**BÀI TẬP**

*Làm việc nhóm với* ***GIT***

*Xây dựng ứng dụng* ***Chuẩn hóa dữ liệu***

**Đề bài**

Bài tập này gồm 2 phần: **chương trình ứng dụng** và **dữ liệu**.

**Chương trình ứng dụng** ở dạng Windows Form, tác vụ chính sẽ là **chuẩn hóa dữ liệu văn bản**.

**Dữ liệu** sẽ là các **câu văn bản**.

Cả lớp sẽ cùng làm bài tập này.

Các em tạo tài khoản trên github.com sau đó truy cập kho chứa dự án ở địa chỉ: <https://github.com/minhtrikhtn/BaiTapFOSSK33>

Thông tin tài khoản github được lưu trữ trong file [Danh\_sach\_Tai\_khoan\_Github\_K33.xls](https://github.com/minhtrikhtn/BaiTapFOSSK33/blob/master/Danh_sach_Tai_khoan_Github_K33.xls)

Trước tiên, các em thu thập và nhập dữ liệu (văn bản) cho **các file (chủ đề).txt theo nhiều chủ đề khác nhau**.

Dữ liệu văn bản được chọn là các **câu văn tiếng Việt, câu văn tiếng Hoa được sắp theo cặp, có nghĩa tương đương với nhau**.

Dữ liệu các em lưu trữ trong thư mục: [Data](https://github.com/minhtrikhtn/BaiTapFOSSK33/tree/master/Data). Thầy đã tạo trước 2 file mẫu là [Amthuc.txt](https://github.com/minhtrikhtn/BaiTapFOSSK33/blob/master/Data/Amthuc.txt) và [Thethao.txt](https://github.com/minhtrikhtn/BaiTapFOSSK33/blob/master/Data/Thethao.txt) trong thư mục Data này, **nội dung như trong 2 file này là đúng định dạng**.

Lớp trưởng sẽ kiểm soát việc thu thập và nhập dữ liệu này. Đảm bảo các chủ đề không bị trùng nhau.

Các thành viên trong lớp sử dụng GIT để quản lý, theo dõi và xây dựng kho dữ liệu song song với việc làm ứng dụng.

Các em cài đặt phần mềm **Microsoft Visual Studio 2010 Professional** để lập trình ứng dụng. Thầy đã có tạo sẵn mã nguồn cơ bản ban đầu cho ứng dụng trong thư mục: [ChuanHoaDuLieuNN](https://github.com/minhtrikhtn/BaiTapFOSSK33/tree/master/ChuanHoaDuLieuNN). Các em có thể tải về và chạy thử rồi tiếp tục phát triển, nâng cấp chương trình lên các phiên bản cao hơn. Em nào làm thì nhớ lưu lại đóng góp của mình theo như hướng dẫn bên dưới.

Các em vào thư mục: [RoleAndContribute](https://github.com/minhtrikhtn/BaiTapFOSSK33/tree/master/RoleAndContribute) xem sẽ có file tên là **<mssv>.docx** chính là *file lưu lại những đóng góp của mình*. Các em thường xuyên cập nhật file này theo những ngày mà mình có đóng góp thay đổi trên kho project này. Đây sẽ là cơ sở tính điểm bài tập cho các em. Lưu ý: Thầy đã có tool check nên các em **cần ghi trung thực, không ghi dối** trong file này!

==================================================================

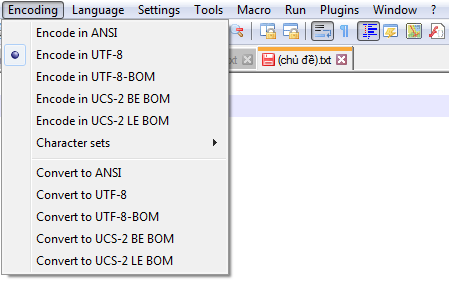
*Các em có thể nhập dữ liệu thô vào (copy / paste từ web hoặc lấy từ các nguồn khác như video, kênh chia sẻ trực tuyến, gõ vào từ các quyển sách…). Yêu cầu phải có 2 câu Hoa – Việt (1 cặp câu) có nghĩa tương đương nhau là được.*

*Sau đó là xây dựng một ứng dụng Windows Form (C#) để* ***làm sạch*** *lại theo đúng định dạng như yêu cầu bên dưới.*

*==================================================*

*Cách làm:*

* *Dùng notepad++ tạo các file txt theo chủ đề có định dạng là Unicode UTF8.*



*Ví dụ: Âm nhạc.txt, Hội họa.txt, Ẩm thực.txt, Thể thao.txt, Điện ảnh.txt…*

* *Cấu trúc giống như file* ***mẫu.txt***
* *Kết thúc câu tiếng Việt phải có dấu . ở cuối (sát vào, không có khoảng trắng. Ví dụ câu: “\** Món này không ngon .*” là sai định dạng, câu đúng định dạng phải là câu: “\** Món này không ngon.*”).*
* *Kết thúc câu tiếng Trung phải có dấu* 。*ở cuối (sát vào, không có khoảng trắng. Ví dụ câu: “+* 你点菜吧 。*” là sai định dạng, câu đúng định dạng phải là câu: “+* 你点菜吧。*”).*
* *Câu tiếng Việt luôn bắt đầu bằng dấu \**

*Ví dụ câu:*

*\** Món này không ngon.

