**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**Viện Công Nghệ Thông Tin Và Truyền Thông**



**ĐỒ ÁN I**

**ĐẾM SỐ KÝ TỰ TRONG VĂN BẢN TIẾNG VIỆT  
VÀ THỐNG KÊ XÁC SUẤT CỦA TỪNG KÝ TỰ**

Sinh viên thực hiện : **Đỗ Văn Thành -20173378**

Thanh.dv173378@sis.hust.edu.vn

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | PGS. TS. ĐẶNG VĂN CHUYẾT  Chữ ký của GVHD |
|  |  |
| **Viện:** | Công nghệ thông tin và truyền thông |
| **HÀ NỘI, 12/2019** | |

**ĐỀ TÀI PROJECT I**

**Mục lục**

**I.Đề tài**

**II.Kỹ thuật sử dụng**

**1.Ngôn ngữ lập trình C**

**2, DEV-C++ (đủ tính năng cho cả C và C++)**

**3, Một số hàm và thư viện cần dùng trong chương trình**

**III. Tìm hiểu về cách đọc FILE UTF-8 ( Dành cho tiếng việt Vietnamese locale)**

**1.Cách đọc file UTF8**

**2.File UTF8 BOM**

**3.Phân vùng của tiếng việt trong UTF8 (Vietnamese locale)**

**IV.Vận dụng**

**1.Thuật toán của chương trình**

**2.Chạy chương trình**

**I.Đề tài**

**-** Sử dụng ngôn ngữ C viết 1 chương trình đọc các văn bản thuần tiếng việt   
- Tính xác suất xuất hiện các ký tự trong văn bản và hiện thị các xác suất này theo ký tự xác suấ giảm dần

**II. Kỹ thuật sử dụng**

**1.Ngôn ngữ lập trình C**

- **Ngôn ngữ lập trình C** là một ngôn ngữ [mệnh lệnh](https://vi.wikipedia.org/wiki/L%E1%BA%ADp_tr%C3%ACnh_ki%E1%BB%83u_m%E1%BB%87nh_l%E1%BB%87nh) được phát triển từ đầu [thập niên 1970](https://vi.wikipedia.org/wiki/Th%E1%BA%ADp_ni%C3%AAn_1970) bởi [Dennis Ritchie](https://vi.wikipedia.org/wiki/Dennis_Ritchie) để dùng trong [hệ điều hành](https://vi.wikipedia.org/wiki/H%E1%BB%87_%C4%91i%E1%BB%81u_h%C3%A0nh) [UNIX](https://vi.wikipedia.org/wiki/Unix). Từ đó, ngôn ngữ này đã lan rộng ra nhiều hệ điều hành khác và trở thành một những ngôn ngữ phổ dụng nhất. C là ngôn ngữ rất có hiệu quả và được ưa chuộng nhất để viết các [phần mềm hệ thống](https://vi.wikipedia.org/wiki/Ph%E1%BA%A7n_m%E1%BB%81m_h%E1%BB%87_th%E1%BB%91ng), mặc dù nó cũng được dùng cho việc viết các [ứng dụng](https://vi.wikipedia.org/wiki/Ph%E1%BA%A7n_m%E1%BB%81m_%E1%BB%A9ng_d%E1%BB%A5ng). Ngoài ra, C cũng thường được dùng làm phương tiện giảng dạy trong [khoa học máy tính](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_m%C3%A1y_t%C3%ADnh) mặc dù ngôn ngữ này không được thiết kế dành cho người nhập môn.

- **Tổng quan**

+/C là một ngôn ngữ lập trình tương đối nhỏ gọn vận hành gần với phần cứng và nó giống với ngôn ngữ Assembler hơn hầu hết các ngôn ngữ bậc cao. Hơn thế, C đôi khi được đánh giá như là "có khả năng di động", cho thấy sự khác nhau quan trọng giữa nó với ngôn ngữ bậc thấp như là Assembler, đó là việc mã C có thể được dịch và thi hành trong hầu hết các máy tính, hơn hẳn các ngôn ngữ hiện tại trong khi đó thì Assembler chỉ có thể chạy trong một số máy tính đặc biệt. Vì lý do này C được xem là ngôn ngữ bậc trung.

