**ĐẠI HỌC THỦ DẦU MỘT**



**BÀI TẬP LỚN MÔN  
THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU**

**THU THẬP VÀ TIỀN XỬ LÍ DỮ LIỆU TUYỂN DỤNG LAO ĐỘNG TẠI VIỆT NAM**

Sinh viên thực hiện: **HOÀNG KIM TUYẾN**

Mã số sinh viên: **1824801040043**

Lớp**: D18HT01**

***Bình Dương, tháng 11 năm 2020***

1. **LỜI CAM ĐOAN**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của ThS. Hồ Ngọc Trung Kiên. 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 báo cáo 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 tôi xin hoàn toàn chịu trách nhiệm về nội dung báo cáo của mình.** Trường Đại học Thủ Dầu Một không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*Bình Dương, ngày 11 tháng 7 năm 2020*

*Người thực hiện  
 (ký tên và ghi rõ họ tên)*

**MỤC LỤC**

[PHẦN 1 4](#_Toc56772246)

[1.1. Mục đích đề tài 4](#_Toc56772247)

[1.2. Câu hỏi nghiên cứu 5](#_Toc56772248)

[1.4. Phân tích dữ liệu 6](#_Toc56772249)

[1.5. Dự đoán 12](#_Toc56772250)

[ Xử lí *StopWord* 12](#_Toc56772251)

[ TF-IDF 14](#_Toc56772252)

[ Tìm kiếm 16](#_Toc56772253)

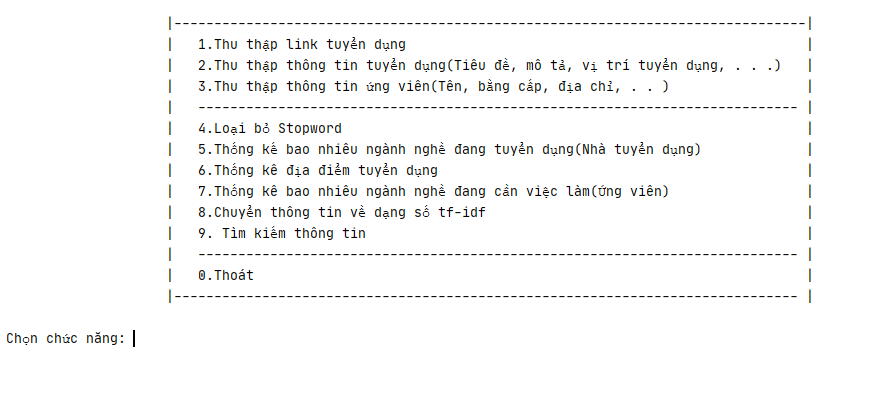
[PHẦN 2: TỰ CHẤM 18](#_Toc56772254)

[DANH MỤC TÀI LIỆU THAM KHẢO 18](#_Toc56772255)

# PHẦN 1

## 1.1. Mục đích đề tài

Thu thập thông tin tuyển dụng trên website tìm việc nhằm thống kê nhu cầu tuyển dụng lao động tại Việt Nam

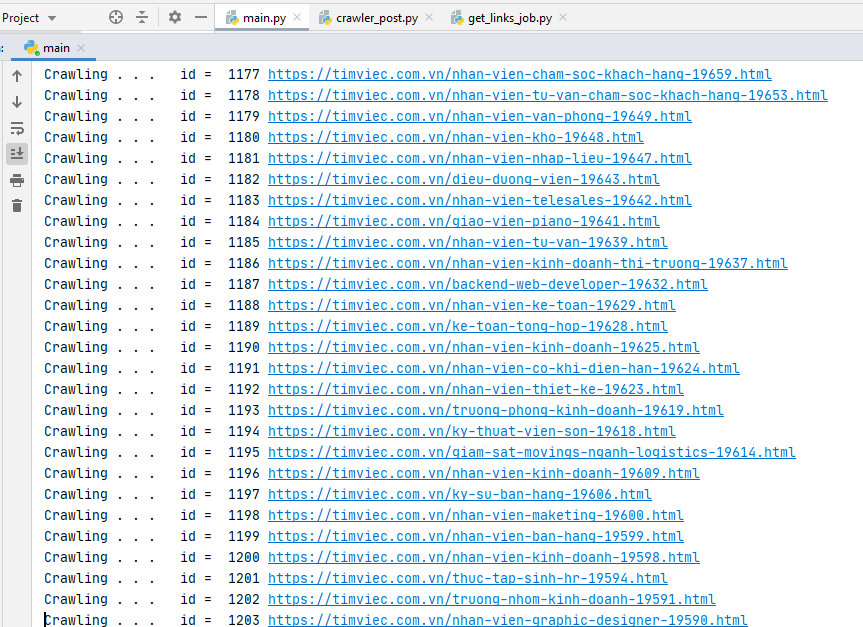
Từ các thông tin, dữ liệu thu thập được, tiến hành xử lí, phân tích. Từ đó đưa ra các dự đoán về nhu cầu tuyển dụng và xu hướng tìm việc trong tuyển dụng lao động tại Việt Nam.

