Natural Language Processing and Data Science in Cambodia

PEANG RATANA,, AFFILIATE: STEM CLUB CAMBODIA , CORRESPONDENT AUTHOR: CAMBODIASTEMCLUB@GMAIL.COM OR PRATANA1@PUTHISASTRA.EDU.KH

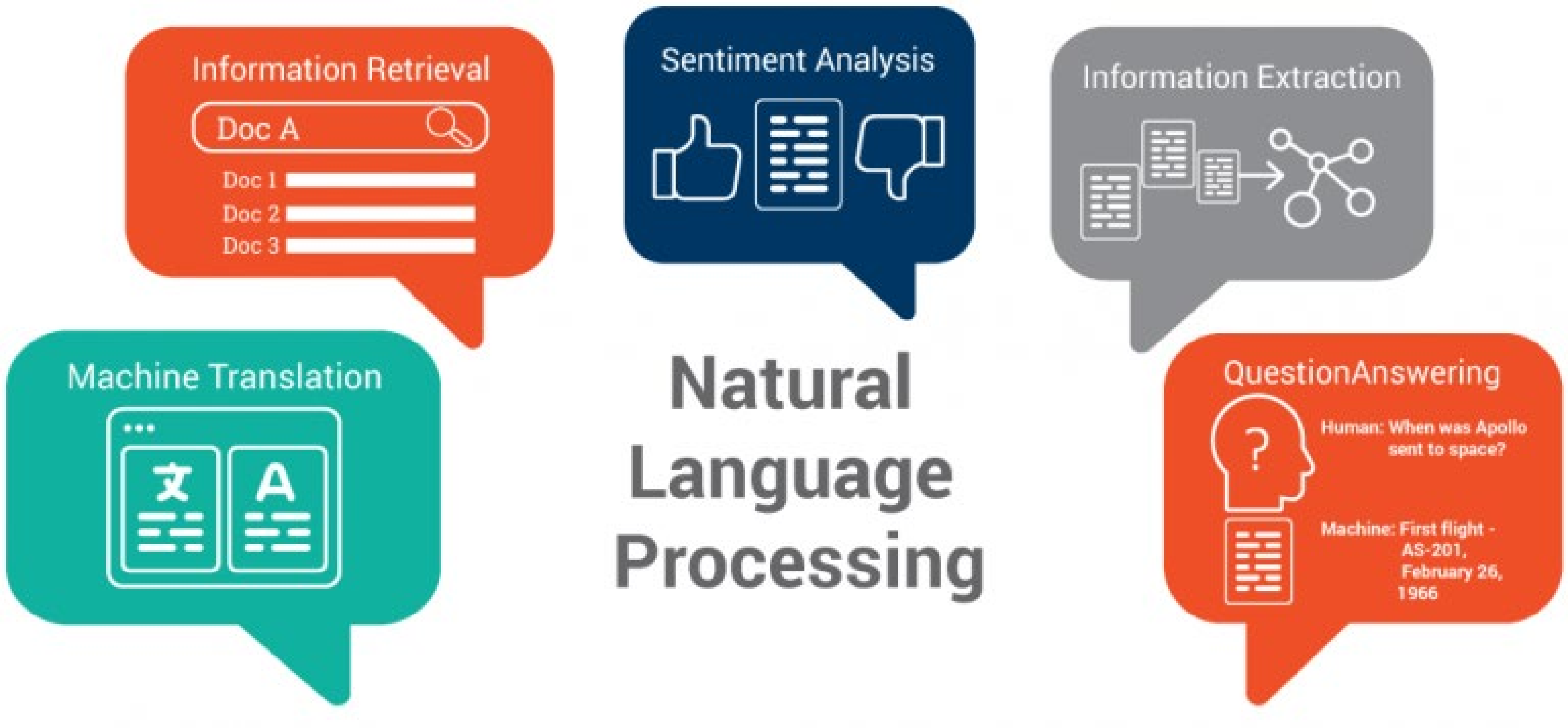
Abstract

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data. Data Science can cultivate learners "reflective inquiry and improve their linguistics ability. How this cultivation is achieved by Natural language processing and Data Science activities remains under-explored. This paper aims to explore and provide insights into how the obstacle of learning Natural language processing, Data science experience and is based on a pilot project which involves three groups of students or experts, researchers, officials in three content-based (i.e. how does jobs market for data science, what do they use Natural language processing and how did you use Data Science). The study employs ongoing observations, online interactions and three focus-group interviews as the methods for data collection. The finding on the obstacle of learning data science show that students rarely knew what is Natural language processing, Data Science and think it has small jobs market, and researchers, officials knew how they use Data Science for research in limited.