+/C đã được tạo ra với một mục tiêu là làm cho nó thuận tiện để viết các chương trình lớn với số lỗi ít hơn trong mẫu hình lập trình thủ tục mà lại không đặt gánh nặng lên vai người viết ra trình dịch C, là những người bề bộn với các đặc tả phức tạp của ngôn ngữ. Cuối cùng C có thêm những chức năng sau:

* Một ngôn ngữ cốt lõi đơn giản, với các chức năng quan trọng chẳng hạn như là những hàm hay việc xử lý tập tin sẽ được cung cấp bởi các bộ thư viện các thủ tục.
* Tập trung trên mẫu hình lập trình thủ tục, với các phương tiện lập trình theo kiểu cấu trúc.
* Một hệ thống kiểu đơn giản nhằm loại bỏ nhiều phép toán không có ý nghĩa thực dụng.
* Dùng ngôn ngữ tiền xử lý, tức là các câu lệnh tiền xử lý C, cho các nhiệm vụ như là định nghĩa các macro và hàm chứa nhiều tập tin mã nguồn (bằng cách dùng câu lệnh tiền xử lý dạng #include chẳng hạn).
* Mức thấp của ngôn ngữ cho phép dùng tới bộ nhớ máy tính qua việc sử dụng kiểu dữ liệu pointer.
* Số lượng từ khóa rất nhỏ gọn.
* Các tham số được đưa vào các hàm bằng giá trị, không bằng địa chỉ.
* Hàm các con trỏ cho phép hình thành một nền tảng ban đầu cho tính đóng và tính đa hình.
* Hỗ trợ các bản ghi hay các kiểu dữ liệu kết hợp do người dùng từ khóa định nghĩa struct cho phép các dữ liệu liên hệ nhau có thể được tập hợp lại và được điều chỉnh như là toàn bộ.

Một số chức năng khác mà C không có (hay còn thiếu) nhưng có thể tìm thấy ở các ngôn ngữ khác bao gồm:

* An toàn kiểu,
* Tự động Thu dọn rác,(dễ thấy java có thu dọn rác).
* Các lớp hay các đối tượng cùng với các ứng xử của chúng (xem thêm hướng đội tượng),
* Các hàm lồng nhau,
* Lập trình tiêu bản hay Lập trình phổ dung,
* Quá tải và Quá tải toán tử,
* Các hỗ trợ cho đa luồng, đa nhiệm và mang.

Mặc dù C còn thiếu nhiều chức năng hữu ích nhưng lý do quan trọng để C được chấp nhận vì nó cho phép các trình dịch mới được tạo ra một cách nhanh chóng trên các nền tảng mới và vì nó cho phép người lập trình dễ kiểm soát được những gì mà chương trình (do họ viết) thực thi. Đây là điểm thường làm cho mã C chạy hiệu quả hơn các ngôn ngữ khác. Thường thì chỉ có ngôn ngữ ASM chỉnh bằng tay chạy nhanh hơn (ngôn ngữ C), bởi vì ASM kiểm soát được toàn bộ máy. Mặc dù vậy, với sự phát triển các trình dịch C, và với sự phức tạp của các CPU hiện đại có tốc độ cao, C đã dần thu nhỏ khác biệt về tốc độ này.

Một lý do nữa cho việc C được sử dụng rộng rãi và hiệu quả là do các trình dịch, các thư viện và các phần mềm thông dịch của các ngôn ngữ bậc cao khác lại thường được tạo nên từ C.

