**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

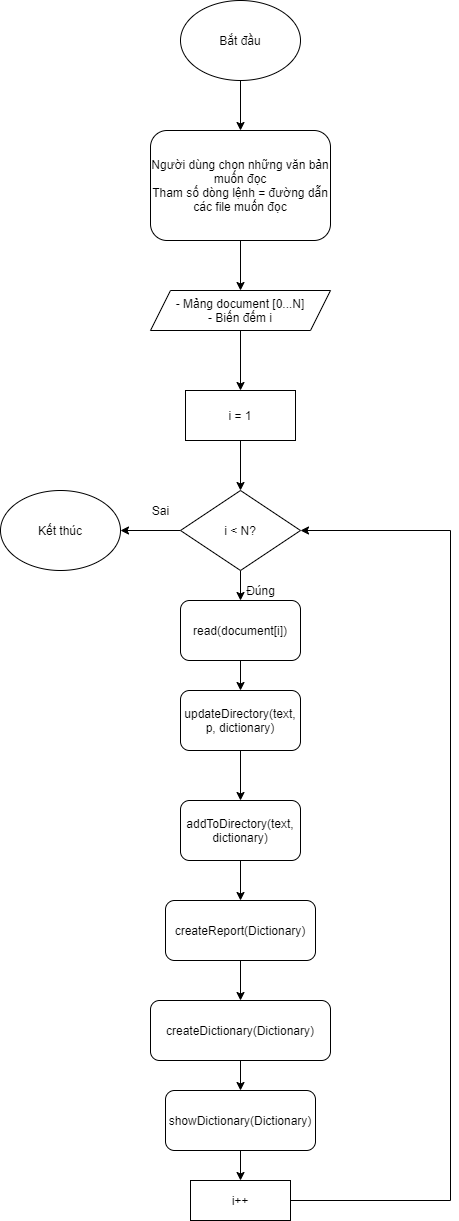
**Giảng viên hướng dẫn**: **TS. Đặng Văn Chuyết**

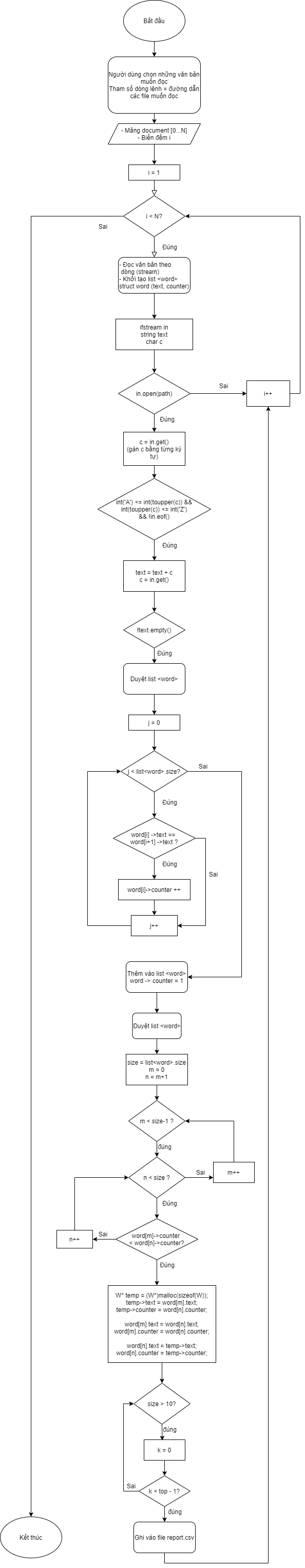
**Sinh Viên Thực Hiện** : **Công Đào Minh Uyên – 20176063**

*Hà Nội, ngày 19 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;

};

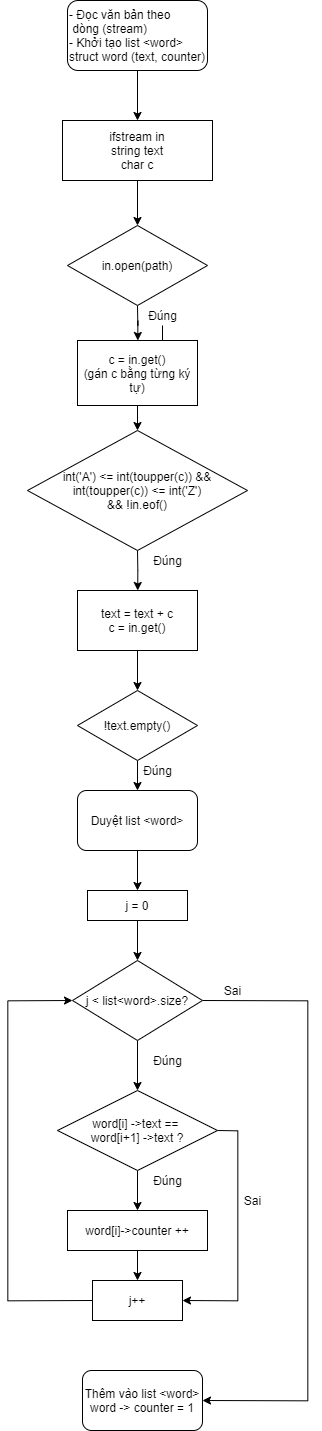
Cấu trúc của 1 từ

1. **Thủ tục và hàm**

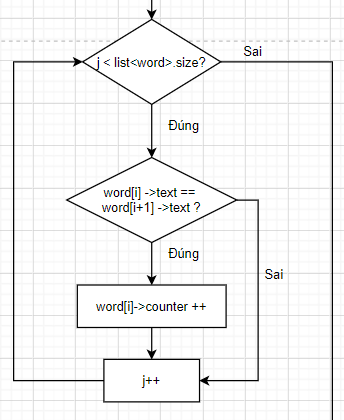
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên hàm | Tham số | Kiểu trả về | Mô tả chức năng |
| 1 | read | string path, list<Word> dictionary | void | Đọc văn bản |
| 2 | addToDirectory | string text, list<Word> dictionary | list<Word> | Thêm vào danh sách các từ |
| 3 | updateDirectory | 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**

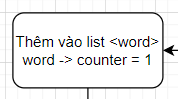
* **Read**

****

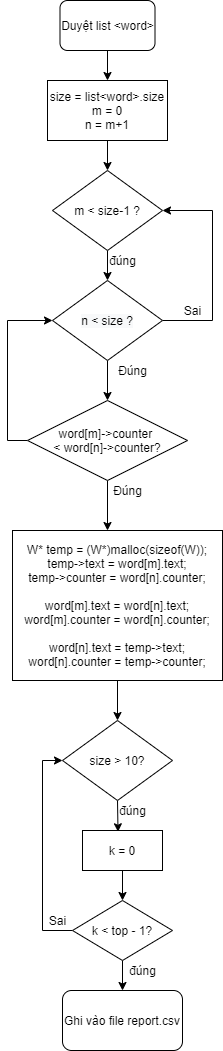
* **updateDirectory**



* **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

+ fstream

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

-

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

- Chương trình chạy và liệt kê các từ khác nhau trong văn bản

- Tuy nhiên nếu file có các ký tự đặc biệt dạng UTF-8 hay UTF-16 thì chương trình sẽ liệt kê thiếu từ

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

- Người dùng nhập vào đường dẫn đến file văn bản cần đọc. Để đọc nhiều văn bản, các file ngăn cách bằng dấu cách

* 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