Keywords: NLP, Data, Science, learning, Cambodia, Market, Research, Effect
Citation: Peang Ratana (2021). Natueal Language and Data Science in Cambodia

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. a Natural Language Processing system that will detect ancient Khmer language script in Angkor Wat, the temple complex in Siem Reap in north Cambodia, and analyze it. “There are only a few specialists in the world who can readthose writings,” De Vos said, “and they’re typically based in affluent countries (**M.G. Zimeta** October 25, 2018). Data Science is the "fourth model" of science that "everything about science is changing because of the influence of information technology," said Jim Gray, winner of the Computer Science Award.. Identifying the need for such innovation, the Ministry of education, Youth and Sport (M0EYS) has been promoting the use of information and communication technology (ICT)in higher education institute (HEIs) classrooms, making innovative thinking, communication, problem solving skills, research and information retrieval, and processing skills the focal points in teaching and learning (M0EYS, 2009-2013). There has been a recent shift away from teacher-centered teaching, such as lectures, in Cambodian schools. This has resulted in an opportunity to implement innovative instruction methods that best fulfill the needs of both individuals and society in the present and the future (M0EYS, 2014).

In Cambodia, in 2020 the ASEAN Foundation introduced a program entitled ASEAN Online Training on Data Science for Cambodia, which had the main aim of using Data Science to help country to enhance their educational quality in order to meet the demands of understanding and sharing information in education, Francis Bacon (1561-1626) once said " Knowledge is power." Data Science describes the process in which data/information is collected, analyzed and interpreted in order to advance our understanding of a topic or issue or to find a new solution for a particular problem. Data Science findings are very useful for establishing communication and a dialogue with researchers, policy makers, planners and practitioners and to guide practical development interventions. Data Science Is the most widely used technique among Artificial Intelligence and Machine Learning to analyze data and make predictions about the future. The overarching objective of this study is to help build students confidence in the use of Data Science in learning in their class. This study identified that, given the amount of resources and other learning aid choices that higher education institution in Cambodia currently have in place, determining the fundamental requirements to achieve the full learning of Data Science. This study implemented Data Science as a mean to assist the learning process in higher education institution, placing Data Science skills, working in a team, solving complex problems at the center of the investigation.


Information Retrieval
Doc A
Doc 1
Doc 2
Doc 3
Sentiment Analysis
Information Extraction
Machine Translation
文
A
Natural Language Processing
QuestionAnswering
?
Human: When was Apollo sent to space?
Machine: First flight - AS-201, February 26, 1966