**2, DEV-C++ (đủ tính năng cho cả C và C++)**

**-Dev-C++** là một [môi trường phát triển tích hợp](https://vi.wikipedia.org/wiki/M%C3%B4i_tr%C6%B0%E1%BB%9Dng_ph%C3%A1t_tri%E1%BB%83n_t%C3%ADch_h%E1%BB%A3p) [tự do](https://vi.wikipedia.org/wiki/Ph%E1%BA%A7n_m%E1%BB%81m_t%E1%BB%B1_do) (IDE) được phân phối dưới hình thức [giấy phép Công cộng GNU](https://vi.wikipedia.org/wiki/Gi%E1%BA%A5y_ph%C3%A9p_C%C3%B4ng_c%E1%BB%99ng_GNU) hỗ trợ việc lập trình bằng [C](https://vi.wikipedia.org/wiki/C_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh))/[C++](https://vi.wikipedia.org/wiki/C%2B%2B). Nó cũng nằm trong bộ trình dịch [mã nguồn mở](https://vi.wikipedia.org/wiki/Ph%E1%BA%A7n_m%E1%BB%81m_ngu%E1%BB%93n_m%E1%BB%9F) [MinGW](https://vi.wikipedia.org/w/index.php?title=MinGW&action=edit&redlink=1). Chương trình IDE này được viết bằng ngôn ngữ [Delphi](https://vi.wikipedia.org/wiki/Delphi_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh)).

- Dự án phát triển Dev-C++ được lưu trữ trên [SourceForge](https://vi.wikipedia.org/wiki/SourceForge). Dev-C++ nguyên được phát triển bởi một lập trình viên có tên là Colin Laplace và chỉ chạy trên [hệ điều hành](https://vi.wikipedia.org/wiki/H%E1%BB%87_%C4%91i%E1%BB%81u_h%C3%A0nh) [Microsoft Windows](https://vi.wikipedia.org/wiki/Microsoft_Windows).

- Bloodshed Dev-C++ là một Môi trường Phát triển Tích hợp (IDE) có hỗ trợ đầy đủ tính năng cho ngôn ngữ lập trình C/C++. Nó sử dụng trình MinGW của [GCC](https://vi.wikipedia.org/wiki/B%E1%BB%99_tr%C3%ACnh_d%E1%BB%8Bch_GNU) (Bộ trình dịch GNU) làm trình biên dịch. Dev-C++ cũng có thể được dùng kết hợp với [Cygwin](https://vi.wikipedia.org/wiki/Cygwin) hay bất kỳ trình dịch nền tảng GCC nào khác.[[1]](https://vi.wikipedia.org/wiki/Dev-C%2B%2B#cite_note-1)

- Chương trình cho người dùng có cảm nhận hơi giống với hình thức của chương trình [Microsoft Visual Studio](https://vi.wikipedia.org/wiki/Microsoft_Visual_Studio) vốn được sử dụng rộng rãi hơn. Dev-C++ có một đặc điểm phụ đó là nó sử dung DevPaks, là một phần gồm các gói mở rộng so với môi trường tiêu chuẩn, bao gồm các thư viện, mẫu, và các tiện ích được đưa thêm vào. DevPaks thường có, nhưng không nhất định, tiện ích [GUI](https://vi.wikipedia.org/wiki/Giao_di%E1%BB%87n_ng%C6%B0%E1%BB%9Di_d%C3%B9ng_%C4%91%E1%BB%93_h%E1%BB%8Da) (giao diện người dùng đồ họa), bao gồm các công cụ phổ biến như [GTK+](https://vi.wikipedia.org/wiki/GTK%2B), [wxWidgets](https://vi.wikipedia.org/w/index.php?title=WxWidgets&action=edit&redlink=1), và [FLTK](https://vi.wikipedia.org/w/index.php?title=FLTK&action=edit&redlink=1). Có những DevPaks có chứa các thư viện với thậm chí nhiều hàm chức năng cao hơn.

- Dev-C++ nói chung là một chương trình chỉ chạy trên [Windows](https://vi.wikipedia.org/wiki/Microsoft_Windows). Tuy nhiên cũng có một [phiên bản cho Linux](http://freshmeat.net/projects/dev-cpp), nhưng vẫn trong giai đoạn [alpha](https://vi.wikipedia.org/w/index.php?title=V%C3%B2ng_%C4%91%E1%BB%9Di_c%E1%BB%A7a_ph%E1%BA%A7n_m%E1%BB%81m&action=edit&redlink=1) và chưa được cập nhật trong vòng hơn 6 năm qua.

**3, Một số hàm và thư viện cần dùng trong chương trình**

-Thư viện <locale.h> chứa hàm setlocale(LC\_CTYPE,”vi\_VN.UTF8”) để cài đặt ngôn ngữ địa phương là Vietnamese locale.

-Hàm fopen:

+/Hàm này trả về một con trỏ FILE. Nếu không, NULL được trả về và biến toàn cục errno được thiết lập để chỉ lỗi.