*là đúng định dạng.*

* *Câu tiếng Trung luôn bắt đầu bằng dấu +*

*Ví dụ câu:*

+ 你点菜吧。

*là đúng định dạng.*

* *Câu tiếng Việt và câu tiếng Trung luôn đi cùng với nhau thành 1 cặp.*
* *Sau 1 cặp câu tiếng Việt – tiếng Trung thì đến 1 dòng trống. Sau đó tiếp tục đến 1 cặp khác 🡪 không để thứ tự bị ngược lại.*

*Ví dụ:*

*+*这是菜单。

*\** Đây là menu.

*là sai định dạng! Đúng phải là:*

*\** Đây là menu.

*+*这是菜单。

* *Mỗi dòng tiếng Việt chỉ có 1 câu, mỗi dòng tiếng Trung chỉ có 1 câu. Nếu có nhiều câu trên một dòng 🡪 tách ra thành các cặp khác.*

*Ví dụ:*

\* Đây là menu. Mời quý khách chọn món.

+ 这是菜单。你点菜吧。

*Được tách thành 2 cặp là:*

\* Đây là menu.

+ 这是菜单。

\* Mời quý khách chọn món.

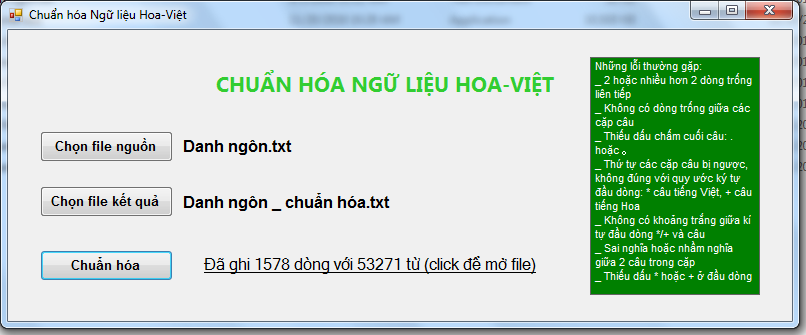
+ 你点菜吧。

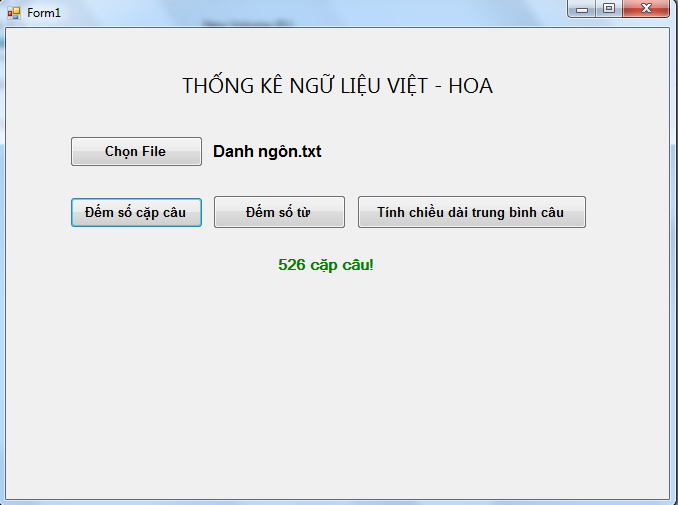
***Mục tiêu cần đạt được của bài tập:***