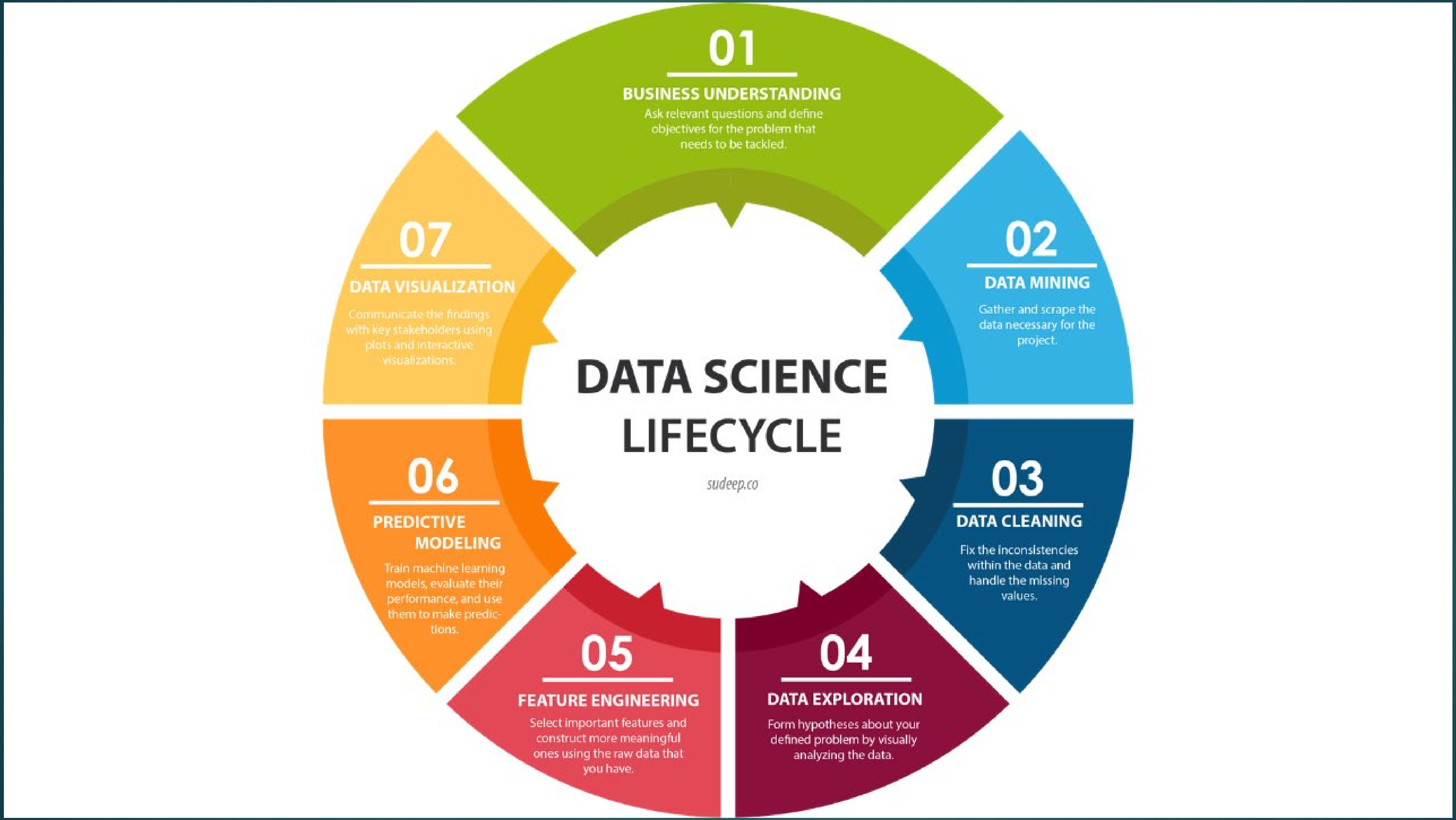
The Natural Language Processing

Natural language processing (NLP) is concerned with enabling computers to interpret, analyze, and approximate the generation of human speech. Typically, this would refer to tasks such as generating responses to questions, translating languages, identifying languages, summarizing documents, understanding the sentiment of text, spell checking, speech recognition, and many other tasks. The field is at the intersection of linguistics, AI, and computer science. The field is at the intersection of linguistics, artificial intelligence, and computer science. The goal? Enabling computers to interpret, analyze, and approximate the generation of human languages (like English or Spanish). NLP got its start around 1950 with Alan Turing's test for artificial intelligence evaluating whether a computer can use language to fool humans into believing it's human. But approximating human speech is only one of a wide range of applications for NLP! Applications from detecting spam emails or bias in tweets to improving accessibility for people with disabilities all rely heavily on natural language processing techniques.

The Roles of learning of Data Science in Higher Education Institution

“It used to be that if you were a mathematician you became a teacher, and then you became a geologist because your knowledge helped find oil, and then you went over to Wall Street,” said Jim Sterne, chairman of the Digital Analytics Association. “And now, the quote is, the best minds of our generation are being put to work in advertising.” Some part of Data Science have been widely integrated into the local universities’ curricula in Cambodia. In fact, some part of Data Science instruction has been used in language teaching, learning, analyzing as a key for identifying and eliciting knowledge from various existing sources including documents, datasets that can be verified and utilized.

