TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN 1**

**Đề Tài:**

**XÂY DỰNG CẶP CÂU SONG NGỮ**

**NGA-VIỆT CHO BÀI TOÁN DỊCH MÁY**

*Người hướng dẫn*: **Nguyễn Chí Thiện**

*Người thực hiện*: **Trương Thuận Lộc – MSSV 51480144**

**Quý Duy – MSSV 51383036**

Lớp **: 13850301 – 14850301**

Khoá**: CD 09-10**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2018**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN 1**

**Đề Tài:**

**XÂY DỰNG CẶP CÂU SONG NGỮ**

**NGA-VIỆT CHO BÀI TOÁN DỊCH MÁY**

*Người hướng dẫn*: **Nguyễn Chí Thiện**

*Người thực hiện*: **Trương Thuận Lộc – MSSV 51480144**

**Quý Duy – MSSV 51383036**

Lớp**: 13850302 - 14850302**

Khoá**: CD 09-10**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2018**

LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn các Thầy Cô trong khoa Công Nghệ Thông Tin đã tận tình giảng dạy, trang bị cho em những kiến thức quý báu trong suốt quá trình học cũng như đã tạo điều kiện cho em thực hiên đề tài này. Đặc biệt chúng em xin gửi lời cảm ơn sâu sắc đến thầy Nguyễn Chí Thiện. Thầy đã tận tình hướng dẫn, chỉ bảo, giúp đỡ cho em để chúng em trong suốt quá trình hoàn thành được đề tài này.

Cảm ơn các ta đã ủng hộ, đóng góp ý kiến để em hoàn thành đề tài của mình. Dù đã rất nỗ lực để hoàn thành đề tài nhưng chắc chắn đề tài em làm vẫn còn nhiều thiếu sót, em mong sẽ nhận đuợc sự góp ý, phê bình của quý Thầy Cô, các anh chị và các ta để đề tài hoàn thiện hơn.

Em xin chân thành cảm ơn!

ĐỒ ÁN ĐƯỢC HOÀN THÀNH

TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của Thầy Nguyễn Chí Thiện. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong luận văn còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào chúng em xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Trương Thuận Lộc*

*Quý Duy*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**MỤC LỤC**

[LỜI CẢM ƠN i](#_Toc515731220)

[ĐỒ ÁN ĐƯỢC HOÀN THÀNH ii](#_Toc515731221)

[TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG ii](#_Toc515731222)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc515731223)

[**MỤC LỤC** 1](#_Toc515731224)

[CHƯƠNG 1 GIỚI THIỆU ĐỀ TÀI 3](#_Toc515731225)

[1.1 GIỚI THIỆU 3](#_Toc515731226)

[1.2 LÝ DO CHỌN ĐỀ TÀI 3](#_Toc515731227)

[1.3 KIẾN THỨC LIÊN QUAN 3](#_Toc515731228)

[1.3.1 PYTHON 3](#_Toc515731229)

[1.3.2 IDE PYCHARM 4](#_Toc515731230)

[1.3.3 THƯ VIỆN XlsxWriter 5](#_Toc515731231)

[1.3.4 PYTHON TRANSLATE APIs 5](#_Toc515731232)

[CHƯƠNG 2 XÂY DỰNG 6](#_Toc515731233)

[2.1 CÀI ĐẶT & SỬ DỤNG PYTHON TRANSLATE APIs 6](#_Toc515731234)

[2.2 ĐỌC GHI FILE & XỬ LÝ CHUỔI TRONG PYTHON 7](#_Toc515731235)

[2.3 GHI DANH SÁCH CÁC CÂU VÀO FILE EXCEL ĐỀ TẠO TẬP SONG NGỮ NGA-VIỆT 9](#_Toc515731236)

[CHƯƠNG 3 TỔNG KẾT 9](#_Toc515731237)

[3.1 KẾT QUẢ ĐẠT ĐƯỢC 9](#_Toc515731238)

[TÀI LIỆU KHAM KHẢO 10](#_Toc515731239)

[PHỤ LỤC 11](#_Toc515731240)

# GIỚI THIỆU ĐỀ TÀI

## GIỚI THIỆU

Trong môi trường kinh doanh quốc tế hiện nay, các doanh nghiệp thuộc các quốc gia khác nhau sẽ có nhu cầu kết nối để cùng tiềm kiếm nguồn lợi nhuận khổng lồ. Nhưng sự bất tiện lớn nhất là rào cảng ngôn ngữ. Nhu cầu thông dịch ngày càng nhiều nhưng số lượng nhân sự chưa đáp ứng được vì nhiều lý do khác nhau, ....

Cùng với sự phát triển của Công Nghệ Thông Tin ngày nay, thì máy móc cũng là một trợ thủ đắc lực cho những công việc như tính toán, quản lý , lưu trữ thì bên cạnh đó việc thông dịch một ngôn ngữ khác củng đang được các lập trình viên xây dựng và phát triển.

## LÝ DO CHỌN ĐỀ TÀI

Với những điều nêu trên thì đồ án này chúng em xây dựng nhằm tạo ra tập dữ liệu song ngữ Nga-Việt. Tập dữ liệu này góp phần giúp ích cho các lập trình viên đang xây dựng ứng dụng hoặc website thông dịch.