+/Hàm **FILE \*fopen(const char \*filename, const char \*mode)** trong Thư viện C chuẩn mở file được trỏ tham số filename bởi sử dụng chế độ **mode** đã cho.

+/ ở đây ta dùng mode là “r” mở file để đọc. File phải tồn tại

-Hàm fgetc() cho phép đọc từng bytes(8 bit t) trong FILE \*fp :

+/Hàm này trả về ký tự được đọc dưới dạng một unsigned char được ép kiểu thành một int hoặc EOF hoặc lỗi

+/ Cú pháp: int fgetc(FILE \*stream);

**III. Tìm hiểu về cách đọc FILE UTF-8 ( Dành cho tiếng việt Vietnamese locale)**

**1.Cách đọc file UTF8**

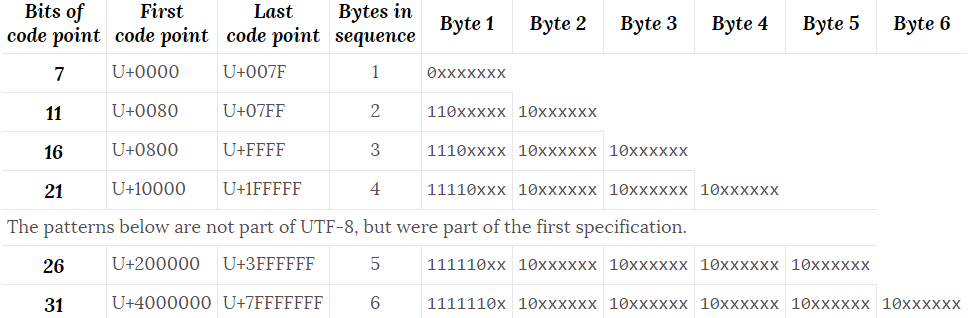
**-**  UTF-8 có quy tắc riêng để định nghĩa số lượng byte cho một ký tự Unicode.

- Trước hết nói về lý do lịch sử, khi bảng ký tự ASCII ra đời thì các ký là ký tự tiếng Anh. Khi đó chỉ cần 1 byte, 255 ký tự, là có thể định nghĩa đầy đủ bảng ký tự.

- Về sau, khi người ta cần thể hiện các ký tự của nhiều ngôn ngữ khác nhau thì họ mới dùng 2 bytes, 65534 ký tự. Rồi vấn đề lại phát sinh khi ta muốn thể hiện tiếng Trung Quốc, Nhật Bản, Hàn Quốc và nhiều loại ngôn ngữ khác nhau thì 2 bytes cũng không đủ.

- Nếu chỉ là 2 bytes không đủ, cứ dùng 4 bytes sẽ đủ, 4294967295 ký tự. Nhưng bây giờ lại phát sinh một vấn đề khác, nếu sử dụng 4 bytes cho tất cả các ký tự. Thì khi viết một văn bản tiếng Anh có 10 ký tự, ta phải dùng 40 bytes thay vì 10 bytes như ban đầu.

- Và từ đây UTF-8, bản mã tối ưu nhất, đã ra đời. Cách hoạt động của UTF-8 như sau:



Đối với ký tự có

* 1 byte, thì bit đầu tiên sẽ là 0
* 2 bytes, thì các bit dầu tiên sẽ là 110, byte tiếp theo 10
* 3 bytes, thì các bit dầu tiên sẽ là 1110, các byte tiếp theo 10 và 10
* 4 bytes, thì các bit dầu tiên sẽ là 11110, các byte tiếp theo 10 và 10 và 10

-Từ nhận xét trên ta có dùng 1 byte bắt đầu là 0 nên giá trị val của bytes đầu là val < 128

-Tương tự dùng 2 bytes, thì các bit dầu tiên sẽ là 110, byte tiếp theo 10 nên giá trị val của bytes đầu là val 128 < val < 224

-Tương tự dùng 3 bytes, thì các bit dầu tiên sẽ là 1110, các byte tiếp theo 10 và 10

nên giá trị val của bytes đầu là val 128 < val < 224

- Tương tự 4 bytes, thì các bit dầu tiên sẽ là 11110, các byte tiếp theo 10 và 10 và 10 nên giá trị val của bytes đầu là val 224 < val < 256

