# Thông tin bài tập cuối kỳ

Tên sinh viên: Nguyễn Đức Kiên

Lớp : D10CNPM1

MSV: 1021040365

Tên bài : tweetfeel



Menu

[Thông tin bài tập cuối kỳ 1](#_Toc376522399)

[1. Cách chạy chương trình 3](#_Toc376522400)

[2. Chức năng chính : 3](#_Toc376522401)

[2.1 Lấy comment từ các trang sau 3](#_Toc376522402)

[2.2 Manual add. 4](#_Toc376522403)

[2.3 Detect : chức năng phân loại 5](#_Toc376522404)

[2.4 Các tùy chọn định dạng dữ liệu 6](#_Toc376522405)

[a. Remove url : 6](#_Toc376522406)

[b. Remove html tag 7](#_Toc376522407)

[c. Remove special character 8](#_Toc376522408)

[d. Translate to vietnamese 9](#_Toc376522409)

[2.5 View chart: hiển thị biểu đồ thống kê kết dữ liệu trong bảng 10](#_Toc376522410)

[2.6 Add to learn data 12](#_Toc376522411)

[2.7 Chức năng quản lý 12](#_Toc376522412)

[a. Quản lý thư viện câu. 13](#_Toc376522413)

[b. Quản lý thư viện từ 15](#_Toc376522414)

[b. Abbreviation lib: quản lý thư viện từ viết tắt 16](#_Toc376522415)

[c. Effect word 16](#_Toc376522416)

[3. Thư viện sử dụng 18](#_Toc376522417)

[4. Kết quả thu được 18](#_Toc376522418)

# Cách chạy chương trình

**chạy file Main/classMain.java**

# Chức năng chính :

## Lấy comment từ các trang sau

Cách sử dụng : điền từ khóa hoạc link vào ô “keyword/url” và bấm nút “fetch data”

|  |  |  |
| --- | --- | --- |
| stt | Tên | Ví dụ - Mô tả |
| 1 | Twitter search api | Sử dụng từ khóa để lấy dữ liệu thông qua search api của twitter |
| 2 | Twitter stream api | Sử dụng từ khóa để lấy dữ liệu thông qua stream api của twitter |
| 3 | VnExpress URL | +Lấy comment của trang báo điện tử vnexpress  + Dữ liệu được kéo về dưới dạng json với link có dạng sau :  <http://interactions.vnexpress.net/index/get?userid=01&objecttype=1&siteid=1000000&limit=10000&objectid>=  Dữ liệu đầu vào: link bài viết vnexpress  Vd: http://vnexpress.net/tin-tuc/thoi-su/xe-cho-bia-bi-lat-dan-giup-tai-xe-thu-gom-2929182.html |
| 4 | VnExpress RSS | +Lấy comment của trang báo điện tử vnexpress.net  + Dữ liệu được kéo về dưới dạng json , xml  Dữ liệu đầu vào: link rss:  Vd: <http://vnexpress.net/rss/the-gioi.rss> |
| 5 | Ngoisao.net url | +Lấy comment của trang báo điện tử ngoisao.net  + Dữ liệu được kéo về dưới dạng json  + Dữ liệu đầu vào: link bài viết ngoisao.net  Vd: <http://ngoisao.net/tin-tuc/hau-truong/chau-a/10-my-nhan-chau-a-duoc-yeu-men-nam-2013-2929052.html> |
| 6 | Ngoisao.net rss | +Lấy comment của trang báo điện tử ngoisao.net  + Dữ liệu được kéo về dưới dạng json , xml  Dữ liệu đầu vào: link rss:  Vd: <http://ngoisao.net/rss/the-thao.rss> |
| 7 | News.zing.vn url | +Lấy comment của trang báo điện tử news.zing.vn ,sử dụng facebook api  + Dữ liệu được kéo về dưới dạng json  Dữ liệu đầu vào: link bài viết  Vd: <http://news.zing.vn/Dan-truoc-2-ban-U19-Viet-Nam-thua-nguoc-HAGL-46-post380744.html#home_featured.noibat> |
| 8 | News.zing.vn rss | +Lấy comment của trang báo điện tử news.zing.vn sử dụng facebook api  + Dữ liệu được kéo về dưới dạng json ,xml  Dữ liệu đầu vào: link rss  Vd: http://news.zing.vn/rss/tin-moi.rss |
| 9 | Haivl.com | +Lấy comment của trang giải trí haivl.com sử dụng facebook api  + Dữ liệu được kéo về dưới dạng json  Dữ liệu đầu vào: link bài viết  Vd: http://www.haivl.com/photo/1910809 |
| 10 | Ringring.vn url | +Lấy comment của trang báo điện tử news.ringring.vn  + Dữ liệu được kéo về dưới dạng json  Dữ liệu đầu vào: link bài viết  Vd: http://news.ringring.vn/sao/bao-anh-toi-khong-phai-gai-ngoan/42660.html?ref=home |
| 11 | Ringring.vn rss | +Lấy comment của trang báo điện tử news.zing.vn  + Dữ liệu được kéo về dưới dạng json ,xml  Dữ liệu đầu vào: link rss  Vd: <http://m.news.ringring.vn/rss/cong-nghe/1> |