Giao diện các chứng năng của đề tài(Pycharm)

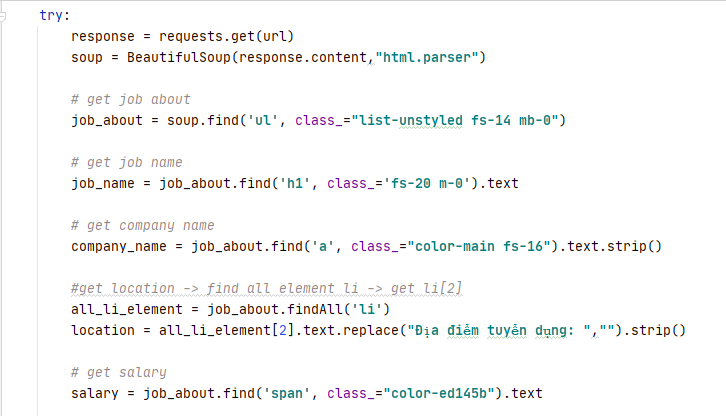
## 1.2. Câu hỏi nghiên cứu

* Sử dụng thư viện trong Python để thu thập và lưu trữ 3000 tin tuyển dụng và 3000 ứng viên.
* Thu thập dữ liệu tại trang website tuyển dụng: **timviec.com.vn**
* Các trường dữ liệu thu thập *thông tin tuyển dụng* gồm:
* Tiêu đề, mô tả
* Link bài viết
* Tên công ty, địa điểm tuyển dụng
* Ngành nghề tuyển dụng
* Mức lương
* Các trường dữ liệu thu thập *thông tin ứng viên* gồm:
* Họ và tên ứng viên
* Tên công việc ứng tuyển
* Địa điểm ứng tuyển
* Số năm kinh nghiệm

## 1.4. Phân tích dữ liệu

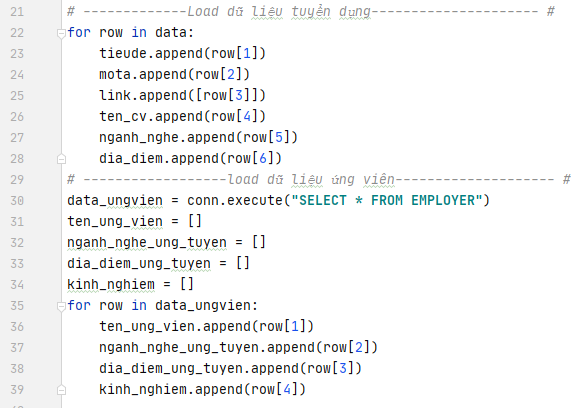
Sử dụng các thư viện **BeautifulSoup**, **requests**, **newspaper**, **sqlite3** để thu thập thông tin tuyển dụng, các thư viện **nltk** để xử lí ngôn ngữ tự nhiên.

