable information

for the expert, it is not that valuable at this exploration stage for student researchers.

Ultimate Research Assistant (URA) is a free Web-based system. Like GS, you specify a search phrase to begin exploration. Optionally, using the *Advanced Search Feature* you may specify search scope (e.g., Wikipedia, National Institutes of Health, educational sites, military sites) and how many retrieved documents to be used for text analysis. The results provided by URA are quite different from those provided by GS. URA provides the following views of retrieved information: Research Report Summary, Tag Cloud, Taxonomy, Mind Map, and a Bar Chart. Essentially all of the above views convey the same information but each uses a different presentation mode. Since URA is a generic tool without domain-specific knowledge, information retrieved is not of high quality. However, it is a useful tool to explore a research topic and obtain a quick conceptual view. The different modes used for presenting the same information caters to the preferences of diverse users.

The next step is to gain a more focused understanding of the domain at a conceptual level. We have used sources such as PLOS journals,[32] ACM Digital Library,[2] IEEE Digital Library,[21] Google Scholar,[19] CiteSeer,[45] and Mendeley[28] to collect full text papers based on key phrases. We downloaded in PDF format 17 papers related to Bigdata Genomics, 40 papers on Bigdata foundations, and 285 papers on Content-based Image Retrieval (CBIR) from the above sources. Plain text from these PDF files is extracted using Tika,[43] an open source content analysis toolkit. Plain text extracted from the Bigdata Genomics papers was input to Wordle™, a Web-based system for generating *word clouds* (a.k.a *tag clouds*). The resulting word cloud is shown in figure 1.

Figure 1: Word cloud of Bigdata Genomics research papers

The figure clearly depicts several domain terms such as RNA, DNA, genome, CloVR, BioGRID, microbial, protein, acids, and bioinformatics. However, the figure also shows terms such as new, using, work, data, and although. The latter are general terms of the language and are not specific to the domain. To improve the word cloud, we manually edited the plain text that is extracted from the PDF files. Linux tools such as *tr* and *sed* were used. More specifically, all upper case letters are replaced with their corresponding lower case, and all occurrences of the words such as technology, work, although, data are removed from the plain text. The latter is again input to

Wordle™. The resulting word cloud is shown in figure 2. Now domain terms such as bases, curation, biocuration, biocurators, assembly, gene, reads, cloud, hadoop, clovr are displayed more prominently.

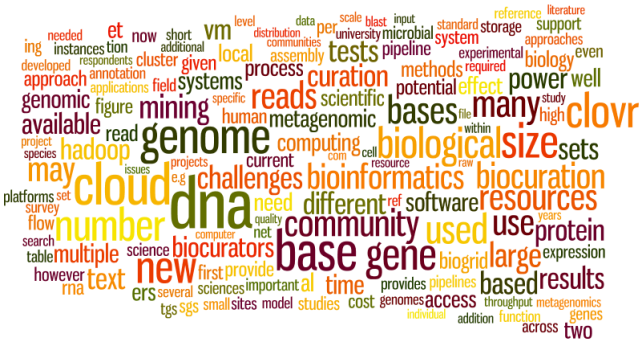

Figure 2: Enhanced word cloud of Bigdata Genomics research papers

Wordle™ offers very little control to the user on how it analyzes plain text. Users may choose options such as removing numbers, making all words lower-case, and removing common English words. However, Wordle™ does not reveal what those common English words are. It does offer various choices for fonts and color, and layout/orientation options for the word cloud. The quality of the generated word cloud critically depends on two factors: ability to use *user-provided* common English words rather than the built-in and fixed list; and, recognizing multiple word domain terms and concepts. Another feature that would be of great benefit to the user is for Wordle™ to permit specification of cut-off point for the most frequently occurring words (e.g., top 50, top 100). *Stemming* is another feature that will help to reduce inflectional and derivationally related forms of a word to a common base form. It is through this type of interactive exploration the user will recognize the conceptual dimensions of the research topic. It should be noted that word clouds can also be formed using only title text or abstract.

There are several other Web-based systems for word cloud generation including Tagxedo, Tagcloud, Tagclouds, and tagcloud-generator. Shown in figure 3 is the word cloud generated using Tagxedo.[42] This word cloud is based on title text of the CBIR papers. It depicts surprisingly large number of domain terms such as content, visual, clustering, classification, indexing, feedback, feature, color, texture, evaluation, distance, and similarity.

