Diagram

Description automatically generated

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH

KHOA TOÁN

✻ ✼ ✽ ✾ ✿ ❀ ❁ ❂ ❃ ❇ ❈ ❉ ❊ ❋



Bài Tập Lớn Môn Phát triển phần mềm hướng đối tượng

Đề tài: Thiết kế và cài đặt một chương trình cho phép phân tích một tệp tin văn bản, xuất ra tần số xuất hiện của các từ có trong tệp tin văn bản đó

Icon

Description automatically generated with medium confidence

Tên thành viên:

Bùi Minh Phương 18110186

Nguyễn Tiến Thông 18110229

Giáo viên hướng dẫn:

Nguyễn Ngọc Long

Diagram

Description automatically generated

Mục lục

[**Lời nói đầu** 3](#_Toc92587880)

[**I.** **Mô tả chương trình:** 4](#_Toc92587881)

[**1.** **Giới thiệu về chương trình:** 4](#_Toc92587882)

[**2.** **Thông tin chi tiết:** 4](#_Toc92587883)

[**II.** **Phân tích và thiết kế chương trình:** 4](#_Toc92587884)

[**1.** **Phân tích chức năng:** 4](#_Toc92587885)

[**2.** **Chức năng mở rộng:** 4](#_Toc92587886)

[**3.** **Sơ đồ UML:** 4](#_Toc92587887)

[**4.** **Phân tích chương trình:** 6](#_Toc92587888)

[**5.** **Lợi thế và hạn chế của chương trình:** 6](#_Toc92587889)

[**6.** **Giao diện minh họa:** 7](#_Toc92587890)

[**III.** **Phân công công việc:** 8](#_Toc92587891)

# **Lời nói đầu**

Hiện nay, ngành công nghệ thông tin ở nước ta càng ngày càng phát triển, nhiều nhân tài ở ngành này đã sáng tạo và phát triển nên những phần mềm, những chương trình có ích cho xã hội. Tuy nhiên, không phải ai cũng hoàn hảo, bên cạnh việc giỏi về công nghệ, nguồn nhân lực về công nghệ thông tin ở nước ta vẫn còn rất nhiều khuyết điểm như trình độ tiếng anh. Chính vì lẽ đó, nhóm tụi em đã chọn đề bài trên cho bài tập lớn môn “Phát triển phần mềm hướng đối tượng” với yêu cầu thiết kế một chương trình có khả năng phân tích tần số của từ tiếng anh trong một tệp tin cho trước.

Chương trình giúp chúng ta thống kê được số lượng từ ngữ đã sử dụng trong tệp tin, tần số xuất hiện của chúng, từ đó có thể phục vụ cho từng mục đích sử dụng khác nhau. Ví dụ, biết được tần số từ trong một đoạn văn có thể giúp ta tránh được lỗi lặp từ, lọc ra những từ giống nhau trong một đoạn văn để người dùng có thể dễ dàng tổng hợp được các từ vựng đã được sử dụng. Không những thế, chương trình có thể loại bỏ đi các từ sai chính tả trong một đoạn văn để người dùng không cần mất thời gian để tra từ điển xem đó có là từ có nghĩa không.

Sau một thời gian tìm hiểu và thực hiện, cuối cùng, nhóm chúng em cũng đã có thể hoàn thành được bài tập và cho ra được một chương trình hoàn chỉnh. Nội dung về chương trình và các phân tích và thiết kế chi tiết sẽ được ghi lại toàn bộ trong bài báo cáo tổng kết này.

Qua bài tập, thứ chúng em nhận được không chỉ dừng lại ở việc tạo ra được một sản phẩm có ý nghĩa, mà còn được ứng dụng những gì đã được học ở bộ môn “Phát triển phần mềm hướng đối tượng” và những bộ môn khác. Nhóm chúng em xin gửi lời cảm ơn chân thành đến thầy Nguyễn Ngọc Long, người đã giảng dạy chúng em để nhóm có thể hoàn thành bài tập này một cách tốt nhất. Tuy nhiên, do thời gian và năng lực có hạn nên chương trình của nhóm em còn gặp nhiều thiếu sót, mong thầy thông cảm và đóng góp thêm nhiều ý kiến để không chỉ là bài tập này mà những dự án hay những đồ án lớn trong tương lai của nhóm chúng em có thể thành công tốt đẹp.