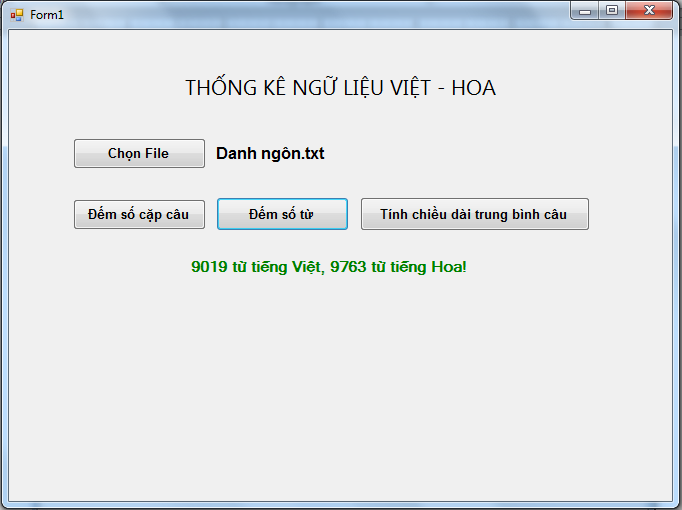
1. ***Về dữ liệu***

* *Lớp phấn đấu sau khi kết thúc bài tập này của môn học chúng ta sẽ có được 500000 – 1000000 cặp câu trong thư mục* [Data](https://github.com/minhtrikhtn/BaiTapFOSSK33/tree/master/Data)*. Nếu đạt mốc 1 triệu cặp câu, cả lớp sẽ được 10đ bài tập này.*
* *Nếu* ***có một em nào một mình*** *thu thập được 100000 cặp câu (200000 câu) trở lên với ít sai sót sẽ được chấm* ***9.5 đ bài tập*** *ngay lập tức.*
* *Dữ liệu được tách ra nhiều file txt khác nhau. Mỗi file txt đều phải đúng định dạng như đã quy ước (xem file mẫu.txt).*
* *Thầy đã có đủ tool để kiểm tra, nếu có xảy ra quá nhiều lỗi trên file dữ liệu txt thì phần mềm sẽ đánh dấu, và từ đó sẽ tìm ra ai là chủ nhân của những sai sót này. Nên các em cũng cần kiểm tra cho kỹ các dữ liệu nhé!*

1. ***Về chương trình ứng dụng***

**

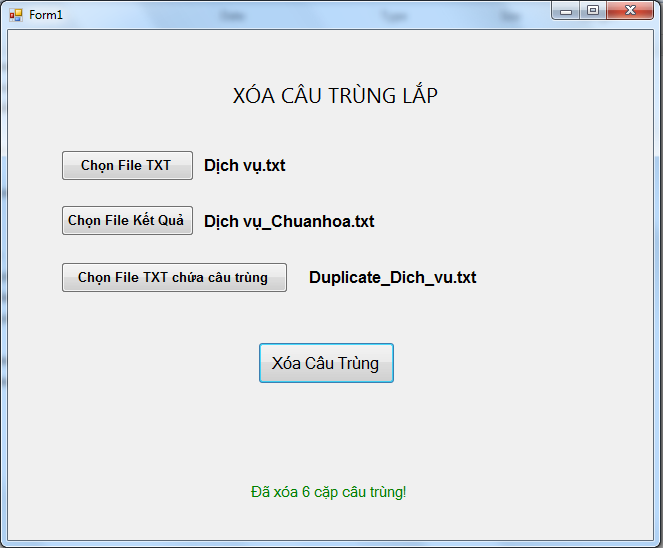
**

**

* *Nếu không làm được ứng dụng chuẩn hóa dữ liệu thì các em cứ làm sạch bằng tay, copy / paste / delete thủ công các lỗi cũng được.*
* *Lập trình chương trình cho phép làm sạch bộ dữ liệu đã thu thập bằng cách xử lý những chỗ sai định dạng và các lỗi phổ biến thường gặp như sau:*

