TRƯỜNG ĐH SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

Logo

Description automatically generated

**TIỂU LUẬN CUỐI KỲ**

**Môn học: Cấu trúc dữ liệu và giải thuật**

**Xây dựng trò chơi nối từ Tiếng Việt**

Giảng viên: PGS.TS. Hoàng Văn Dũng

**Danh sách sinh viên thực hiện**

|  |  |  |
| --- | --- | --- |
| **Mã số SV** | **Họ và tên** | **Tỷ lệ**  **đóng góp** |
| 20133079 | Lã Minh Phúc | 120% |
|  |  |  |
|  |  |  |

*TP. Hồ Chí Minh, tháng .... năm .......*

# MỤC LỤC

# MỞ ĐẦU

Ngày nay với sự phát triển của internet thì mạng xã hội đã trở thành một phần trong cuộc sống mỗi chúng ta. Sử dụng mạng xã hội trong mục các mục đích như giao tiếp, chia sẻ, giải trí giúp chúng ta luôn cập nhập thông tin một cách nhanh chóng. Tuy nhiên chính vì nhu cầu trao đổi thông tin nhanh chóng nên nhiều người dùng đã sử dụng các từ vay mượn từ tiếng nước ngoài trong quá trình chia sẻ thông tin nhằm rút ngắn thời gian nhập liệu. Điều này dẫn tới việc lạm dụng từ mượn làm mất đi vẻ đẹp và sự phong phú của Tiếng Việt khi ngày càng nhiều các từ Tiếng Việt dần bị thay thế như “tạm biệt”, “xin chào”,

# CƠ SỞ LÝ THUYẾT

## Công cụ và môi trường để lập trình

### Xây dựng chương trình

* Visual Studio Community 2019, Desktop Development with C++
* Thư viện dpp (hỗ trợ kết nối với bot Discord)

### Xử lý, phân tích dữ liệu

* Python 3.8
* Anaconda, Jupyter Notebook
* Từ điển Việt Việt; Hồ Ngọc Đức

## Các phương pháp, kĩ thuật sử dụng

* Danh sách sử dụng mảng
* Danh sách sử dụng liên kết đơn
* Bảng băm
* Cây nhị phân tìm kiếm
* Tìm kiếm nhị phân

# PHÂN TÍCH, THIẾT KẾ GIẢI PHÁP

## Dữ liệu đầu vào của chương trình

### dictionary.txt

Là một tệp văn bản chứa các từ vựng có hai âm tiết của Tiếng Việt mỗi từ được ngăn cách nhau bằng dấu phẩy (kí hiệu: ,), các từ được sắp xếp tăng dần theo giá trị mã hoá từ của chương trình.

Các từ vựng trong trò chơi được xem là có nghĩa nếu tồn tại trong tệp này. Trong chương trình tệp sẽ được đọc và ghi dữ liệu vào một bảng băm.

Là kết quả của quá trình phân tích Từ điển tiếng Việt.

### start\_word.txt

Là một tệp văn bản chứa các âm để bắt đầu một từ trong Tiếng Việt. Dòng đầu tiên là số lượng phần tử các âm, dòng thứ hai là các âm để bắt đầu bằng một từ trong Tiếng Việt.

Các âm trong tệp được chương trình sử dụng làm âm ngẫu nhiên để bắt đầu trò chơi. Trong chương trình tệp sẽ được đọc và ghi vào một mảng.

Là kết quả của quá trình phân tích Từ điển tiếng Việt.

### encoder.txt

Là một tệp văn bản có cấu trúc được chương trình sử dụng để phân tách và mã hoá các âm trong Tiếng Việt.

Là kết quả của quá trình phân tích dữ liệu bằng Python.

