**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**Viện Công nghệ thông tin và Truyền thông**

------



**BÁO CÁO PROJECT 1**

Đề tài: Đọc và thống kê các từ trong văn bản

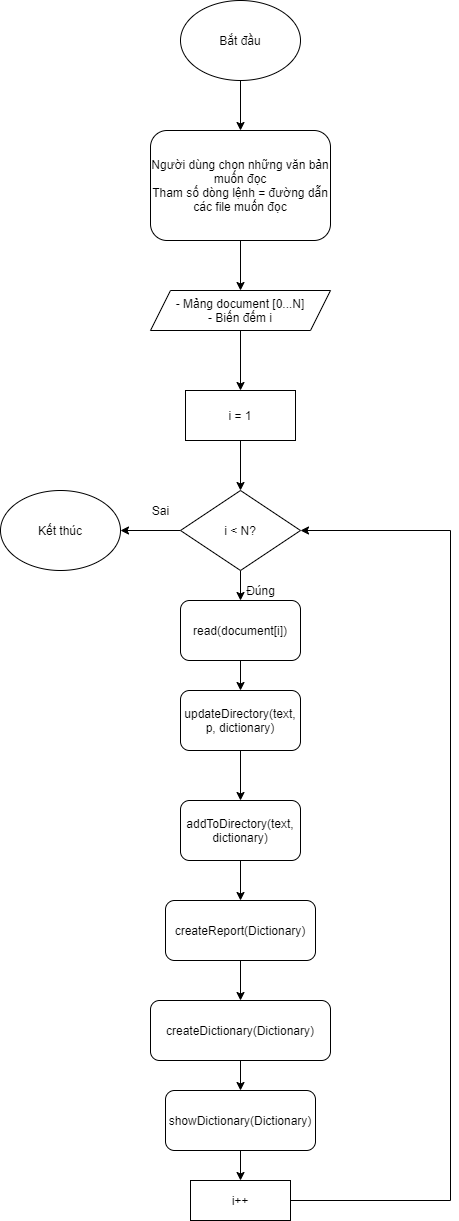
**Giảng viên hướng dẫn**: **TS. Đặng Văn Chuyết**

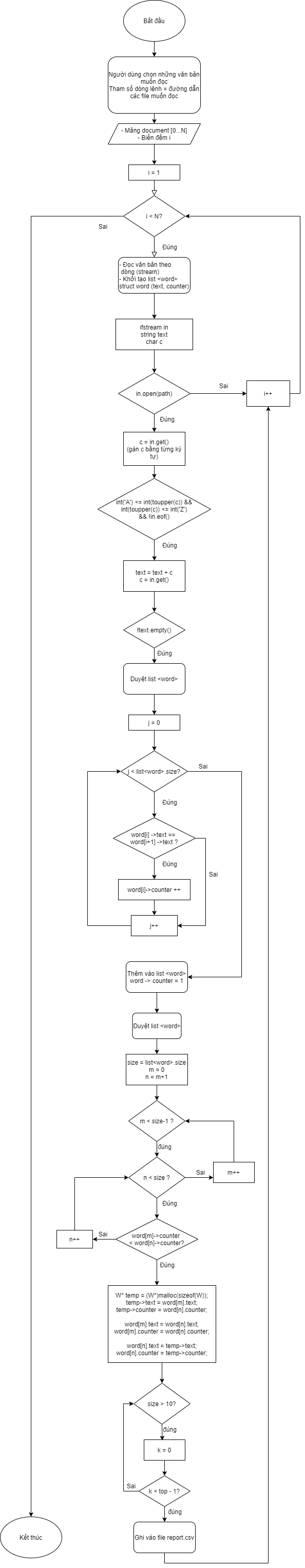
**Sinh Viên Thực Hiện** : **Công Đào Minh Uyên – 20176063**

*Hà Nội, ngày 5 tháng 1 năm 2021*

1. **Lưu đồ thuật toán**

Theo hàm số





1. **Mô tả hằng số, biến số, các thủ tục** 
   1. **Hằng số**
   2. **Biến số**

struct Word {

string text;

int counter;

};

struct W {

string\* text;

int counter;

};

