Ao Li 798657

aol3@student.unimelb.edu.au

Project proposal

Clinical Text De-identification and Colander

# 1 Background

Since, with the development of Electronic Health Records (EHR) system, a large amount of patient data nowadays can be easily provided for clinicians, professors or authorities for several purposes, such as diagnosis, analysis and statistics, and the confidentiality of the information to a healthcare provider has been guaranteed as a cornerstone to make the trust relationship to the patients (Meystre et al, 2010). So, the protection of the confidentiality data should be think highly of as it has become more readily accessible and operational.

Currently, there are two major type of methods for protecting the confidentiality of the data, which are *De-Identification* Approach and *Colander* Approach.

1. De-Identification Approach --- is basically a process to remove patients’ confidentiality information from the clinical texts so as to protect their personal details.
2. Colander Approach --- is an opposite way of processing to the de-identification approach. Key idea is to pre-process and only keep the terms that we want in the clinical texts, which means that it considers all other terms as the Protected Health Information (PHI) and then get rid of all of them.

Thus, based on the two approaches, essentially, there are two aspects contained in this project:

1. Building external specific dictionary for improving the de-identification approaches’ performance. Some dictionaries will be created for some of the important PHI category and provide some analysis based on the dictionaries.
2. Implementing colander approach for some clinical analysis, such as popularity of the disease for some specific regions etc. Some scenarios will be built and test the efficiency of the colander approach.

# 2 Methodology

To well implement the project, the following steps of process will be conducted:

|  |  |
| --- | --- |
| Method | Process |
| De-identification Approach | 1. Establish various external specific dictionaries on different PHI category. 2. Evaluate the sensitive of the performance to each of the specific dictionary. 3. Estimate the size of samples are needed for learning a good dictionary. 4. Compare the difference performance on the established dictionaries with online open source dictionaries. 5. Critical analyse the overall performance of de-identification approach. |
| Colander Approach | 1. Establish the protected words dictionary. 2. Estimate the performance of the colander approach. 3. (optional) Extend Colander Approach |

# 3 Time Schedule and Preliminary Results

To ensure the quality of each part in the project, the following plan will be considered:

|  |  |
| --- | --- |
| Time Schedule | Preliminary Results |
| 01/03/2018 - 01/04/2018 | 1. Complete all of the analysis on de-identification approach part. 2. Finish the corresponding part in report. |
| 01/04/2018 - 01/05/2018 | 1. Complete all of the analysis on colander approach part. 2. Finish the corresponding part in report. |
| 01/05/2018 - 01/06/2018 | 1. Check if each part need to do the extension. 2. Compare and evaluate the two approaches’ advantages and disadvantages. 3. Conclude the analysis and complete the whole report. |

# 4 Reference

Meystre, SM, Friedlin, FJ, South, BR, Shen, S, & Samore, MH, 2010, ‘Automatic de-identification of textual documents in the electronic health record: a review of recent research’, *Bmc Medical Research Methodology,* vol.10, pp.70