## KIẾN THỨC LIÊN QUAN

### PYTHON

Chúng em xây dựng tập dữ liệu này trên ngôn ngữ Python , các thư viện và APIs hỗ trợ.

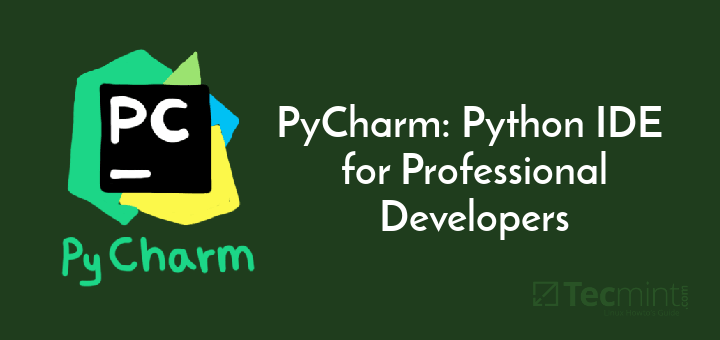
Python là một ngôn ngữ lập trình thông dịch (interpreted), hướng đối tượng (object-oriented), và là một ngôn ngữ bậc cao (high-level)  ngữ nghĩa động (dynamic semantics). Python hỗ trợ các module và gói (packages), khuyến khích chương trình module hóa và tái sử dụng mã. Trình thông dịch Python và thư viện chuẩn mở rộng có sẵn dưới dạng mã nguồn hoặc dạng nhị phân miễn phí cho tất cả các nền tảng chính và có thể được phân phối tự do.

Các điểm mạnh của Python:

* Ngữ pháp đơn giản, dễ đọc.
* Vừa hướng thủ tục (procedural-oriented), vừa hướng đối tượng (object-oriented)
* Hỗ trợ module và hỗ trợ gói (package)
* Xử lý lỗi bằng ngoại lệ (Exception)
* Kiểu dữ liệu động ở mức cao.
* Có các bộ thư viện chuẩn và các module ngoài, đáp ứng tất cả các nhu cầu lập trình.
* Có khả năng tương tác với các module khác viết trên C/C++ (Hoặc Java cho Jython, hoặc .Net cho IronPython).
* Có thể nhúng vào ứng dụng như một giao tiếp kịch bản (scripting interface).

### IDE PYCHARM

PyCharm được phát triển bởi Jet Brains, cung cấp cho người dùng bản Community miễn phí, dùng thử 30 ngày cho phiên bản chuyên nghiệp, $213 – $690 phí đăng ký hằng năm. Khả năng hỗ trợ code toàn diện và phân tích làm cho PyCharm là IDE tốt nhất cho các lập trình Python tất cả các cấp độ. PyCharm cũng hỗ trợ các ngôn ngữ khác và hoạt động trên nhiều nền tảng, vì vậy thực tế bất cứ ai cũng có thể sử dụng nó.



**Ngôn ngữ được hỗ trợ:** AngularJS, Coffee Script, CSS, Cython, HTML, JavaScript, Node.js, Python, TypeScript và các ngôn ngữ mẫu

**Ưu điểm:**

* Tương thích với Windows, Linux và Mac OS
* Đi kèm với Django IDE
* Dễ dàng tích hợp với Git, Mercurial và SVN
* Giao diện tùy biến với mô phỏng VIM
* Trình sửa lỗi JavaScript, Python và Django
* Hỗ trợ Google App Engine

### THƯ VIỆN XlsxWriter

XlsxWriter vẫn còn trong giai đoạn trứng của nó như là một mô-đun, nhưng hỗ trợ tính năng tạo XLSX cơ bản.  
Điều này bao gồm định dạng tế bào, ký hiệu tế bào, nhiều bảng tính thao tác, công thức, ngày tháng, số và viết văn bản.  
Tất cả các bảng tính được xuất ra tương thích với các tiêu chuẩn tập tin XLSX

### PYTHON TRANSLATE APIs

Googletrans là thư viện python miễn phí và không giới hạn được triển khai API Google Translate. Nó sử dụng API Ajax của Google Translate để thực hiện cuộc gọi đến các phương thức như phát hiện và dịch.

Tính năng:

* Nhanh chóng và đáng tin cậy - nó sử dụng cùng một máy chủ mà translate.google.com sử dụng
* Tự động phát hiện ngôn ngữ
* Bản dịch hàng loạt
* URL dịch vụ tùy chỉnh
* Kết nối tổng hợp (lợi thế của việc sử dụng requests.Session)
* Hỗ trợ HTTP / 2

# XÂY DỰNG

## CÀI ĐẶT & SỬ DỤNG PYTHON TRANSLATE APIs

Để cài đặt, hãy sử dụng những thứ như pip với gói “googletrans” hoặc tải xuống gói và đặt thư mục “googletrans” vào đường dẫn python.