*Cấu trúc tệp:*

* Dòng thứ nhất: Các số lượng phần tử của các dữ liệu được ngăn cách nhau bằng khoảng trắng, các dữ liệu tương ứng lần lượt là: các âm đặc biệt, các kí tự bắt đầu của một âm, các kí tự kết thúc của một âm, các kí tự có dấu trong Tiếng Việt.
* Dòng thứ hai: Các chuỗi được ngăn cách nhau bằng dấu phẩy, chứa các âm đặc biệt không thể phân tách và mã hoá theo cách thông thường của chương trình.
* Dòng thứ ba: Các chuỗi được ngăn cách nhau bằng dấu phẩy, chứa các kí tự bắt đầu của một âm trong Tiếng Việt, được sắp xếp theo mã hoá UTF8.
* Dòng thứ bốn: Các chuỗi được ngăn cách nhau bằng dấu phẩy, chứa các kí tự kết thúc của một âm trong Tiếng Việt, được sắp xếp theo mã hoá UTF8.
* Dòng thứ năm: Các chuỗi được ngăn cách nhau bằng dấu phẩy, chứa các kí tự có dấu trong Tiếng Việt.

## Các cấu trúc dữ liệu

### Sử dụng bảng băm để chứa các từ vựng (class Dictionary)

*Cấu trúc bảng băm:*

* Khoá (key): Âm bắt đầu.
* Giá trị (value): Một danh sách mảng (con trỏ) chứa các mã hoá của âm kết thúc của một từ tương ứng với giá trị băm của âm bắt đầu. Mảng được sắp xếp tăng dần theo giá trị mã hoá của âm.
* Hàm băm (hash function): Hàm mã hoá âm Tiếng Việt. Giá trị trong đoạn từ 0 đến 38877. Là hàm băm không có va chạm.

Shape

Description automatically generated with medium confidence

*Ví dụ tra cứu từ “gà” trong bảng băm*

*Đọc tệp “dictionary.txt”:*

* Lấy lần lượt các từ vựng trong tệp
* Với một một khoá tạo một danh sách liên kết chứa các mã hoá của các âm kết thúc tương ứng
* Chuyển danh sách liên kết chứa các âm kết thúc thành con trỏ danh sách mảng chứa các âm kết thúc để lưu trữ và hỗ trợ thực hiện tìm kiếm nhị phân.

*Các phương thức:*

bool checkMeaning(const std::wstring& word) // Kiểm tra tính có nghĩa của một từ có 2 âm tiết

MyArray<uint16\_t>\* getLastSounds(const int& index) // Lấy danh sách mảng các encode của âm kết thúc

### Sử dụng cây nhị phân tìm kiếm lưu trữ các từ đã sử dụng trong màn chơi (class History)

Kiểu dữ liệu lưu trữ: số nguyên không âm 32 bits (Giá trị mã hoá một từ Tiếng Việt 2 âm tiết).

*Các phương thức:*

bool isExisted(const uint32\_t& encode) // Kiểm tra từ đã tồn tại trong cây trước đó

bool add(const std::wstring& x) // Thêm giá trị encode của x vào cây

void removeAll() // Xoá tất cả các node của cây

### Danh sách liên kết đơn (class ArrayBuilder)

Được sử dụng để lưu trữ dữ liệu sau đó có thể chuyển qua danh sách mảng khi số lượng phần tử không đổi nhằm tiết kiệm bộ nhớ và thực hiện tìm kiếm nhị phân.

*Các phương thức:*

void add(T x) // Thêm một phần tử vào đầu danh sách liên kết đơn

MyArray<T>\* toMyArray() // Chuyển thành danh sách sử dụng mảng (con trỏ), danh sách mảng trả về sẽ ngược với danh sách liên kết

### Danh sách mảng (class MyArray)

*Các phương thức:*

T& operator[](const int& index) // Truy cập đến một phần tử của mảng

int getSize() // Lấy số lượng phần tử của mảng

## Thuật toán

### Tách dấu một âm trong Tiếng Việt

*Đọc dữ liệu dấu từ tệp:*

Trong tệp *encoder.txt*, phần tử cuối của dòng đầu tiên sẽ là số lượng các kí tự có thể có dấu trong Tiếng Việt giá trị này được ghi vào biến *sizeTones* trong chương trình. Dòng cuối cùng của tệp là chuỗi các kí tự có dấu trong Tiếng Việt, các dấu lần lượt là thanh, sắc, huyển, hỏi, ngã, nặng, các chuỗi này được ghi vào mảng *tones* trong chương trình.