**2.File UTF8 BOM**

**-** Để hiểu file UTF8 BOM là gì thì ta cần biết BOM là gì.

-Theo “Unicode Standard“, BOM là viết tắt của “**Byte Order Mark**“ là dấu hiệu byte nhận biết

- Để dễ hiểu thì mỗi File UTF8 BOM khác vớii UTF8 là FILE UTF8-Bom luôn có 3 bytes đầu *“BF BB BF”* ( 239 187 191)

- Từ kiến thức trên ta có thể dễ dàng kiểm tra xem 1 FILE có phải là File UTF8 hay không

**3.Phân vùng của tiếng việt trong UTF8 (Vietnamese locale)**

**Chú ý** : ở đây tôi chỉ xét hệ 10 ( decimal) để phục vụ cho chương trình

Bảng 1. Chữ cái thường bảng

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ký tự | Byte 1 | Byte 2 | Byte 3 | Ký tự | Byte 1 | Byte 2 | Byte 3 | Ký tự | Byte 1 | Byte 2 | Byte 3 |
| a | 97 | 0 | 0 | ễ | 225 | 187 | 133 | ợ | 225 | 187 | 163 |
| á | 195 | 161 | 0 | ể | 225 | 187 | 131 | ở | 225 | 187 | 159 |
| à | 195 | 160 | 0 | f | 102 | 0 | 0 | ỡ | 225 | 187 | 161 |
| ã | 195 | 163 | 0 | g | 103 | 0 | 0 | p | 112 | 0 | 0 |
| ạ | 225 | 186 | 161 | h | 104 | 0 | 0 | q | 113 | 0 | 0 |
| ă | 196 | 131 | 0 | i | 105 | 0 | 0 | r | 114 | 0 | 0 |
| ắ | 196 | 131 | 0 | í | 195 | 173 | 0 | s | 115 | 0 | 0 |
| ằ | 225 | 186 | 177 | ì | 195 | 172 | 0 | t | 116 | 0 | 0 |
| ẵ | 225 | 186 | 181 | ị | 225 | 187 | 139 | u | 117 | 0 | 0 |
| ẳ | 225 | 186 | 179 | ỉ | 225 | 187 | 137 | ú | 195 | 186 | 0 |
| ặ | 225 | 186 | 183 | ĩ | 196 | 169 | 0 | ù | 195 | 185 | 0 |
| â | 195 | 162 | 0 | j | 106 | 0 | 0 | ũ | 197 | 169 | 0 |
| ấ | 225 | 186 | 165 | k | 107 | 0 | 0 | ủ | 225 | 187 | 167 |
| ầ | 225 | 186 | 167 | l | 108 | 0 | 0 | ụ | 225 | 187 | 165 |
| ẫ | 225 | 186 | 171 | m | 109 | 0 | 0 | Ư | 198 | 176 | 0 |
| ậ | 225 | 186 | 173 | n | 110 | 0 | 0 | ứ | 225 | 187 | 169 |
| ẩ | 225 | 186 | 169 | o | 111 | 0 | 0 | ừ | 225 | 187 | 171 |
| b | 98 | 0 | 0 | ó | 195 | 179 | 0 | ữ | 225 | 187 | 175 |
| c | 99 | 0 | 0 | ò | 195 | 178 | 0 | ử | 225 | 187 | 173 |
| d | 100 | 0 | 0 | õ | 195 | 181 | 0 | ự | 225 | 187 | 177 |
| đ | 196 | 145 | 0 | ỏ | 225 | 187 | 143 | v | 118 | 0 | 0 |
| e | 101 | 0 | 0 | ọ | 225 | 187 | 141 | w | 119 | 0 | 0 |
| é | 195 | 169 | 0 | ô | 195 | 180 | 0 | x | 120 | 0 | 0 |
| è | 195 | 168 | 0 | ố | 225 | 187 | 145 | y | 121 | 0 | 0 |
| ẻ | 225 | 186 | 187 | ổ | 225 | 187 | 149 | ý | 195 | 189 | 0 |
| ẽ | 225 | 186 | 189 | ỗ | 225 | 187 | 151 | ỳ | 225 | 187 | 179 |
| ẹ | 225 | 186 | 185 | ộ | 225 | 187 | 153 | ỹ | 225 | 187 | 185 |
| ê | 195 | 170 | 0 | ồ | 225 | 187 | 147 | ỵ | 225 | 187 | 181 |
| ế | 225 | 186 | 191 | ơ | 198 | 161 | 0 | ỷ | 225 | 187 | 183 |
| ề | 225 | 187 | 129 | ớ | 225 | 187 | 155 | z | 225 | 187 | 185 |
| ệ | 225 | 187 | 135 | ờ | 225 | 187 | 157 | ả | 225 | 186 | 163 |

