All Sciences Proceedings

1st International Conference on Scientific and Innovative Studies

April 18-20, 2023 : Konya, Turkey

ICSIS

All Sciences Proceedings
http://as-proceeding.com/

https://as-proceeding.com/index.php/icsis

Dataset for Software Engineering Learning Resources

Muddassira Arshad*, Muhammad Murtaza Yousaf [2] and Syed Mansoor Sarwar [3]

[1]Department of Software Engineering, University of the Punjab, Pakistan
[2]Department of Software Engineering, University of the Punjab, Pakistan
[3]Punjab University College of Information Technology, University of the Punjab, Pakistan

*(muddassira@pucit.edu.pk) Email of the corresponding author

Abstract **–** In the current digital age, an abundance of digital resources is readily available to learners. With the ongoing COVID pandemic and prevalent economic crises, a significant number of learners prefer to engage in self-learning. To develop customized self-learning applications and guide learners to utilize resources based on their learning preferences, a dataset containing learning resources and their prerequisite relationships is required. Several learning resource datasets exist for Machine Learning (ML), Information Retrieval (IR), and Natural Language Processing (NLP). To contribute to this area, we present the Software Engineering Learning Resource Dataset (SELRD), which is a publicly available dataset specifically designed for learning Software Engineering (SE). We have extracted the data for SELRD from multiple sources, including edX, my-mooc, and textbooks. The SE learning resources (SELR) are organized based on topics, and the dataset includes 602 SELRs referring to 302 topics. We have extracted the content from lectures and books available in presentation files (pptx) and Portable Document Format (PDF) using Python libraries. Additionally, we have computed the expected reading time for each SELR, which would facilitate learners by guiding them on the time required to read each respective resource. The SELRD comprises 692 prerequisite pairs, including 592 positive pairs and 100 negative pairs. This data can be used along with machine learning algorithms to generate learning paths that would facilitate self-learners. Additionally, the SELRD can also serve as a repository of SE learning resources. In the future, we plan to add best practices and examples for each SELR, making it even more useful for learners.

Keywords – Software Engineering Learning Resource Corpus, Reading Recommender dataset for Learning Software Engineering, Readability Assessment of Software Engineering Learning Resources, Software Engineering self-learning repository with readability guidance, Software Engineering Concept Map

I. INTRODUCTION

Self-learning has been in-practice for the many years and have been intensified due to COVID and associated economic conditions. In this digital era, several learning resources are also available using which learner may learn and grasp the topics. However, collecting these resources and managing the resources plethora increases information overload.

Several domains have gained attention with the increase of freelancing. Many tasks related to Software Engineering domain are also the in top ten list of freelancing trends[1]. Learners may enhance their learning skills by learning concepts of Software Engineering (SE). Therefore, learning broad field of SE provides an opportunity to improve the Software Development Life Cycle Activities as well as Project management skills

[1] https://www.freelancing.com.au/top-10-in-demand-freelancing-job-skills/

which would facilitate to practice software engineering.

Keeping in view the significance of Software Engineering, we are presenting dataset comprising of learning resources that is topic wise presentations, book chapters for learning software engineering. Our dataset "Software Engineering Learning Resources Dataset (SELRD)" have several usecases. It can not only facilitate customized learning, and could serve as a resource repository, but would also be useful in learning path generation as well as optimization. In addition, it could also be incorporated as a standardization mechanism for learning SE concepts. In addition, it can serve as a valuable resource for data-centric resource recommendation models.

Our designed SELRD provides provenance of the learning resources along with expected reading time. These resources are extracted from publicly available lecture notes, online courses, course slides and Wikipedia pages. In addition, it also encapsulates prerequisite pairs to represent the relationships between the topics. Our dataset will not only serve the purpose of aggregating the resources, but also provide reference to them. SELRD can also be useful for implementing machine learning algorithms to solve concept prerequisite learning, prerequisite chain learning and topic learning recommender systems.

Our paper is organized to cover the existing datasets in Section II, mechanism to generate the dataset in Section III, benchmark characteristics which we have ensured in our dataset are specified in section IV, whereas results are presented in section V. We have summarized our studies in Section VI with concluding remarks in section VII.

II. Existing Datasets

In literature few datasets exist which serve as a repository of learning resources and are helpful in generating the prerequisite chain relationships. These datasets include Lecturebank [1], Tutorial bank [2], NPTEL MOOC[2], University Course datasets (UCD) [3] and ALCPL [4]. Although MOOCCube [5] and MOOCCubeX [6] also provide courses as well as prerequisite information but the publicly available datasets needs translation from Chinese to English language.

