Ho Chi Minh University of Education

**Department of Information Technology**

|  |
| --- |
|  |
|  |  |

**Report No.2 – 09/02/2020**

**CodEbot-AITeacher**

**Main Supervisor: Professor Hung Nguyen**

**Students: Nguyen Vuong, Tam Luong**

Copyright © 2020 by Nguyen Vuong, Tam Luong

# CONTENTS

[CONTENTS 3](#_Toc32184856)

[FIGURES 3](#_Toc32184857)

[Chapter 1. Survey Questions 4](#_Toc32184858)

[1.1. Programming FAQ (for building dataset) 4](#_Toc32184859)

[1.2. Side information (for papers’ motivation) 4](#_Toc32184860)

[Chapter 2. Data Labelling Pipeline 5](#_Toc32184861)

[2.1. Question answering 5](#_Toc32184862)

[a. Expected dataset layout (2 cols) 5](#_Toc32184863)

[b. Expected size of dataset: 5](#_Toc32184864)

[2.2. Error explaining and solving 5](#_Toc32184865)

[a. Expected dataset layout (4 cols) 5](#_Toc32184866)

[b. Expected size of dataset 5](#_Toc32184867)

[References 6](#_Toc32184868)

# FIGURES

***No table of figures entries found.***

# Survey Questions

## Programming FAQ (for building dataset)

* “Hãy cho biết khi mới học lập trình, bạn thường hay đặt những câu hỏi như thế nào? (Cung cấp 5 câu hỏi về lý thuyết lập trình hoặc thực hành với C/C++ hoặc Python)  
  **Một vài câu hỏi ví dụ:**  
  TLE là gì?  
  Làm thế nào để giải quyết TLE?”
* “Lỗi bạn thường gặp khi mới học lập trình C/C++ hoặc Python là gì? (Cung cấp 5 câu báo lỗi điển hình)”

## Side information (for papers’ motivation)

* “Khi gặp một lỗi cú pháp trong lập trình, các bạn thường làm gì để tự gỡ lỗi? Việc tự gỡ lỗi đó có thường làm khó bạn hay lấy của bạn quá nhiều thời gian (vài tiếng) hay không?”
* “Theo bạn, một phần mềm có thể giải quyết hầu hết những câu hỏi trên cho bạn thay vì bạn phải đi tìm người giải đáp thắc mắc cho mình hay tốn thời gian tìm kiếm một lời giải tương tự trên Internet có cần thiết hay không? (đánh giá theo thang điểm 5)”
* “Giả sử tồn tại một phần mềm như thế, bạn mong muốn nó sẽ tập trung giải đáp thắc mắc về mặt lý thuyết của lập trình hay giải thích lỗi cú pháp và hướng dẫn sửa lỗi thường gặp?”

# Data Labelling Pipeline

## Question answering

Each question from the survey will be extended to a few extra questions by labelers. After that, for each question in the extended question set, labelers will provide an accurate answer for it.

### Expected dataset layout (2 cols)

|| Question || Answer ||

### Expected size of dataset:

100\*5\*20 = 10000 (samples)

## Error explaining and solving

For each syntactic error statement, labeler should provide a simple explanation, the most usual solution the problem, and a recommendation for questioner to avoid this type of error in the future.

### Expected dataset layout (4 cols)

|| Error statement || Explanation || Solution || Recommendation ||

### Expected size of dataset

100\*5 = 500 (samples)

# References

1. D. A. Ferrucci, "Introduction to “This is Watson”," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 1:1-1:15, May-June 2012.
2. A. Lally et al., "Question analysis: How Watson reads a clue," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 2:1-2:14, May-June 2012.
3. M. C. McCord, J. W. Murdock and B. K. Boguraev, "Deep parsing in Watson," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 3:1-3:15, May-June 2012.
4. J. Chu-Carroll, J. Fan, N. Schlaefer and W. Zadrozny, "Textual resource acquisition and engineering," in IBM Journal of Research and Development, vol. 56, no. .4, pp. 4:1-4:11, May-June 2012.
5. J. Fan, A. Kalyanpur, D. C. Gondek and D. A. Ferrucci, "Automatic knowledge extraction from documents," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 5:1-5:10, May-June 2012.
6. J. Chu-Carroll, J. Fan, B. K. Boguraev, D. Carmel, D. Sheinwald and C. Welty, "Finding needles in the haystack: Search and candidate generation," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 6:1-6:12, May-June 2012.
7. J. W. Murdock et al., "Typing candidate answers using type coercion," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 7:1-7:13, May-June 2012.
8. J. W. Murdock, J. Fan, A. Lally, H. Shima and B. K. Boguraev, "Textual evidence gathering and analysis," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 8:1-8:14, May-June 2012.
9. C. Wang, A. Kalyanpur, J. Fan, B. K. Boguraev and D. C. Gondek, "Relation extraction and scoring in DeepQA," in IBM Journal of Research and Development, vol. 6, no. 3.4, pp. 9:1-9:12, May-June 2012.
10. A. Kalyanpur et al., "Structured data and inference in DeepQA," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 10:1-10:14, May-June 2012.
11. J. M. Prager, E. W. Brown and J. Chu-Carroll, "Special Questions and techniques," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 11:1-1:13, May-June 2012.
12. J. Chu-Carroll, E. W. Brown, A. Lally and J. W. Murdock, "Identifying implicit relationships," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 12:1-12:10, May-June 2012.
13. A. Kalyanpur, S. Patwardhan, B. K. Boguraev, A. Lally and J. Chu-Carroll, "Fact-based question decomposition in DeepQA," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 13:1-13:11, May-June 2012.
14. D. C. Gondek et al., "A framework for merging and ranking of answers in DeepQA," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 14:1-14:12, May-June 2012.
15. E. A. Epstein, M. I. Schor, B. S. Iyer, A. Lally, E. W. Brown and J. Cwiklik, "Making Watson fast," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 15:1-15:12, May-June 2012.
16. G. Tesauro, D. C. Gondek, J. Lenchner, J. Fan and J. M. Prager, "Simulation, learning, and optimization techniques in Watson's game strategies," in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 16:1-16:11, May-June 2012.
17. B. L. Lewis, "In the game: The interface between Watson and Jeopardy!" in IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 17:1-17:6, May-June 2012.