|  |
| --- |
| encoder.txt |
| 11 28 152 12  …  aáàảãạ,eéèẻẽẹ,iíìỉĩị,oóòỏõọ,uúùủũụ,yýỳỷỹỵ,âấầẩẫậ,êếềểễệ,ôốồổỗộ,ăắằẳẵặ,ơớờởỡợ,ưứừửữự |

*Dữ liệu để đọc mảng tones của tệp encoder.txt*

Thuật toán sẽ duyệt lần lượt các kí tự của chuỗi âm, khi đến kí tự có dấu sẽ thay thế bằng kí tự không dấu và trả về giá trị tương ứng với dấu của kí tự.

/// <summary>

/// Thay một kí tự có dấu trong âm bằng kí tự không dấu và trả về giá trị của dấu.

/// </summary>

/// <param name="sound">Âm cần tách dấu, sau khi tách sẽ mất dấu</param>

/// <returns>Giá trị của dấu</returns>

int removeTone(std::wstring& sound)

{

    int lenght = sound.size();

    int tone = 0;

    for (int i = lenght - 1; i >= 0; i--)

    {

        if (sound[i] >= 224) // Giá trị bé nhất của một kí tự có dấu

        {

            for (int j = 0; j < sizeTones; j++)

            {

// Duyệt các kí tự có dấu (trừ dấu thanh)

                for (int k = 1; k < 6; k++)

                {

                    if (sound[i] == tones[j][k])

                    {

                        sound[i] = tones[j][0];

                        tone = k;

                        return tone;

                    }

                }

            }

        }

    }

    return tone;

}

Code C++ hàm tách dấu của một âm Tiếng Việt

### Mã hoá một âm trong Tiếng Việt

Phân tách một âm trong Tiếng Việt thành 3 thành phần: phần bắt đầu, phần kết thúc và phần dấu.

Sử dụng dữ liệu *unsigned short (uint16\_t)* để lưu trữ 3 thành phần này, cụ thể:

* 8 bits đầu: Phần kết thúc
* 5 bits tiếp theo: Phần bắt đầu
* 3 bits còn lại: Phần dấu

Bỏ qua phân biệt kí tự hoa, thường, vị trí đăt dấu trong âm. Ví dụ: âm “Hoá” sẽ có cùng mã hoá với âm “hóa”

Các âm đặc biệt được tính phần bắt đầu là 31, phần kết thúc là vị trí của âm trong mảng các âm đặc biệt.

A picture containing diagram

Description automatically generated

Lưu đồ khối thuật toán mã hoá một âm trong Tiếng Việt

Hình Nal: Mô tả phân tách và mã hoá âm “thầy”

### Mã hoá một từ có hai âm tiết trong Tiếng Việt

Tách một từ thành hai âm tiết sử dụng 16 bits đầu để lưu trữ từ đầu tiên và 16 bits tiếp theo để lưu trữ từ thứ hai.

Graphical user interface

Description automatically generated

Lưu đồ khối thuật toán mã hoá một từ hai âm tiết

# TÀI LIỆU THAM KHẢO

* Hồ Ngọc Đức (2004), Dự án Từ điển tiếng Việt miễn phí, <http://informatik.uni-leipzig.de/~duc/Dict/>. Dữ liệu từ điển của trò chơi