Datasets we are considering [2] [1] [3] covers the topics of Machine Learning (ML), Information Retrieval (IR), Natural Language Processing (NLP), Data Mining, Pre-Calculus, Physics, or Computer Science (CS) in general. However we have developed the dataset encapsulating the learning resources of Software Engineering.

A. *Lecturebank*

Li [1] contributed the manually collected dataset comprising of 1352 lecture files for 60 courses in the domains of Natural Language Processing (NLP), Machine Learning (ML), Artificial Intelligence (AI), Deep Learning (DL), and Information Retrieval (IR) from 60 courses referenced from renowned universities. With the vocabulary of 1221 terms, covered in 51939 slides with the total of 2546.65 tokens per lecture. This dataset is rich collection of NLP resources. The dataset also contain annotated list of prerequisite relations mapping the 42750 prerequisite and post-requisite pairs out of which 921 prerequisite relations exist.

B. *Tutorialbank*

In order to facilitate self-learning, keeping in view diversified users who learn with different methods, learning resources of different pedagogical significance, for added to yield Tutorialbank [2] dataset. It not only serves as a collection of diversified learning resources of varied pedagogical significance but also provide massive dataset of diversified field. While Lecturebank comprises of lecture files only, Tutorialbank includes resources categorized to 305 taxonomy topics, and includes surveys, long papers, tutorials, corpus, code bases and libraries from github, link sets, tutorials and NACLOs. Link set represent resource representing the collection of links, whereas NACLOs refer to linguistic puzzles from North American Computational Linguistics Olympiad. Moreover, tutorials include slide deck from conference tutorials. It serves as an updated collection of learning resources for AI, ML, IR,

[2] http://nptel.ac.in/

C. NPTEL MOOC

The National Program on Technology Enhanced Learning (NPTEL) [2,3] dataset comprises of videos based transcription on open learning material on science and technology including Computer Science, Biotechnology, and Engineering disciplines.
It has total of 19500 crawled videos. The sample dataset has over 382 lectures transcripts with vocabulary size of 345, and average transcription size of approximately 28000.

D. University Course Dataset

The rich collection of the courses offered along with their description is aggregated in UCD[4]. It presents the course description of 654 courses. However, it only represents the syllabus to be taught in the courses offered at Princeton, MIT, Stanford, Illinois, Carnegie Melon, Princeton, Maryland, Penn State University (PSU) and Iowa State University. Average course description of the text is 710 characters.

E. AL-CPL

Liang [4] suggested the dataset development using the Wiki Concept Map (WCM) [28]. WCM is the collection of Wiki Concepts based on the concepts acquired from textbooks of Data Mining, Geometry12, Physics 13 and Calculus 14. With 120, 89, 153, and 224 concepts referred in domains of Data Mining, Geometry, Physics, and Pre-Calculus, total of 586 Wikipedia concepts were referred and their data is extracted.

We have selected these datasets as they not only provide reference to the downloadable resources but also prerequisite relationships. However, these datasets refer to the domains of NLP, AI, ML, DL, and Computer Science (CS) but does not include Software Engineering (SE). So we presented the dataset SELRD for learning SE. This dataset follows the recommendation of Lecturebank dataset.

III. Data Extraction and Dataset Generation

First of all, search engines Google, Google Scholar, Bing were used to check which of the top 5 world ranked universities are offering the course of "Software Engineering", "Object Oriented Analysis and Design:, "Software Quality" or "Software Project Management". Later, the publicly available course contents/ lecture notes were searched from these resources. The URLS of these files were noted. The courses of Massachusetts Institute of Technology (MIT), Virginia Tech, and Rutgers were downloaded. In addition, the textbooks were searched and it was noted that Software Engineering by Ian Somerville, Software Engineering by Marcus, have online versions as well as online presentations. Moreover, Software Engineering Body of Knowledge (SWEBOK)[5] provides extensive knowledge about Software Engineering approaches. In addition, sample chapters of "Object Oriented Analysis and Design"[6] by Craig Larman were also publicly available. The publicly available content by the book's authors on their website was noted and downloaded. Content was also extracted and organized with respect to topics from "Software Engineering: A Hands on Approach by Roger Y Lee"[7]. 302 key topics from these books/ lecture slides were noted and were searched using Wikipedia. Their content was also scrapped using Python library "Wikipedia". Two annotators (PhD scholars of Computer Science) annotated the topics and identified prerequisite relationships between the topics with Cohen Kappa measure of 0.86.

