VIETNAM GENERAL CONFEDERATION OF LABOUR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**MID TERM EXERCISE**

**KNOWLEDGE DISCOVERY AND DATA MINING**

**MINING AND CONVERTING**

**TDTU STUDENT AFFAIRS DATA**

**INTO SQuAD FORMAT**

*Instructor*: **Ph.D TRẦN THANH PHƯỚC**

*Student*: **NGUYỄN QUỐC SƠN – 51800921**

**VÕ QUỐC SƠN – 51800922**

Class **: 18050301**

**HO CHI MINH CITY, 2020**

VIETNAM GENERAL CONFEDERATION OF LABOUR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**MID TERM EXERCISE**

**KNOWLEDGE DISCOVERY AND DATA MINING**

**MINING AND CONVERTING**

**TDTU STUDENT AFFAIRS DATA**

**INTO SQuAD FORMAT**

Instructor: **Ph.D TRẦN THANH PHƯỚC**

Student: **NGUYỄN QUỐC SƠN**

**VÕ QUỐC SƠN**

Class **: 18050301**

**HO CHI MINH CITY, 2020**

TABLE OF CONTENTS

[TABLE OF CONTENTS 1](#_Toc52709368)

[CHAPTER 1 – INTRODUCTION TO SQUAD 2](#_Toc52709369)

[1.1 What is SQuAD ? 2](#_Toc52709370)

[1.2 SQuAD format 2](#_Toc52709371)

[CHAPTER 2 – STUDENT AFFAIRS DATA CRAWLING 3](#_Toc52709372)

[2.1 Data resource 3](#_Toc52709373)

[2.2 Crawling methods 3](#_Toc52709374)

[CHAPTER 3 – CONVERT DATA INTO SQUAD FORMAT 4](#_Toc52709375)

CHAPTER 1 – INTRODUCTION TO SQUAD

1.1 What is SQuAD ?

SQuAD (**S**tanford **Qu**estion **A**nswering **D**ataset) is a dataset on which we train a learning model to answer questions from a given comprehension. SQuAD was created out of crowdsourcing, where crowd workers were posed questions on Wikipedia articles for which the answers were answerable from the comprehension. Learning algorithms can then be trained on this model and then could be tested. This dataset has been extensively used in the research community for evaluation purposes.[1]

1.2 SQuAD format

SQuAD format as a JSON file which contains:

* version : (version of the dataset)
* data :
  + title : (title of the document)
  + paragraphs :
    - context : (content of the paragraph)
    - qas :
* question : (content of the question)
* id : (question\_id)
* answer :
* answer\_start : (the start index of answer span in paragraph)
* text : (content of the answer)

CHAPTER 2 – STUDENT AFFAIRS DATA CRAWLING

2.1 Data resource

The data comes from several website :

* <https://student.tdtu.edu.vn/gioi-thieu-phong-cthssv>
* <https://student.tdtu.edu.vn/chinh-sach>
* TDTU student portal
* <https://www.tdtu.edu.vn/giao-duc/cong-khai-thong-tin>
* <https://student.tdtu.edu.vn/>

The dataset includes: student affair data, scholarship data, rule-regulation data and department data (for model’s knowledge enrichment purpose).

2.2 Crawling methods

Programing language: Python.

Framework: beautifulSoup.

To get content of the website, we prefer using request. After that, the content needs to be filtered to get the needed data. To do that, we use the beautifulSoup framework, the powerful framework for data filtering. BeautifulSoup filter data base on html tag, id or class. Finally, we store data in a csv file for converting purpose.

CHAPTER 3 – CONVERT DATA INTO SQUAD FORMAT

Firstly, we use the function “df2squad” provided by **cdQA** repository on github to covert the pandas dataframe to a blank SQuAD json file with no pairs of question, answer.

Second step, we use the web-based annotation tool also provided **cdQA** to create the pairs of question and answer of each paragraphs. The tool will automaticaly generate the answer\_start and the question\_id. After get that job done, just click the “Download” button and we got the complete JSON file in SQuAD format.

**REFERENCES**

[1] <https://medium.com/@deepaknarayanan120199/squadrun-squad-2-0-from-stanford-c620b12918a4> (Access time: 10/4/2020)

[2] <https://github.com/cdqa-suite> (Access time: 10/4/2020)

[3] <https://towardsdatascience.com/how-to-create-your-own-question-answering-system-easily-with-python-2ef8abc8eb5> (Access time: 10/4/2020)