Concept extraction,[12,16] single- and multi-document summaries,[16] clustering and classification,[8] and collection understanding[9,15] are means for identifying higher-level domain concepts and to acquire unified view of a research topic. Performing these tasks requires programming expertise with Python and NLTK library for natural language processing.[8,23] Other tools for these tasks include LingPipe[3] and GATE.[46] Open Text Summarizer (OTS) is a ready to use application for single document summarization.[35] A user may specify three options for OTS: size of the summary as a percentage of document length, request for displaying document keyword list, and request for displaying summary in highlighted HTML form. We experimented these tools on foundational Bigdata papers. As these tasks involve rather lengthly steps, details of these processes are omitted.

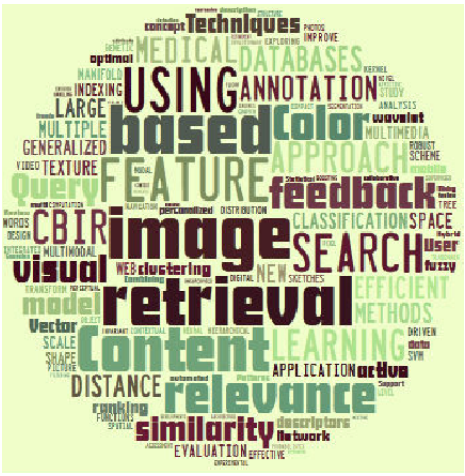

Figure 3: Word cloud of title text of CBIR papers

Ability to write well in general, and technical writing in particular, takes a prolonged period of learning and practice. One aspect of technical writing is maintaining an accurate and relevant bibliography. However, collecting and maintaining a bibliography collection (or database) is a tedious and error-prone task. Of late, there has been a proliferation of software tools for bibliography management, citation, and formatting to suit various bibliographic citation standards. Many of these tools strive to reduce the tedium of authors' bibliography related activities. Furthermore, they maintain a clear distinction between bibliography content and its rendering. Therefore, the same content can be rendered according to various styles as required. This feature promotes bibliography database reuse across multiple related publications and contributes to authors' productivity. Effortless reusability feature is critical to the success of especially beginning researchers.

Though there are many formats available for storing bibliography information, we find BibTeX format versatile. There are several open source tools available which use BibTeX as the underlying format. The tools that we have used include BibDesk[38] and JabRef.[39] BibDesk provides flexible search features, versatile export features, associating abstract, full text paper, and user annotations. Shown in figure 4 is BibDesk managing a collection of 1240 bibliographic records.

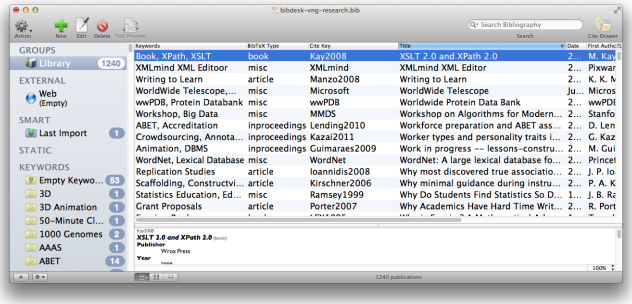

Figure 4: BibDesk bibliography reference manager

4 Analysis, Interpretation, and Write Up

Selecting a suitable research problem to work on, and devising and implementing a solution both require guidance of an experienced researcher. No tools can replace this role at least in the foreseeable future. In the last several years a plethora of sophisticated open source tools have appeared. R for statistical computing,[33] SAGE for symbolic computation,[36] and Octave for linear algebra problems[14] to name a few. Furthermore, numerous Java and Python libraries for tasks ranging from numerical analysis, image processing, to natural language processing are available.[10,22,37,40,41,44] It will serve young researchers well in the long term if they develop expertise with a subset of the above for their field of study. For example, R has over 2,500 packages and 50,000 functions in addition to the core functionality. It takes many years of active work with R to develop even intermediate-level of expertise with it. In additional to its computational capabilities, R provides a wide assortment of very high quality presentation graphics of various types. These aspects of R can be used to convey research results in compelling ways.

As an author, one should be concerned about the structure and content of document rather than its text formatting and layout details. For improving technical writers' productivity, document preparation systems should provide features such as automatic resolution of cross-referencing, generation of table of contents and lists of tables and figures, bibliography management and citation, and index generation. LaTeX is a very high quality typesetting system and provides the above features and more. It is a free software and is available for virtually all computer systems. LaTeX documents are markup documents like HTML, and are portable across many different computer systems. Hundreds of add-ons, called *packages*, extend the core functionality of LaTeX.

LaTeX contributes to effective writing in a number of ways. First, it allows the authors to focus on the document structure rather than its formatting and layout. Since there is a clear separation between document contents and how it is rendered, an author can effortlessly change the way the document is rendered. Major document revisions are easily accomplished since resolution of cross-referencing, and generation of bibliography, index, table of contents, lists of figures and tables are are performed automatically. Therefore, there is merit in using LaTeX as it substantially contributes to improving authors' productivity.