|  |  |
| --- | --- |
| **Lỗi** | **Ví dụ** |
| *Có 2 hoặc nhiều hơn 2 dòng trống liên tiếp sau cặp câu* | \* Tôi sẽ gọi lại cho bạn sau 8:30.  + 那我晚上8点半以后再打给你吧。  \* Được, nói chuyện sau nhé.  + 好的，待会儿聊。 |
| *Không có dòng trống giữa các cặp câu* | \* Tôi sẽ gọi lại cho bạn sau 8:30.  + 那我晚上8点半以后再打给你吧。  \* Được, nói chuyện sau nhé.  + 好的，待会儿聊。 |
| *Thiếu dấu chấm cuối câu: . hoặc 。* | \* Vâng, ở đây thật sự rất bận rộn  hoặc  + 是的，今天真忙 |
| *Thứ tự các cặp câu bị ngược, không đúng với quy ước ký tự đầu dòng.* | *Ví dụ cặp câu:*  \* 你点菜吧。  + Mời quý khách chọn món.  *là sai định dạng quy ước kí tự đầu dòng! Đúng ra phải là:*  \* Mời quý khách chọn món.  + 你点菜吧。 |
| *Không có khoảng trắng giữa kí tự đầu dòng (\* hoặc +) và câu* | \*Mời quý khách chọn món.  *hoặc*  +你点菜吧。 |
| *Sai nghĩa hoặc nhầm nghĩa giữa 2 câu trong cặp 🡪 không tương ứng 1-1, câu tiếng Việt và câu tiếng Hoa cùng 1 cặp nhưng khác nghĩa!* | \* Chúng ta chạy đua thử xem ai chạy nhanh hơn nhé?  + 去吧，此后多行善事不可做恶。  *Nghĩa của 2 câu này hoàn toàn không có liên quan gì nhau* |
| *Thiếu dấu \* hoặc + ở đầu dòng* | Vâng, ở đây thật sự rất bận rộn.  是的，今天真忙。 |
| *Trong cùng 1 câu bị trùng ý, trùng vế 🡪 bỏ trùng* | \* Xin lỗi bạn, xin lỗi bạn, tôi không cố ý.  + 对不起你，对不起你，我不故意的。  *Được tách thành:*  \* Xin lỗi bạn, tôi không cố ý.  + 对不起你，我不故意的。 |
| *Có câu tiếng Việt nhưng thiếu câu tiếng Trung tương ứng và ngược lại.* | \* Tôi sẽ gọi lại cho bạn sau 8:30.  \* Được, nói chuyện sau nhé.  + 好的，待会儿聊。  + 拜拜。 |
| *Câu kết thúc bằng dấu ) hay ’ hay ” thì cũng phải có dấu . ở cuối.* | \* Căn phòng này được gọi là “Uyên Ương Sảnh”.  + 这个厅堂就叫做“鸳鸯厅”。 |
| *Câu tiếng Trung kết thúc bằng dấu gì thì câu tiếng Việt kết thúc bằng dấu đó (tương quan 1-1). Ví dụ: các dấu có thể kết câu là ! ; : .* | \* Mẹ, mấy món đồ nội thất này thật tốt, thật là đẹp!  + 妈妈，这里的家具好棒，真漂亮啊！ |
| *Câu kết thúc bằng dấu phẩy , thì bỏ dấu phẩy , thay bằng dấu chấm .* | \* Dì ơi, dì hiểu biết thật là nhiều,  + 阿姨，您懂得真多啊，  *Được đổi thành:*  \* Dì ơi, dì hiểu biết thật là nhiều.  + 阿姨，您懂得真多啊。 |
| *Câu kết thúc bằng | / \ { } [ ] @ # $ % ^ & - \_ thì bỏ hết và thay bằng dấu chấm . (ở giữa câu mà các dấu trên xuất hiện mà có nghĩa thì vẫn giữ lại nhé, như địa chỉ nhà: Số 23/4/156 Đường A… với dấu / là có nghĩa)* |  |
| *Dấu () trong tiếng Trung là （）hoặc【 】* |  |
| *Khoảng trắng quá nhiều trong câu tiếng Việt / quá nhiều giữa các từ 🡪 bỏ bớt, giữa các từ tiếng Việt chỉ có 1 khoảng trắng.* | \* Dì ơi, tại sao phòng để tiếp khách nam phải đẹp hơn vậy ?  *Được đổi thành:*  \* Dì ơi, tại sao phòng để tiếp khách nam phải đẹp hơn vậy? |
| *Từ tiếng Việt bị dính nhau.* | *Ngàymai bạnd ự định đi đ âuthế? 🡪 cái này các em chỉnh bằng tay nhé!* |
|  |  |
|  |  |
|  |  |

* *Em nào* ***xóa được tất cả những cặp câu trùng nhau*** *trong các file txt kết quả sẽ được cộng thêm nhiều điểm.*



* *Lớp trưởng quản lý, cho các bạn* ***lựa chọn chủ đề****, đảm bảo* ***không bị trùng lắp*** *chủ đề nhé! Nếu làm chung một chủ đề thì chỉ cần chỉnh sửa trên một file txt của chủ đề đó thôi*
* *Cả lớp có thời gian làm bài tập này nhiều nhất tối đa có thể, đến trước khi Thầy tổng kết điểm môn này nhé! (có thể đến sau cả thi học kỳ này luôn!)*

Mẫu trình bày kết quả cuối trong file **<mssv>.docx**:

**Kết quả sau cùng (đã xử lý làm sạch) là … <số lượng> … file txt theo chủ đề: …**

*Ví dụ:* **Kết quả sau cùng (đã xử lý làm sạch) là 4 file txt theo chủ đề: Cưới hỏi, Nhà hàng, Công ty, Giao thông.**

* Đã thu thập được … cặp câu Việt – Trung chứa trong các file txt là: ….
* Đã áp dụng kỹ thuật làm sạch dữ liệu để xử lý các lỗi định dạng như là:

+ Lỗi 1…

+ Lỗi 2…

…

* (Cộng điểm thêm, nếu không làm được thì khỏi ghi phần này) Đã lập trình một chương trình đơn giản giúp xử lý các lỗi định dạng
  + Chương trình chạy theo nguyên tắc … (ví dụ: đọc file dữ liệu txt nguồn/thô 🡪 nhấn nút 🡪 xử lý ngoại lệ và ghi ra file kết quả, chẳng hạn!)