# Thông tin bài tập cuối kỳ

Tên sinh viên: Nguyễn Đức Kiên

Lớp : D10CNPM1

MSV: 1021040365



# Chạy chương trình bằng cách :

**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.

## + 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.

## + Remove url :

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

## + Remove html tag :

-Loại bỏ các thông tin có dạng <\*\*\*> và </\*\*\*>

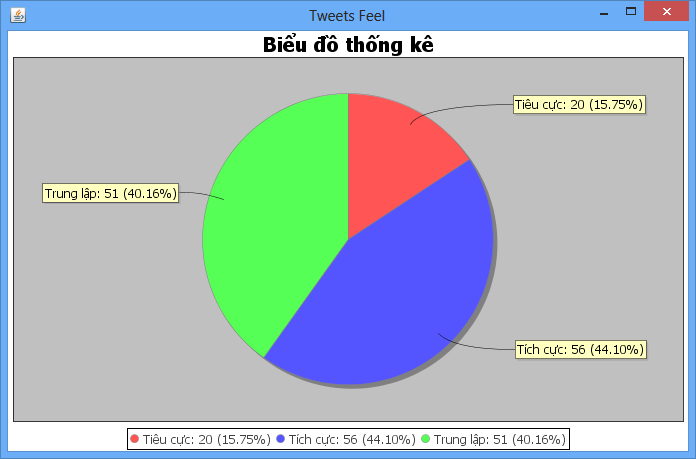
## + Remove special character

loại bỏ các ký tự đặc biệt trong bảng ascii

## + Translate to vietnamese

Dịch dữ liệu lấy về sang tiếng việt

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



## + 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

## + Chức năng quản lý :

## +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.

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

“0 Xấu hơn của Samsung”

## + 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.

## + Abbreviation lib: quản lý thư viện từ viết tắt

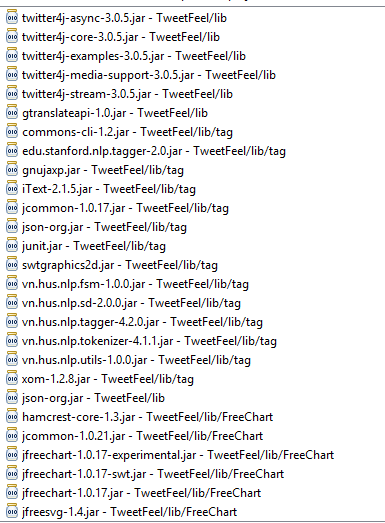
* Abbreviation: từ viết tắt
* Original : từ nguyên gốc.
* Tác dụng: Sau khi dữ liệu được kéo về từ internet , các từ viết tắt sẽ được thay thế bằng các từ gốc.

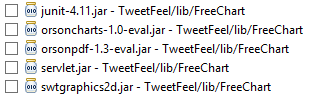
## + Effect word

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 .