5 Conclusions

Knowledge and skills needed to efficiently conduct independent research, accurately and concisely write research results, and present research findings at professional conferences is not only desirable but also expected of today's undergraduate students. Acquiring these abilities at an undergraduate level pose serious challenges. Based on decades of our own experiences as well as those of the students that we taught and mentored, the solutions we have proposed for a subset of the tasks that comprise the *research workflow* are effective and have withstood the test of time. Furthermore, our solutions are based on open source software that is freely available for everyone. Finally, our solutions are equally beneficial for graduate students and faculty members as well.

Some people seem to have unfounded hesitation to use open source software. They believe that there are hidden costs associated with such software based on the saying that *open source software is free as in getting a free puppy but not free as in getting free beer*. On the contrary, open source software is ideal for use at academic and research institutions. Not because it is free, but more importantly it offers an open laboratory to explore, modify and enhance the software as needed. This process offers unparalleled opportunities to examine the design, architecture, and inner workings of production quality software. The latter are invaluable for students majoring in science and engineering disciplines.

One drawback in the solutions that we have proposed is that they require using various pieces of open source software for each task of the *research workflow*. An integrated workbench that seamlessly unifies various software tools will help to make researchers more productive. Developing such a workbench using foundational tools such as R,[33] Lucene,[17] Tika,[43] and NLTK[8] is our next step.

References

[1] The council of undergraduate research. http://www.cur.org/, 2012.

[2] ACM. ACM Digital Library. `http://dl.acm.org/`, September 2012.

[3] alias-i.com. LingPipe. `http://alias-i.com/lingpipe/`, September 2012.

[4] Andy Hoskinson, LLC. Ultimate Research Assistant. `http://www.ultimate-research-assistant.com/`, September 2012.

[5] Reynold Bailey, Guy-Alain Amoussou, Tiffany Barnes, Hans-Peter Bischof, and Thomas Naps. Relevant real-world undergraduate research problems: lessons from the NSF-REU trenches. In *SIGCSE '10: Proceedings of the* 41st *ACM technical symposium on Computer science education*, pages 62–63, New York, NY, USA, 2010. ACM.

[6] L. Barker. Student and faculty perceptions of undergraduate research experiences in computing. *Trans. Comput. Educ.*, 9(1):5:1–5:28, March 2009.

[7] Jon Beck, Brent Buckner, and Olga Nikolova. Using interdisciplinary bioinformatics undergraduate research to recruit and retain computer science students. In *Proceedings of the 38th SIGCSE technical symposium on Computer science education*, SIGCSE '07, pages 358–361, New York, NY, USA, 2007. ACM.

[8] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[9] Olga (Olha) Buchel. Designing map-based visualizations for collection understanding. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 429–430, New York, NY, USA, 2011. ACM.

[10] Wilhelm Burger and Mark J. Burge. *Digital Image Processing: An Algorithmic Introduction Using Java*. Springer, 2008.

[11] Teresa Dahlberg, Tiffany Barnes, Audrey Rorrer, Eve Powell, and Lauren Cairco. Improving retention and graduate recruitment through immersive research experiences for undergraduates. *SIGCSE Bull.*, 40(1):466–470, March 2008.

[12] Dmitry Davidov, Roi Reichart, and Ari Rappoport. Superior and efficient fully unsupervised pattern-based concept acquisition using an unsupervised parser. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 48–56, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[13] Mary DeVito, Christine Hofmeister, Michael Jochen, and N. Paul Schembari. Undergraduate research in computer forensics. In *Proceedings of the 2011 Information Security Curriculum Development Conference*, InfoSecCD '11, pages 61–68, New York, NY, USA, 2011. ACM.

[14] John W. Eaton. GNU Octave. `http://www.gnu.org/software/octave/`, September 2012.

[15] Melanie Feinberg, Gary Geisler, Eryn Whitworth, and Emily Clark. Understanding personal digital collections: an interdisciplinary exploration. In *Proceedings of the Designing Interactive Systems Conference*, DIS '12, pages 200–209, New York, NY, USA, 2012. ACM.

[16] James M. Foster, Md. Arafat Sultan, Holly Devaul, Ifeyinwa Okoye, and Tamara Sumner. Identifying core concepts in educational resources. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '12, pages 35–42, New York, NY, USA, 2012. ACM.

[17] The Apache Software Foundation. Apache Lucene. `http://lucene.apache.org/`, 2012.

[18] Jinzhu Gao. Developing a web-based collaborative research environment for undergraduates. *J. Comput. Sci. Coll.*, 26(2):39–46, December 2010.