Data Science is an effective choice for learning because it motivates students to analyze, to deriving hypotheses from data (*exploratory setting*), i.e., trying to understand the data before hypothesizing. With the rise of Industry 4.0 manufacturing as we know it will change over the coming years. As a leader in Data, student will be one of the proud Cambodians that helps reshape the Kingdom and its industries. Data Science can identify valuable data sources and automate collection processes, undertake preprocessing of structured and unstructured data, analyze large amounts of information to discover trends and patterns, build predictive models and machine-learning algorithms, combine models through ensemble modeling, present information using data visualization techniques, introduction to Data Science in real world, project Life cycle.



DATA SCIENCE LIFECYCLE

sudeep.co

01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.

02
DATA MINING
Gather and scrape the data necessary for the project.

03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.

04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.

05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.

06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.

07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

Conceptualizing the Effect of Data Science Outcomes

By the present technology, companies looking to hire data scientists from within need to think about how to upskill, reskill or preskill their data analysts to perform the roles needed to implement AI and ML fully. A believer that if you provide prescriptive and progressive curricula around the essential topics a budding data scientist needs, you can equip them with the skills and knowledge to make a difference and move them forward in the organization.

Data Acquisition Understand Machine Learning, algorithms Study the tools and techniques of Experimentation, evaluation and Project Deployment, understand the concept of Prediction and Analysis Segmentation through Clustering, understand the basics of Big Data and ways to integrate R with Hadoop Get trained about the roles and responsibilities of a Data Scientist, Explored steps to install IMPALA Live Projects on Data science, analytics and Recommender System, work on data mining, data structures, data manipulation. Data Scientist can become Big Data Specialists, Business Analysts and Business Intelligence professionals Statisticians skills, Developers, Techniques Information Architects looking to learn Predictive Analytics Those looking to take up the roles of Data Scientist and Machine Learning Expert.

Study Area and Methodology

The first National Conference on Research and Innovation in Cambodia (November 21-22, 2019) group and Cambodia AI Community, Startup Jungle are the research site for the study. During the conference there were 106 researchers, officials, Professors, teachers, students and 362 people I n Cambodia AI Community 359, Startup Jungle 194 people.

In Natural language processing (NLP) like information retrieval, machine translation, speech processing, etc.; words boundary is very important (Narin Bi and Nguonly Taing, 2014). And " Around 90% of Cambodian populations speak this language in Cambodia and also some speakers live in Vietnam, Thailand, U.S.A, France, Australia, and Canada. Khmer language (Cambodian) is one of the under resourced Southeast Asian languages for natural language processing (NLP). It is a SVO (Subject, Verb and Object) language. Syntactically it is quite similar to Chinese and English, and also it is similar to Japanese, Chinese, and Myanmar in the word composition" (Soky, 2016).

Data Science were customized to leverage learning interest and outcomes. With this, the study aimed to take part in addressing the aforementioned expectation in the two cases. In these cases, people were assigned to respond the questions. Data exploration or *data mining* is fundamental for the proper usage of analytical methods in Data Science. The most important contribution of statistics is the notion of distribution. It allows us to represent variability in the data as well as (a-priori) knowledge of parameters, the concept underlying Bayesian statistics. Distributions also enable us to choose adequate subsequent analytic models and methods.

The study utilized a qualitative research paradigm as it was exploring experiences and perceptions that are practical in nature (Babbie,1992). In this study, the researcher used both qualitative and quantitative approaches as a mixed method. Qualitative approach that is being used by using interview open question aligned to the qualitative standard questionaires. To get results from qualitative interview. Two kinds of collecting data collection methods are used, survey and in depth interview.

The research methodology is collected both primary and secondary data. The primary data was collected through the questionnaire which were developed in order to get the information through in depth interview and the standard questionaires for quantitative data to explore have they ever used NLP, how does jobs market for Data Science skills, what do they use Data Science for. The secondary data was obtained from existing resources such as books, journal, articles, internet, and previous research records. Furthermore, this paper elaborates the technique and method of data processing and the procedure of data analyzing using statistical tools for quantitative and qualitative analysis that include descriptive and analytical analysis.

The current people in The first National Conference on Research and Innovation in Cambodia is 106, and Cambodia AI Community, there were 359, and Startup Jungle, there were 194. Among them there were researchers, teachers, professors, developers, students.

The data collection was started from 10 st October 2020 to 23 st October 2020. Due to the availability of the participants, in depth interview is helpful to get the qualitative data and 69 persons among them was selected.