## Manual add.

Ngoài chức năng lấy dữ liệu trên internet có thể thêm dữ liệu vào bẳng bằng tay:

* Bấm nút “Manual add” -> thêm dữ liệu vào dòng mới xuất hiện trên bảng.



## 2.3 Detect : chức năng phân loại

* Sau khi thêm dữ liệu vào bảng , bấm nút “Detect”
* Kết quả được hiển thị ở cột “classification”: Giá trị >0 negative , <0 positive , = 0 neutural.



## 2.4 Các tùy chọn định dạng dữ liệu

### a. Remove url :

-Loại bỏ các thông tin có dạng url như sau : \*\*://...../&...?

Vd: <http://google.com> ,ftp://182.162.51.1:21...

+ Thực hiện bằng cách tick vào mục **remove url**



### b. Remove html tag

-Loại bỏ các thẻ html hay các đoạn text có dạng : **<\*\*\*> và </\*\*\*>**

Vd: <br>,<span>,....

-Thực hiện bằng cách tick vào mục **remove html tag**

****

### c. Remove special character

Loại bỏ các ký tự đặc biệt trừ các ký tự dấu câu như . và ,

****

### d. Translate to vietnamese

Dịch dữ liệu lấy về sang tiếng việt

Thực hiện bằng cách tick vào mục translate to vietnamese -> bấm fetch data , mọi dữ liệu hiển thị trong bảng sẽ được dịch sang tiếng việt.

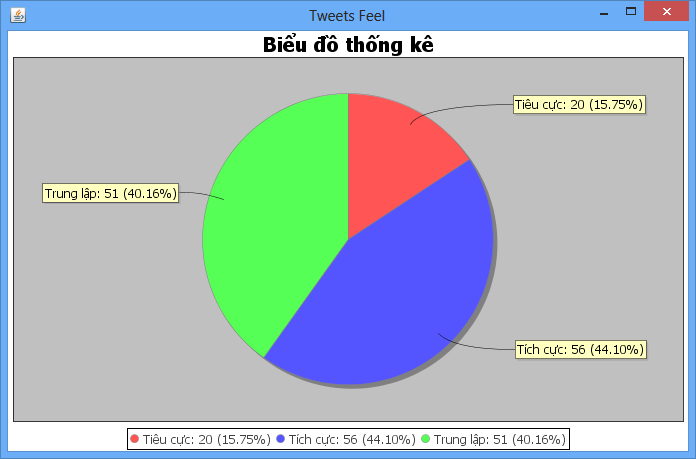


## 2.5 View chart: hiển thị biểu đồ thống kê kết dữ liệu trong bảng

Thực hiện bằng cách bấm nút: View chart



Kết quả thu được là biểu đồ thống kê các kết quả sau khi sử dụng chức năng “Detect”