$ pip install googletrans

Cách sử dụng: Nhập thư viện googletrans để sử dụng chức năng dịch chuỗi, sau đó ta khởi tạo đối tượng translator, đối tượng này sẽ gọi hàm dịch

#-^-encoding=utf-8-^-  
From goolestrans import Translator  
translator = Translator()

Ví dụ: đoạn code dưới đây dịch chuỗi kì tự Hàn ra ngôn ngữ Anh

>>> **from** **googletrans** **import** Translator  
>>> translator = Translator()  
>>> translator.translate('안녕하세요.')  
# <Translated src=ko dest=en text=Good evening. pronunciation=Good evening.>

## ĐỌC GHI FILE & XỬ LÝ CHUỔI TRONG PYTHON

f = open (‘tenfile.txt‘,’r’,encoding = ‘utf-8’)  
str = f.read()

Trước hêt ta load file cần xử lý lên và lưu vào str ký tự “r” có ý nghỉa mở file để đọc, và “encoding” là chỉ định kiểu text utf-8, nếu không có input encoding này chung ta sẽ gặp lỗi.

UnicodeDecodeError: 'charmap' codec can't decode

Vì file của chúng ta là file tiếng Việt gồm nhiều câu kết thúc bằng dấu chấm, khi được đọc và lưu vào biến str nó có dạng chuổi nên chúng em sẽ dùng hàm split để cắt chuổi bằng ký tự chấm.

List = str.split(‘.’)

Duyệt list trên bằng dòng for với item trong list nếu item bằng các chuỗi trống thì bỏ qua không chép, ngược lại thì chép item đó vào file kết quả, điểu kiện if trên xuất phát từ ký tự “…” và “..” hàm split sẽ không phân biệt được với dấu chấm kết thúc câu, vì thế em chọn các bỏ qua chuỗi trống để file kết quả được các câu kết thúc bằng dấu chấm và trên từng dòng khít nhau, tránh các dòng trống.

For item in List:  
 if item == ‘’ or item == ‘ ‘:  
 countinue  
 eles:  
 f.write(c+’.\n’)  
f.close()

Ta dùng dòng for duyệt lần lướt cho file tiếng Nga và file tiếng Việt để có được kết qua mong muốn như ví dụ sau:

|  |  |
| --- | --- |
| File tiếng Việt | File tiếng Nga |
| Nội dung dòng 1.  Nội dung dòng 2.  Nội dung dòng 3. | Nội dung dòng 1.  Nội dung dòng 2.  Nội dung dòng 3. |

Sau khi có hai file với mẩu như trên thì ta mở lần lượt hai file kết quả đó dùng đến vòng lặp for và translate api duyệt qua và dịch từng dòng để dịch cả hai bảng ra tiếng Anh, ngôn ngử thứ ba này giúp ta so sánh xem sư tương đồng nghĩa giửa câu dòng thư i trong n dòng giửa hai file Nga – Việt, kêt quà được mong đời là cả hai câu sẽ tướng số thứ tự dòng và cả ngữ nghĩa của chúng.

For item in List:  
 result = translator.translate(item,dest=’en’  
 result.text = result.encod(‘ascii’,’ignore).decode(‘ascii’)  
 f.write(result.text+’.\n’)  
f.close()

Hàm encode dưới dây để mã hỏa chuổi ký tự được đọc từ file text thành ascii và sẽ báo lội ignore nếu gặp lội sau đó decode để giải mã:

result.text = result.encod(‘ascii’,’ignore).decode(‘ascii’)

Vì ở tài liệu gốc tiếng Nga và phiên bảng tiếng Việt của có sứ khác nhau về câu cú trình bày do phiên dịch nên sự so sánh cần được làm thủ công để tương biết chính xác có tương ứng nghỉa với nhau hay không, mẩu xem xét cần phải đủ lớn để tao ra độ tin cậy cao.

## GHI DANH SÁCH CÁC CÂU VÀO FILE EXCEL ĐỀ TẠO TẬP SONG NGỮ NGA-VIỆT

Để Python có thể làm việc với file Excel ta cần dùng đến thư viện XlsxWriter, trước tiên ta vẫn cài đặt bằng pip.

$ pip install XlsxWriter

Sau đó ta nhập thư viện XlsxWriter và tạo ra file excel result.xlsx như sau:

workbook = xlsxwriter.Workbook('result.xlsx')  
worksheet = workbook.add\_worksheet()

Sau đó mở lần lượt từng file.text để tạo thành list và chép list và file excel với col = 0 là cột đầu tiên trong excel và row = -1 khi vào vòng for trước khi chép chuỗi nó sẽ được cộng lên 1 đơn vị tương ứng với dòng đầu tiên, điều này tránh việc nhảy dòng khi chép.

col=0  
row=-1  
For item in List:  
 row +=1  
 worksheet.write(row,col,item)  
worksheet.close()

# TỔNG KẾT

## KẾT QUẢ ĐẠT ĐƯỢC

Theo đánh giá chủ quan chúng em đ thực hiện được 80% của kết quả mong đợi tuy nhiên sẽ có sai sót. Trong đề tài này chúng em đã học được rất nhiều kỷ năng xử lý lỗi cũng như ngôn ngữ python.

TÀI LIỆU KHAM KHẢO

PHỤ LỤC

.