Bảng 2 chữ cái in hoa

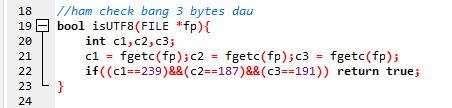
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ký tự | Byte 1 | Byte 2 | Byte 3 | Ký tự | Byte 1 | Byte 2 | Byte 3 | Ký tự | Byte 1 | Byte 2 | Byte 3 |
| A | 65 | 0 | 0 | Ể | 225 | 187 | 130 | Ờ | 225 | 187 | 156 |
| Á | 195 | 129 | 0 | Ễ | 225 | 187 | 132 | Ở | 225 | 187 | 158 |
| À | 195 | 128 | 0 | ÊJ | 225 | 187 | 134 | Ỡ | 225 | 187 | 160 |
| Ả | 225 | 186 | 162 | F | 70 | 0 | 0 | Ợ | 225 | 187 | 162 |
| Ã | 195 | 131 | 32 | G | 71 | 0 | 0 | P | 80 | 0 | 0 |
| Ạ | 225 | 186 | 160 | H | 72 | 0 | 0 | Q | 81 | 0 | 0 |
| Ă | 196 | 130 | 0 | I | 73 | 0 | 0 | R | 82 | 0 | 0 |
| Ắ | 225 | 186 | 174 | Í | 195 | 141 | 0 | S | 83 | 0 | 0 |
| Ằ | 225 | 186 | 176 | Ì | 195 | 140 | 0 | T | 84 | 0 | 0 |
| Ẵ | 225 | 186 | 180 | Ỉ | 225 | 187 | 136 | U | 85 | 0 | 0 |
| Ẳ | 225 | 186 | 178 | Ĩ | 196 | 168 | 0 | Ú | 195 | 154 | 0 |
| Ặ | 225 | 186 | 182 | Ị | 225 | 187 | 138 | Ù | 195 | 153 | 0 |
| Â | 195 | 130 | 0 | J | 74 | 0 | 0 | Ũ | 197 | 168 | 0 |
| Ấ | 225 | 186 | 164 | K | 75 | 0 | 0 | Ủ | 225 | 187 | 166 |
| Ầ | 225 | 186 | 166 | L | 76 | 0 | 0 | Ụ | 225 | 187 | 164 |
| Ẫ | 225 | 186 | 170 | M | 77 | 0 | 0 | Ư | 198 | 175 |  |
| Ẩ | 225 | 186 | 168 | N | 78 | 0 | 0 | Ứ | 225 | 187 | 168 |
| Ậ | 225 | 186 | 172 | O | 79 | 0 | 0 | Ừ | 225 | 187 | 170 |
| B | 66 | 0 | 0 | Ó | 195 | 147 | 0 | Ử | 225 | 187 | 172 |
| C | 67 | 0 | 0 | Ò | 195 | 146 | 0 | Ữ | 225 | 187 | 174 |
| D | 68 | 0 | 0 | Õ | 195 | 149 | 0 | Ự | 225 | 187 | 176 |
| Đ | 196 | 144 | 0 | Ỏ | 225 | 187 | 142 | V | 86 | 0 | 0 |
| E | 69 | 0 | 0 | Ọ | 225 | 187 | 140 | W | 87 | 0 | 0 |
| É | 195 | 137 | 0 | Ô | 195 | 148 | 0 | X | 88 | 0 | 0 |
| È | 195 | 136 | 0 | Ố | 225 | 187 | 144 | Y | 89 | 0 | 0 |
| Ẽ | 225 | 186 | 188 | Ồ | 225 | 187 | 146 | Ý | 195 | 157 | 0 |
| Ẻ | 225 | 186 | 186 | Ỗ | 225 | 187 | 150 | Ỳ | 225 | 187 | 178 |
| Ẹ | 225 | 186 | 184 | Ộ | 225 | 187 | 152 | Ỷ | 225 | 187 | 182 |
| Ê | 195 | 138 | 0 | Ổ | 225 | 187 | 148 | Ỹ | 225 | 187 | 184 |
| Ế | 225 | 186 | 190 | Ơ | 198 | 160 | 0 | Ỵ | 225 | 187 | 180 |
| Ề | 225 | 187 | 128 | Ớ | 225 | 187 | 154 | Z | 90 | 0 | 0 |