## 2.6 Add to learn data

thêm vào dữ liệu học máy của chương trình, các dữ liệu trong bảng có giá trị phân loại khác 0 sẽ được thêm vào dữ liệu. Dữ liệu được lưu vào file “learnData.lib”



## 2.7 Chức năng quản lý

Vào mục quản lý bằng cách bấm nút **View library**

****

### a. Quản lý thư viện câu.

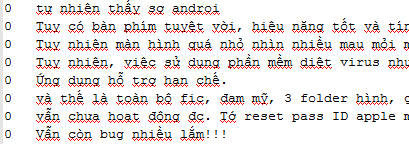
Dữ liệu quản lý gôm có nhưng câu đã được phân loại



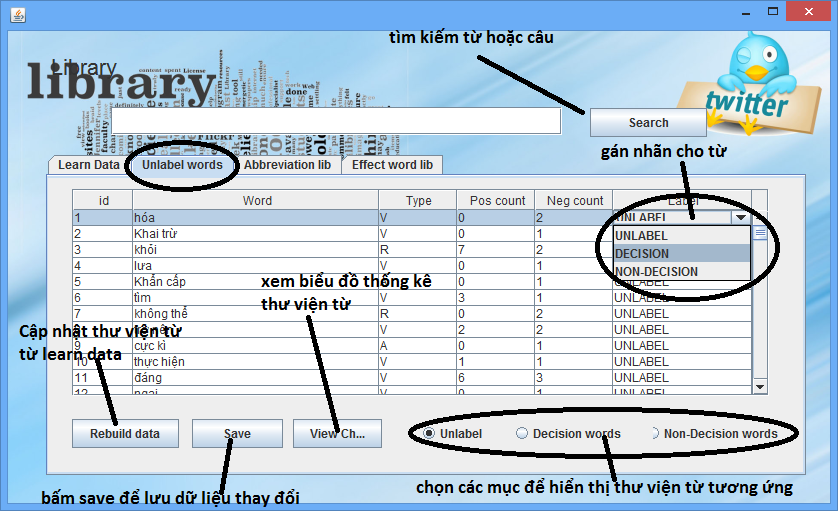
- Xóa bằng cách để trống mục status và bấm save.

- Thêm dữ liệu bằng cách sử dụng chức năng **add to learn** tại giao diện chính.

- Import dữ liệu với file text định dạng mỗi dòng như ví dụ sau:



### b. Quản lý thư viện từ



* Type: V: động từ A: tính từ R: Trạng từ
* Pos count: số lần xuất hiện trong câu positive
* Neg count: số lần xuất hiện trong câu negative
* Label: loại từ. Do dữ liệu được tách từ sử dụng thư viện ko đạt hiệu quả cao, yêu cầu người dùng phân loại bằng tay 1 lần nữa .

Có 3 loại label :

+Unlabel: dữ liệu chưa được phân loại ,

+Decision word: Từ có tác động xác định loại của câu khi phân loại (Dữ liệu được quan tâm tới)

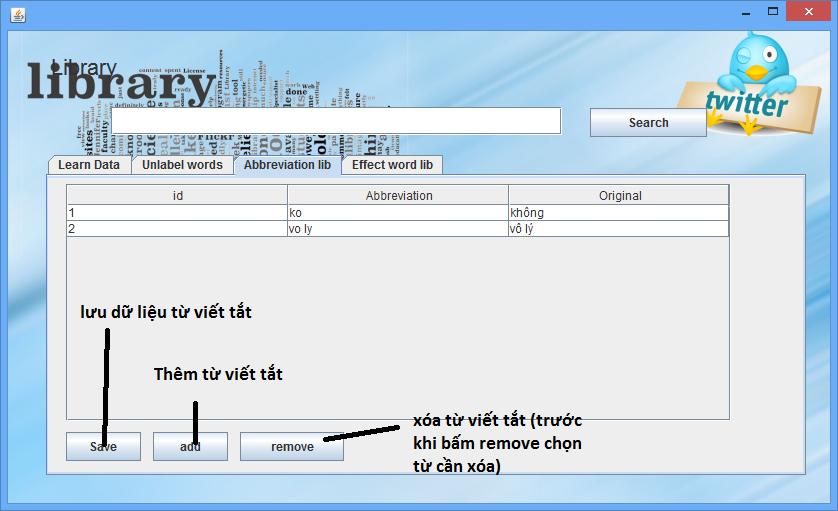
+ Non – Decision word : ngược lại vs decision word (Dữ liệu này không được quan tâm tới khi phân loại , nhưng được sử dụng trong việc rebuild word lib).