F. Meta Data

Meta data comprises of listing 602 resources for learning SELR. These topics are mapped to 302 topics. It consists of following features about each SELR:

[3] https://pypi.org/project/nptel-dl
[4] https://github.com/topics/university-course-dataset
[5] https://www.computer.org/education/bodies-of-knowledge/software-engineering

[6] https://www.craiglarman.com/wiki/index.php?title=Book_Applying_UML_and_Patterns

[7] https://link.springer.com/book/10.2991/978-94-6239-006-5

Table 1 SELRD Metadata Format

Sr.	Field	Description
1	ID	Unique Identifier of SELR
2	URL	In case of lectures referring to the LR/ Wikipedia page. I case of Books' section, it refers to book's URL
3	Topic Title	The main topic presented in SELR
4	Institute	Name of the University from where the lecture notes are extracted
5	Author/ Presented	Author of the book/ Presenter's name in case of the lecture. In case of Wikipedia file it remains empty
6	Reading Time	Reading time required to read the text .Reading time is computed using PyPI[8] 'readtime' library which calculates read time on the average reading speed[9] of an adult which is 265words per minutes.

G. *Pre requisite Annotation*

Using the "Topic Title" extracted in Meta-data specification, prerequisite pairs were identified. Prerequisite annotation comprises of relationship pairs being annotated whether the topic is prerequisite of another. 2 annotators (PhD Scholars at Department of Computer Science) annotated with dataset. The format of each Prerequisite pair is <PreReqRelID, PrereqTopic, postReqTopic, PreReqRelation>. Table 2 presents an overview of the features specified in prerequisite relation.

Table 2 PreReq pairs Annotation Features

Sr. No	Feature	Description
1	PreReqRelID	Unique Identifier for each prerequisite pair
2	PreReqTopic	The topicID of the topic that must be read prior to PostReqTopic
3	PostReqTopic	Topic ID of the topic which requires understanding of PreReqTopic
4	PreReqRelation	PreReqRelation is 1 if the PreReqTopic serves a prerequisite for learning PostReqTopic, otherwise 0.

IV. BENCHMARK CHARACTERISTICS

In literature [7], [8] following characteristics have been considered for specifying the benchmark datasets. We have incorporated the following characteristics in our contributed dataset.

A. *Relevance*

The extracted dataset is relevant to learn the theoretical and conceptual aspects of Software Engineering domain. While Wikipedia specifies the topic's basic idea and its elaboration in non-standard format, books content is also helpful in acquiring knowledge to understand the concepts using examples.

B. *Representativeness*

SE Learning Resources are extracted from diversified resources that is book sections, Wikipedia, and publicly available online Lectures. Therefore, our dataset complies with the representativeness as we have covered several learning platforms.

C. *Non-Redundancy*

In order to remove the redundancy from our dataset, we have not specified book as well as book's section as separate SELR, rather we have split the book into its section. This not only controlled the redundancy but also provide relevance of particular section with the topic.

D. *Experimentally verified cases*

We have verified using experiments that the PreRequisite pairs should not have cyclic relationship indicating, PreReqTopic is a prerequisite of PostReqTopic and vice versa. For 8

[8] https://pypi.org/project/readtime/

[9] https://help.medium.com/hc/en-us/articles/214991667-Read-time

different cases in which this relation existed, we have re-analysed the relationship and removed the weaker relationship.

E. Positive and negative cases

We have included the positive cases where topics pair have prerequisite relation. In addition, we have also specified the pairs which are considered as prerequisites but annotators marks the pair as non-prerequisite. This would facilitate comprehensive assessment of the dataset.

F. Scalability

Our dataset is scalable as simple format is used to specify the dataset. We would increase the dataset size and incorporate data specifically representing the definitions, examples, and case studies referring to the existing specified topics. In addition, more topics can be added in easy to specify mechanism for meta-data as well as prerequisite relations. Therefore, the dataset is vertically as well as horizontally scalable.

G. Reusability

Metadata as well as prerequisite data is presented in csv file formats. Since the dataset is open, and can be publicly accessed, it can be reused in several usecases like curriculum design, machine learning algorithms for recommender systems.

H. Address atleast one clear Machine learning Task

Our dataset is useful in recommender system where the learning path could be recommended to the learner on the basis of prerequisite relationships between the topics.

İ. Open

Our dataset is open to use for non-commercial usage.