**IV.Vận dụng**

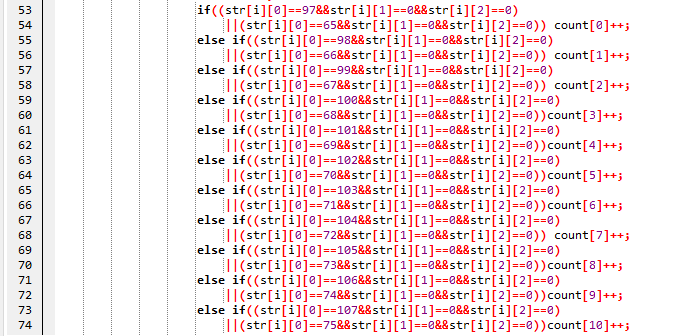
**1.Thuật toán của chương trình**

**-** Hàm kiểm tra xem ký tự được lưu bởi bào nhiêu bytes dựa vào lý thuyết ở phần

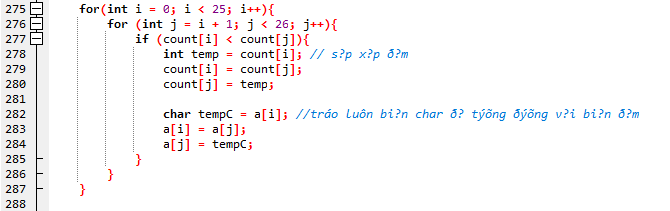
3 mục 1 :



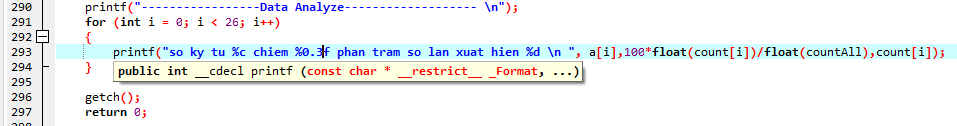
-Hàm kiểm chuỗi đang đọc bởi con trỏ nếu trùng thì công count[i] lên 1