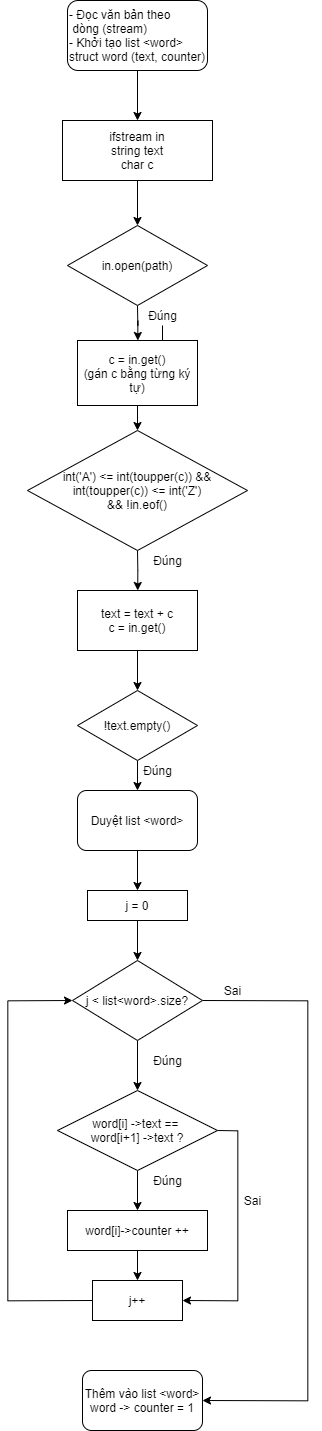
Cấu trúc của 1 từ

1. **Thủ tục và hàm**

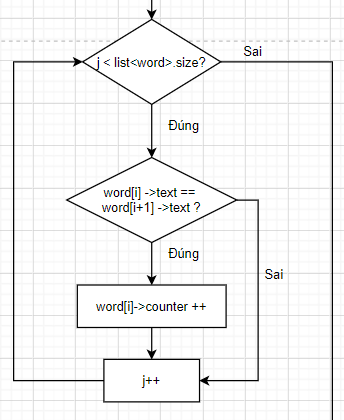
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên hàm | Tham số | Kiểu trả về | Mô tả chức năng |
| 1 | read | string path, list<Word> dictionary | void | Đọc văn bản |
| 2 | addToDirectory | string text, list<Word> dictionary | list<Word> | Thêm vào danh sách các từ |
| 3 | updateDirectory | string text, bool\* status, list<Word> dictionary | bool\* | -So sánh mỗi từ với danh sách từ  -Nếu từ có trong danh sách, tăng bộ đếm counter của từ lên 1, trả về true  - Nếu từ không có trong danh sách, trả về false |
| 4 | createDictionary | list<Word> dictionary | Void | Tạo và ghi vào file Dictionary.csv (chứa danh sách các từ và tần suất xuất hiện của nó) |
| 5 | createReport | list<Word> dictionary | void | Tạo và ghi vào file Report.csv (chứa 10 từ có tần suất xuất hiện cao nhất) |

**Lưu đồ thuật toán tương ứng với các hàm**

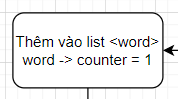
* **Read**

****

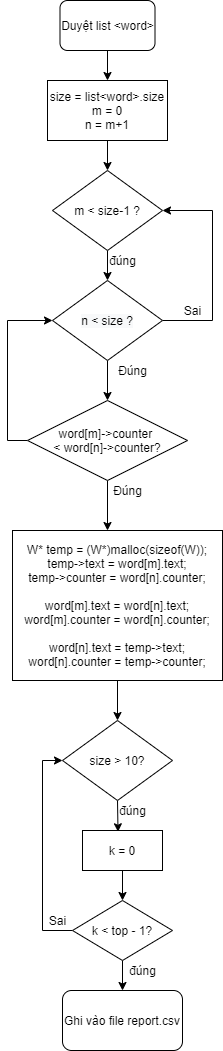
* **updateDirectory**



* **addToDirectory**

****

* **createReport**

****

1. **Mô tả môi trường cài đặt chương trình**

- Chương trình được cài đặt bằng ngôn ngữ C với sự hỗ trợ của IDE Visual studio 2019

- Sử dụng một số thư viện hỗ trợ đọc, ghi file, thao tác với tiếng Việt

+ fstream

+ codecvt

+ io.h

1. **Tích hợp chương trình và quá trình hoàn thành chương trình**

- Thư viện fstream, codecvt có thể đọc file tiếng việt. Tuy nhiên trong khi xử lý các từ, các hàm không hỗ trợ xử lý các chữ cái tiếng việt có dấu nên đôi khi sẽ có lỗi một số chữ cái có dấu sẽ bị chuyển thành dấu ?

1. **Mô tả về thử nghiệm và đánh giá nó**

- Chương trình chạy trên màn hình console sử dụng tham số dòng lệnh, tuy nhiên hơi không thuận tiện khi người dùng phải nhập chính xác đường dẫn của file văn bản

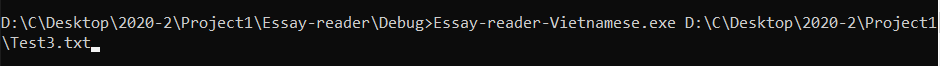
1. **Đóng gói chương trình và hướng dẫn sử dụng**

- Chương trình thực thi là file .exe nằm trong thư mục Debug.

- Để sử dụng chương trình ta khởi chạy file

- Sử dụng tham số dòng lệnh, truyền vào đường dẫn cho các file muốn đọc.

Mỗi file văn bản là 1 tham số dòng lệnh, cách nhau bằng dấu cách

VD : 

* Kết quả được tạo ra sẽ được ghi vào file csv

+ Dictionary.csv : file ghi danh sách các từ và tần suất xuất hiện các từ

+ Report.csv : file ghi top 10 từ có tần suất xuất hiện nhiều nhất