J. Discoverable and accessible

Since the dataset is available on github, it is accessible and discoverable.[10]

K. Enough features to be interesting

Table 1 and Table 2 provides relevant features for generating the recommender system for learning Software Engineering theoretical and conceptual effects.

L. Labels

SELRD is labelled to represent the topics of each SELR, Therefore SELRD can be useful in classifications approaches.

M. De-localized

It can be easily delocalized by removing the Table 1 features like institute, author/presenter, and would still facilitate the Machine Learning Task for recommender system.

N. Not be too big

Since the dataset size of textual content of SELR is 56MB, it is easily downloadable and process able. However, in future we would add more SELR, which would increase dataset size. However, since we only plan to include the text description referring the SELR, the dataset size would not grow rapidly.

O. Well documented

The dataset is well documented with self-descriptive feature names.

P. Clean but not too clean

SELRD is clean as we have minimized blank entries. However, spelling mistakes in the existing SELR are not rectified. Moreover redundant words like “in today’s lecture, “etc.” have been removed.

V. RESULTS

We have evaluated SELRD using the benchmark characteristics with our dataset complies to commonly discussed characteristics of benchmark dataset.

The prerequisite annotation was conducted by two annotators. Both the annotators are PhD Scholars (at Department of Computer Science, University of the Punjab) and have effective technical soundness. In addition, inter-annotator agreement of 0.86. The prerequisite dataset was assessed whether there are any cyclic relations in dataset. In this connection, we have evaluated that if topic A is a prerequisite relation of B, then B should not be the prerequisite relation of A.

[10] https://github.com/MuddassiraSheraz/SELRD

VI. DISCUSSION

In context of prerequisite learning and our objective of contribution of dataset for learning repository of Software Engineering, it is analysed the discussed datasets present the learning repositories as well as prerequisite relations for CS domain in general, or specifically for Machine Learning and associated fields of Natural Language Processing, Deep Learning, and Information Retrieval. ALCPL also discusses domains of Pre-Calculus, Computer Networks, Data mining however, they also didn't exclusively include Software Engineering. Therefore, SELRD provides exclusive repository for learning Software Engineering Learning Resources. However, at present we have included the learning resources from publicly available content from textbooks, PowerPoint presentations and Wikipedia. At present we have not included NACLOs, and link sets. SELRD comprises of major topics covered in standard textbooks and undergraduate lecture series. This would be helpful in generating the prerequisite chain learning, course planning, and developing basic understanding of Software Engineering concepts.

VII. CONCLUSION

We have contributed novel benchmark dataset for learning theoretical aspects of field of Software Engineering, which would not only serve as a repository but would also be used for resource recommendation, learning path generation as well as optimization and curriculum development. We have ensured that SELRD complies with the benchmark characteristics discussed in literature [8]. However, since our dataset does not cover case studies, standards, and research papers, in future we would scale it horizontally as well as vertically by adding more diversified learning resources like standards, case studies, case tool Tutorials and research papers and also by adding more SELR to existing categories respectively.

.

ACKNOWLEDGMENT

The authors would like to acknowledge Prof. Dr. Shahzad Sarwar, HoD CS Department, University of the Punjab, and Prof. Dr. Kamran Malik, Associate Professor, IT Department, University of the Punjab for their support.

REFERENCES

[1] I. Li, A. R. Fabbri, R. R. Tung and D. R. Radev, "*What should I learn first: Introducing lecturebank for NLP education and prerequisite chain learning*," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[2] A. . R. Fabbri, I. Li, P. Trairatvorakul, Y. He, W. Ting, R. Tung, C. Westerfield and D. Radev, "*TutorialBank: A Manually-Collected Corpus for Prerequisite Chains, Survey Extraction and Resource Recommendation*," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018.

[3] C. Liang and J. Ye, "*Recovering Concept Prerequisite Relations from University Course Dependencies*," in Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[4] C. Liang, "*Investigating active learning for concept prerequisite learning*," in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[5] J. Yu and G. Luo, "*{MOOCC}ube: A Large-scale Data Repository for {NLP} Applications in {MOOC}s*," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[6] J. Yu and Y. Wang, "*MOOCCubeX: a large knowledge-centered repository for adaptive learning in MOOCs*," in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021.

[7] A. Sarkar and Y. Yang, "*Variation benchmark datasets: update, criteria, quality and applications*," Database, vol. 2020, 2020.

[8] M. Hall, "*What makes a good benchmark dataset?*," 03 04 2019. [Online]. Available: https://agilescientific.com/blog/2019/4/3/what-makes-a-good-benchmark-dataset. [Accessed 02 03 2023].