Cách phân loại : Chọn mục label sẽ có 1 combo box hiện ra.

* Rebuild data: dữ liệu sẽ được rebuild , cập nhật danh sách unlabel – dữ liệu decision và non decision được giữ nguyên.
* Save: sau khi thay đổi dữ liệu cần bấm nút save.
* Viewchart: xem biểu đồ thống kê thư viện từ.

### b. Abbreviation lib: quản lý thư viện từ viết tắt

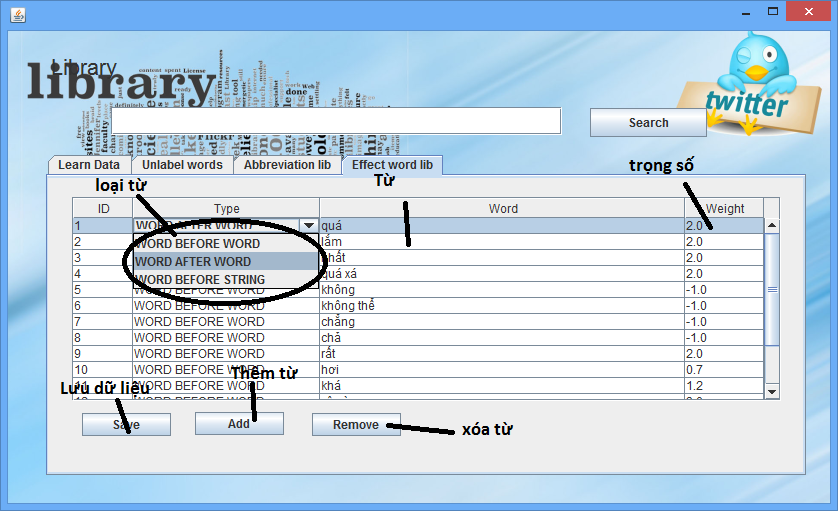
* Abbreviation: từ viết tắt
* Original : từ nguyên gốc.
* Tác dụng: Sau khi dữ liệu được kéo về từ internet , các từ viết tắt sẽ được thay thế bằng các từ gốc.



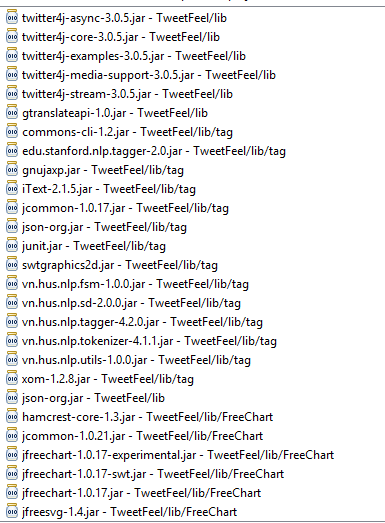
### c. Effect word

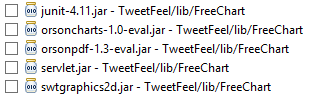
Để tăng độ chính xác khi tính điểm cho câu nên sử dụng một danh sách các từ mang hệ số với 3 loại từ:

* Word after word: Từ tác động vào từ trước nó
* Word before word: Từ tác động vào từ sau nó
* Word before string : từ tác động vào cả câu.



# Thư viện sử dụng





# Kết quả thu được

* Đã có thể phân loại được câu.
* Quản lý các thư viện từ như đã mô tả ở mục 3.
* Lấy các dữ liệu ý kiến người dùng và lưu vào file (các file được write dưới dạng object)
* Hiển thị thống kê dưới dạng biểu đồ .