-Hàm sắp xếp tần suất xuất hiện giảm dần tròn chuỗi char a[] trong chương trình



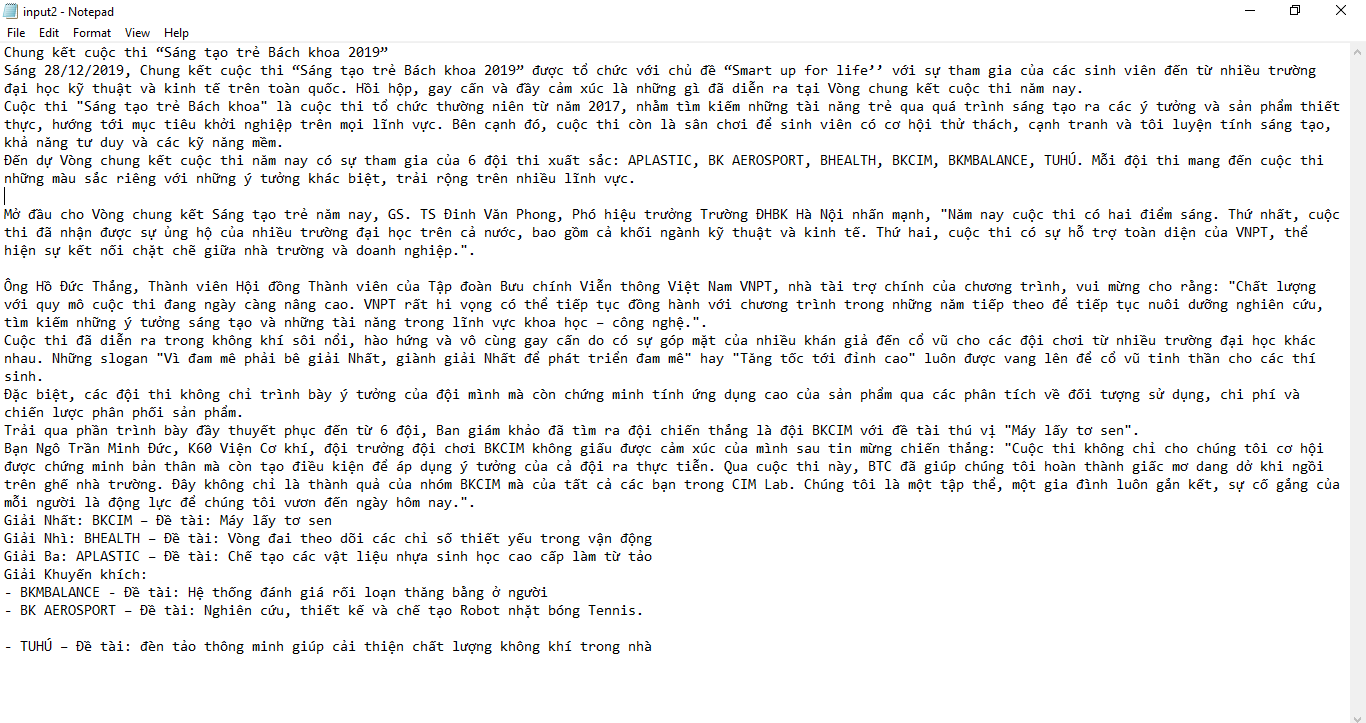
-Cuối cùng là hiển thị tần suất :



Chú ý : Là chương trình này các ký tự là {a,á,à,ả,ã,ạ,ă,â…} thì mình chỉ tính cho ký tự a vì đơn giản là chương trình này có tác dụng thống kê nên quan trong là ta biết được ký tự nào xuất hiện nhiều.

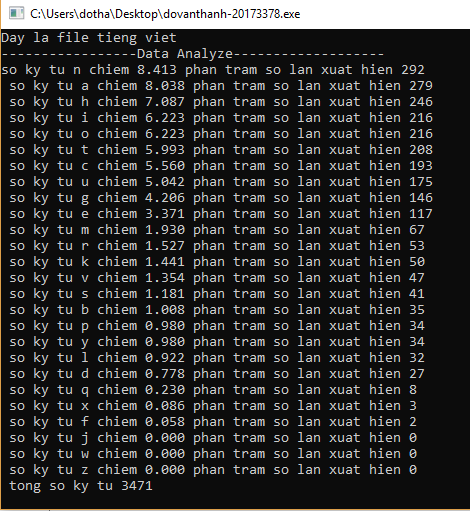
**2.Chạy chương trình**

Ở đây ta đếm số ký tự trong file input2.txt



Đây là 1 văn bản mà ta lấy trên trang chủ cảu BKHN – 1 bài viết (thuần tiếng việt)

Chạy chương trình ta được kết quả như sau



# TÀI LIỆU THAM KHẢO

1.UTF8-BOM : <https://chungnguyen.xyz/posts/co-gi-khac-nhau-giua-utf-8-va-utf-8-without-bom>

2.Convert UTF8 to decimal : <https://onlineutf8tools.com/convert-utf8-to-decimal>

3.Hàm fgetc : <https://vietjack.com/thu-vien-c/ham-fgetc-trong-c.jsp>

4. Hàm fopen : <https://vietjack.com/thu-vien-c/ham-fopen-trong-c.js>

5. Cách đọc và lưu từng ký tự trong file UTF8 : <http://www.wikiwand.com/en/UTF-8#/Sample_code>

6 Tìm hiểu về C và Dev C++ : <https://vi.wikipedia.org/wiki/Trang_Ch%C3%ADnh>