Results and Findings

One participant

Other participant from Mobile App Dev said that the market jobs for Data Science was too small but it was a good skills and high salary He added

" I was recognizing he would study this skill or not , hard decision ".

The total numbers of research respondents: 12 person from Startup Jungle group, 41 person from Cambodia AI Community and 16 person from The first National Conference on Research and Innovation in Cambodia.

Table 1: Have you ever use NLP		
	Frequency	Percent
Yes	3	25%
No	7	58%
Other	2	17%
	Total= 12	

By the result of table 1, the number that is showing in data collection it says that, they ever used NLP is 25%%, never used is 58% and other is 17%.

Table 2: Data Science for		
	Frequency	Percent
Research	11	69%
Data Mining	4	25%
Other	1	6%

By the result of table 2, the number that is showing in data collection it says that, the research is 69%, data mining is 25% and other is 6%. In this survey, found 3 types of using Data Science.

Table 3: How does jobs market for Data Science skills		
	Frequency	Percent
Many	8	20%
Less	28	68%
Other	5	12%

By the result of table 3, the number that is showing in data collection it says that, the jobs market for Data Science skills is very low for them to believe on these skills.

This work was supported by **STEM Club Cambodia (SCC).**

Brief Biographies

Peang Ratana completed a Master Degree in Science in Mathematics from Royal University of Phnom Penh & CIMPA. After completing Master's degree in Cambodia, I taught at Pannasastra University of Cambodia, International University, University of Management and Economics and University of Puthisastra. Here, I also attended Summer School at NIMS, Daejeon, South Korea about "Symplectic Embeddings systolic inequalities and celestial mechanics". Attended ASEAN Science Assembly Diplomats held at Davao City ,Philippine, and attended Common Purpose ASEAN Young Leaders Program at Singapore Institute of Technology, Singapore. In 2019 I published my research paper on STEM Education at Royal University of Agriculture, 2021 publish my research paper on Learning in Digital Era at TCI/MOC Asia Conference at CamEd Business School.

- Website: https://www.researchgate.net/profile/Peang-Ratana
- https://www.linkedin.com/in/peangratana/

References

- [1] Bowman, R. (2018): Teaching and Learning in the Age of Questions, The Clearing House: A journal of Educational Strategies, Issues and Ideas.
- [2] Creswell, J.W. (2003). Research design: Qualitative, quantitative, and mixed approaches. Thousand Oaks, CA: Sage.
- [3] Johnson, D. W., Johnson, R. T., &Smith, K. A. (2014). Cooperative learning: Improving university instruction by basing practice on validated theory. Journal on Excellence in University Teaching, 25(4), 1-26.
- [4] Miranda and et al. (2015). Collaborative learning in higher education: lecturers' practices and beliefs, Research Papers in Education, 30:2, 232-247, DOI: 10.1080/02671522.2014.908407
- [5] Ministry of Education Youth and Sports. (n.d.). Retrieved December 1, 2017, from http://www.moeys.gov.kh/en/education/higher-education.html#.WiDIgdKWbIU
- [6] MoEYS. (2016, October 27). Draft Budget Moves Ahead, Boosts Education, Phnom Penh, Cambodia.

- [7] Prashant G Desai, Saroja “A Study of Natural Language Processing Based Algorithms for Text Summarization” Devi Niranjan N Chiplunkar, Mahesh Kini M.
- [8] Brown, M.S.: Data Mining for Dummies. Wiley, London (2014)
- [9] Donoho, D.: 50Years ofData Science. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf (2015)
- [10] Press, G.: A Very Short History of Data Science. https://www.forbescom/sites/gilpress/2013/05/28/a-very-short-history-ofdata-science/#5c515ed055cf (2013). [last visit: March 19, 2017]
- [11] Wu, J.: Statistics = data science? http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

- [12] Philipp Koehn and Hieu Hoang. Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 868–876, 2007.

- [13] Andreas Stolcke. Srilm-an extensible language modeling toolkit. In Proceedings International Conference on Spoken Language Processing, pages 257– 286, November 2002.

-

- [14] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

- [15] Philipp Koehn and Barry Haddow. Edinburgh's submission to all tracks of the wmt2009 shared task with reordering and speed improvements to moses. In Proceedings of the Fourth Workshop on Statistical Machine Translation, StatMT '09, pages 160–164, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

Any question ?