[19] Google. Google Scholar. `http://scholar.google.com/`, September 2012.

[20] Andy Hoskinson. Creating the ultimate research assistant. *Computer*, 38:97–99, 2005.

[21] IEEE Computer Society. CS Digital Library. `http://www.computer.org/portal/web/csdl`, September 2012.

[22] Philipp K. Janert. *Data Analysis with Open Source Tools*. O'Reilly Media, Sabestopol, CA, 2010.

[23] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, second edition, 2009.

[24] Karen A. Kim, Amy J. Fann, and Kimberly O. Misa-Escalante. Engaging women in computer science and engineering: Promising practices for promoting gender equity in undergraduate research experiences. *Trans. Comput. Educ.*, 11(2):8:1–8:19, July 2011.

[25] Deborah L. Knox, Peter J. DePasquale, and Sarah M. Pulimood. A model for summer undergraduate research experiences in emerging technologies. In *Proceedings of the 37th SIGCSE technical symposium on Computer science education*, SIGCSE '06, pages 214–218, New York, NY, USA, 2006. ACM.

[26] Herman Koppelman, Betsy van Dijk, and Gerrit van der Hoeven. Undergraduate research: a case study. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*, ITiCSE '11, pages 288–292, New York, NY, USA, 2011. ACM.

[27] Rick Matzen and Rad Alrifai. Defining undergraduate research in computer science: a survey of computer science faculty. *J. Comput. Sci. Coll.*, 27(3):31–37, January 2012.

[28] Mendeley Ltd. Mendeley: Reference Manager and Academic Social Network. `http://www.mendeley.com/`, September 2012.

[29] Masoud Milani, S. Masoud Sadjadi, Raju Rangaswami, Peter J. Clarke, and Tao Li. Research experiences for undergraduates: autonomic computing research at fiu. In *The Fifth Richard Tapia Celebration of Diversity in Computing Conference: Intellect, Initiatives, Insight, and Innovations*, TAPIA '09, pages 93–97, New York, NY, USA, 2009. ACM.

[30] Joan Peckham, Fatma Mili, Daniela Stan Raicu, and Ingrid Russell. Reus: undergraduate research experiences and funding. *J. Comput. Sci. Coll.*, 23(5):208–211, May 2008.

[31] Jennifer A. Polack-Wahl and Karen Anewalt. Learning strategies and undergraduate research. In *Proceedings of the 37th SIGCSE technical symposium on Computer science education*, SIGCSE '06, pages 209–213, New York, NY, USA, 2006. ACM.

[32] Public Library of Science (PLOS). PLOS Journals. http://www.plos.org/, September 2012.

[33] R. A Free Software Environment for Statistical Computing and Graphics. www.r-project.org/, May 2012.

[34] Daniela Stan Raicu and Jacob David Furst. Enhancing undergraduate education: a REU model for interdisciplinary research. In *SIGCSE '09: Proceedings of the* 40^{th} *ACM technical symposium on Computer science education*, pages 468–472, New York, NY, USA, 2009. ACM.

[35] Nadav Rotem. Open Text Summarizer. http://libots.sourceforge.net/, September 2012.

[36] SAGE. SAGE Mathematics Software System. http://www.sagemath.org/, September 2012.

[37] Toby Segaran and Jeff Hammerbacher. *Beautiful Data*. O'Reilly Media, 2009.

[38] sourceforge.net. BibDesk: Mac Bibliography Manager. http://bibdesk.sourceforge.net/, September 2012.

[39] sourceforge.net. JabRef Reference Manager. http://jabref.sourceforge.net/, September 2012.

[40] Stanford NLP Group. Stanford NLP Tools. http://nlp.stanford.edu/software/index.shtml, September 2012.

[41] Julie Steele and Noah Iliinsky. *Beautiful Visualization*. O'Reilly Media, 2010.

[42] Tagxedo. Word clouds with styles. http://www.tagxedo.com/, September 2012.

[43] The Apache Software Foundation. Apache Tika: A Content Analysis Toolkit. http://tika.apache.org/, September 2012.

[44] The Apache Software Foundation. openNLP. http://opennlp.apache.org/, September 2012.

[45] The Pennsylvania State University. CiteSeerX. http://citeseerx.ist.psu.edu/index, September 2012.

[46] The University of Sheffield. GATE. http://gate.ac.uk/, September 2012.

[47] Thomas P Way. A virtual laboratory model for encouraging undergraduate research. In *SIGCSE '06: Proceedings of the* 37^{th} *SIGCSE technical symposium on Computer science education*, pages 203–207, New York, NY, USA, 2006. ACM.