# PHỤ LỤC

## Giá trị mã hoá các thành phần của âm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Phần bắt đầu** | **Giá trị** | **Phần bắt đầu** | **Giá trị** | **Phần bắt đầu** | **Giá trị** |
|  | 0 | k | 9 | q | 18 |
| b | 1 | kh | 10 | r | 19 |
| c | 2 | l | 11 | s | 20 |
| ch | 3 | m | 12 | t | 21 |
| d | 4 | n | 13 | th | 22 |
| g | 5 | ng | 14 | tr | 23 |
| gh | 6 | ngh | 15 | v | 24 |
| gi | 7 | nh | 16 | x | 25 |
| h | 8 | ph | 17 | đ | 26 |

Bảng Nal: Giá trị mã hoá phần bắt đầu

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Phần kết thúc** | **Giá trị** | **Phần kết thúc** | **Giá trị** | **Phần kết thúc** | **Giá trị** |
| a | 0 | oc | 45 | ây | 90 |
| ac | 1 | oe | 46 | ê | 91 |
| ach | 2 | oi | 47 | êch | 92 |
| ai | 3 | om | 48 | êm | 93 |
| am | 4 | on | 49 | ên | 94 |
| an | 5 | ong | 50 | ênh | 95 |
| ang | 6 | op | 51 | êp | 96 |
| anh | 7 | ot | 52 | êt | 97 |
| ao | 8 | oăn | 53 | êu | 98 |
| ap | 9 | oăng | 54 | ô | 99 |
| at | 10 | oăt | 55 | ôc | 100 |
| au | 11 | u | 56 | ôi | 101 |
| ay | 12 | ua | 57 | ôm | 102 |
| e | 13 | uc | 58 | ôn | 103 |
| em | 14 | ui | 59 | ông | 104 |
| en | 15 | um | 60 | ôp | 105 |
| eng | 16 | un | 61 | ôt | 106 |
| eo | 17 | ung | 62 | ăc | 107 |
| ep | 18 | up | 63 | ăm | 108 |
| et | 19 | ut | 64 | ăn | 109 |
| i | 20 | uy | 65 | ăng | 110 |
| ia | 21 | uya | 66 | ăp | 111 |
| ich | 22 | uych | 67 | ăt | 112 |
| im | 23 | uynh | 68 | ơ | 113 |
| in | 24 | uyt | 69 | ơi | 114 |
| inh | 25 | uyên | 70 | ơm | 115 |
| ip | 26 | uyêt | 71 | ơn | 116 |
| it | 27 | uân | 72 | ơp | 117 |
| iu | 28 | uât | 73 | ơt | 118 |
| iêc | 29 | uê | 74 | ư | 119 |
| iêm | 30 | uôc | 75 | ưa | 120 |
| iên | 31 | uôi | 76 | ưc | 121 |
| iêng | 32 | uôm | 77 | ưi | 122 |
| iêp | 33 | uôn | 78 | ưng | 123 |
| iêt | 34 | uông | 79 | ưt | 124 |
| iêu | 35 | uôt | 80 | ưu | 125 |
| o | 36 | uơ | 81 | ươc | 126 |
| oa | 37 | y | 82 | ươi | 127 |
| oach | 38 | âc | 83 | ươm | 128 |
| oai | 39 | âm | 84 | ươn | 129 |
| oan | 40 | ân | 85 | ương | 130 |
| oang | 41 | âng | 86 | ươp | 131 |
| oanh | 42 | âp | 87 | ươt | 132 |
| oat | 43 | ât | 88 | ươu | 133 |
| oay | 44 | âu | 89 |  |  |

Bảng Nal: Giá trị mã hoá phần kết thúc

|  |  |
| --- | --- |
| **Phần dấu** | **Giá trị** |
| Thanh | 0 |
| Sắc | 1 |
| Huyền | 2 |
| Hỏi | 3 |
| Ngã | 4 |
| Nặng | 5 |

Bảng Nal: Giá trị mã hoá phần dấu

|  |  |
| --- | --- |
| **Âm đặc biệt** | **Giá trị** |
| gi | 0 |
| yên | 1 |
| yêm | 2 |
| yêu | 3 |

Bảng Nal: Giá trị mã hoá âm đặc biệt