Xin chân thành cảm ơn.

1. **Mô tả chương trình:**
2. **Giới thiệu về chương trình:**

Chương trình có tên là “wfreq” (stands for word frequency program) có nghĩa là “Chương trình tần số từ” với các yêu cầu cơ bản như sau:

* Phân tích một tập tin văn bản. xuất ra tần số xuất hiện của các từ có trong tập tin văn bản.
* Các ký tự ngăn cách bao gồm khoảng trắng chuẩn (dấu ‘,’, ‘.’, ‘;’, ‘:’, ‘?’, ‘!’, ‘...’) và các ký tự ngăn cách dấu câu.
* Từ là từ có nghĩa trong từ điển từ tiếng Anh, có xét đến số ít, số nhiều, quá khứ (ví dụ: mouse, mice được tính là cùng một từ, study, studies, studied, studying được tính là cùng một từ).

1. **Thông tin chi tiết:**

* Đầu vào của chương trình là tên của một file dữ liệu có định dạng \*.txt, đầu ra của chương trình là một bảng báo cáo tần số ở dạng console.
* Nếu như trong chương trình không có tệp văn bản được đưa vào, chương trình sẽ lấy văn bản từ stdin (văn bản được nhập từ bàn phím).
* Khi chỉ tên đường dẫn văn bản được cung cấp, hãy in tần suất của các từ trong văn bản.
* Chương trình phân tích tần suất xuất hiện của các từ trong văn bản xong in ra báo cáo
* Chương trình sử dụng từ điển tiếng Anh được lưu dưới dạng 1 file \*.txt.
* Chương trình sử dụng từ điển không đầy đủ (từ điển chỉ chứa từ có nghĩa, không giải thích nghĩa của từ).

1. **Phân tích và thiết kế chương trình:**
2. **Phân tích chức năng:**

* Chức năng đọc một đoạn văn bản được nhập từ bàn phím xuất ra tần số từ trong văn bản.
* Chức năng đọc một tệp văn bản có sẵn xuất ra tần số từ trong văn bản.

1. **Chức năng mở rộng:**

* Tính tổng số từ tiếng anh có nghĩa trong đoạn văn bản.
* Xem phiên bản của chương trình.
* Tìm ra từ có tần số xuất hiện cao nhất, từ có tần số xuất hiện thấp nhất. (mở rộng sau).

1. **Sơ đồ UML:**

Diagram

Description automatically generated

1. **Phân tích chương trình:**

Chương trình gồm các class được định nghĩa cụ thể như sau:

1. *Word:*

* Dùng để định nghĩa một từ.
* Word là một chuỗi kí tự (String).

1. *WordList:*
2. Được xem như một bảng, mỗi record sẽ là một hàng
3. *Word Dictionary (interface):*

* Dùng để phân tích các từ trong file dictionary, xem file đó như một từ điển.
* Các phương thức trong WordDictionary:

+ setname và get name: Dùng lấy tên của từ điển.

+ find: Tìm một từ trong từ điển.

1. *WordFrequencyList:*

* Chứa danh sách các từ đã tính tần số.
* Phương thức:

+ comWFreq&comFreq: Tính tần số của từ trong danh sách từ.

+ setName và getName: phuong thức lấy tên của danh sách tần số từ

+ getTotalFrequency: lấy tổng tần số từ trong văn bản.

1. *Word Analyzer:*

* Dùng để phân tích các từ tiếng anh.
* Phương thức:

+ set và get: Đặt và lấy dấu phân cách.

+ find: Tìm một từ chứa trong từ điển.

+ addSeperators: Thêm một trình phân tách

+ checkValid: Kiểm tra xem từ đó có là từ hợp lệ hay không.

+ print: In bảng tần số từ.

1. *Record:*

* Gồm các lớp WordRecord, WordFrequencyRecord dùng để lưu lại thông tin của các từ, mỗi lớp record được xem như là một hàng của WordList. Trong đó:

+WordRecord dùng để lưu lại từ và các biến thể của từ (só ít, só nhiều, quá khứ, hiện tại tiếp diễn, v.v).

+ WordFrequencyRecord dùng để lưu lại tần số của từ tương ứng

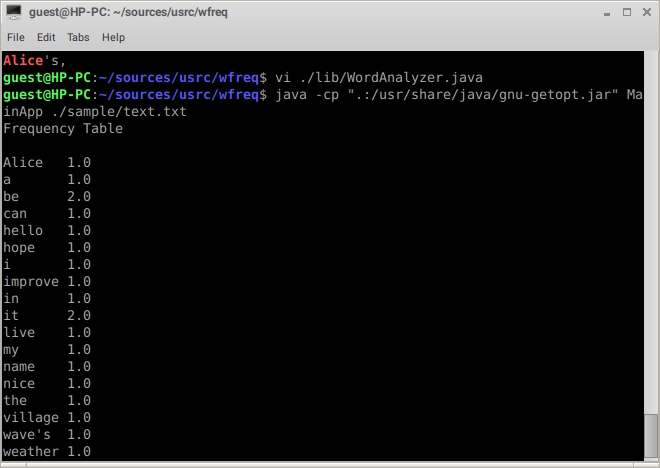
1. **Lợi thế và hạn chế của chương trình:**
2. **Lợi thế:**

* Chương trình có tính di động.
* Từ điển sử dụng loại từ điển không đầy đủ nên tối ưu được thời gian tính toán và phân tích.
* Phân tích chính xác đâu là từ có nghĩa, loại bỏ được các từ không có nghĩa và các kí tự dư thừa.
* Chương trình cho phép xác định các địa danh bằng Tiếng Anh, tên riêng.

1. **Hạn chế:**

* Điều kiện tiên quyết để chương trình hoạt động là phải cài đặt gói gnu.getopt.Getopt.
* Chương trình thiết kế dành cho người dùng quen với giao diện dòng lệnh.
* Chương trình chỉ xuất ra kết quả trên giao diện console, muốn xuất ra dưới dạng khác phải dùng chương trình bên ngoài.
* Chương trình không thể sử dụng loại từ điển khác.
* Chương trình còn nhiều hạn chế trong việc phân loại từ ngữ (nếu trong file \*.txt đầu vào chứa các từ giống nhau về cú pháp nhưng khác nhau về nghĩa, chương trình sẽ cho kết quả sai lệch).
* Chương trình sẽ hoạt động sai nếu đầu vào là một tệp nhị phân.
* Do sử dụng từ điển mặc định nên nếu từ trong văn bản không có trong từ điển, chương trình mặc định từ đó là không có nghĩa.

1. **Giao diện minh họa:**



**A screenshot of a computer

Description automatically generated with medium confidence**

1. **Phân công công việc:**

|  |  |  |
| --- | --- | --- |
|  | Thông | Phương |
| Tuần 1  (13-19/12) | * Chọn đề bài * Phân tích yêu cầu đề (Chọn công nghệ sử dụng, hướng hoạt động của chương trình). | * Phân tích yêu cầu đề (Vẽ sơ đồ UseCase cơ bản) * Bổ sung các object. |
| Tuần 2  (20-26/12) | * Code chức năng phân tách từ trong một tệp văn bản. * Sửa lỗi. | * Kiểm thử và sửa lỗi. * Tìm và chỉnh sửa loại từ điển cho phù hợp. |
| Tuần 3  (27/12-2/1) | * Sửa lại sơ đồ cho phù hợp với chương trình. * Chỉnh sửa từ điển. * Hỗ trợ code chức năng tính tần số. * Kiểm tra lại từ điển. | * Vẽ lại sơ đồ. * Định dạng biến thể từ điển * Kiểm thử và sửa lỗi. * Code chức năng tính tần số. |
| Tuần 4  (3-9/1) | * Kiểm tra lại chương trình * Sửa lỗi. * Viết mô tả chương trình bằng tiếng anh. | * Kiểm tra chương trình * Sửa lỗi. * Viết báo cáo và ppt trình chiếu. |

**VI. Nguồn tham khảo:**

https://www.storyofmathematics.com/frequency-statistic