Tiến hành thu thập dữ liệu từ website tuyển dụng(timviec.com.vn)

* Sử dụng các hàm trong thư viện **BeautifulSoup,** **newspaper**  để tìm ra các khối dữ liệu sau đó tiến hành xử lí, loại bỏ các khối *html*, *dấu cách*, . . . và lưu vào **database(sqlite3)**

Sử dụng thư viện để thu thập thông tin từ website tuyển dụng

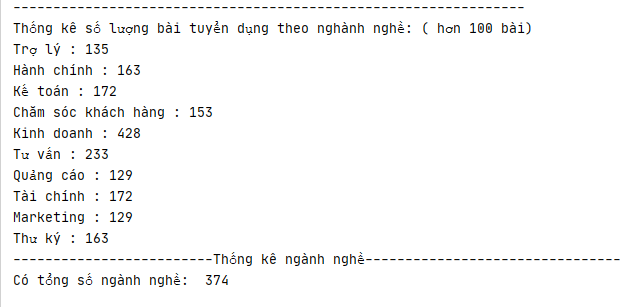
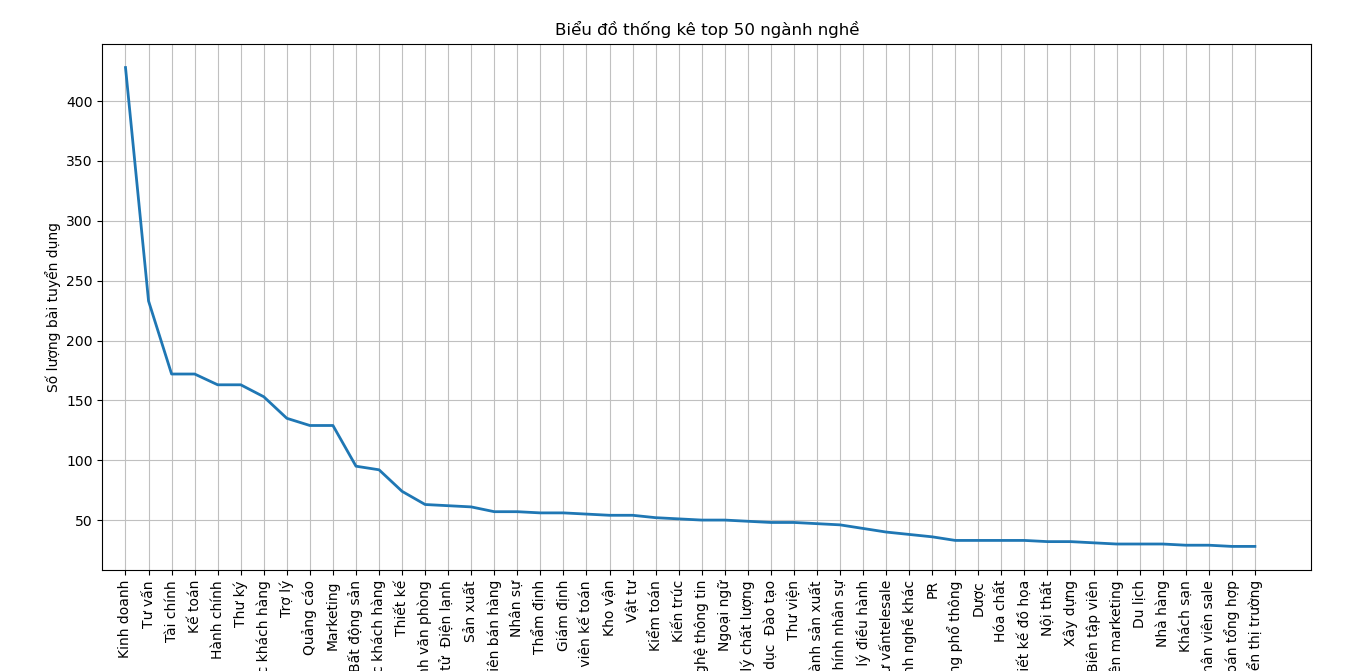
Lưu trữ thông tin thu thập được vào cơ sở dữ liệu

* Sau khi thu thập dữ liệu, lưu trữ dữ liệu, tiến hành load dữ liệu lên và xử lí các yêu cầu như phân tích, thống kê

Tiến hành load dữ liệu

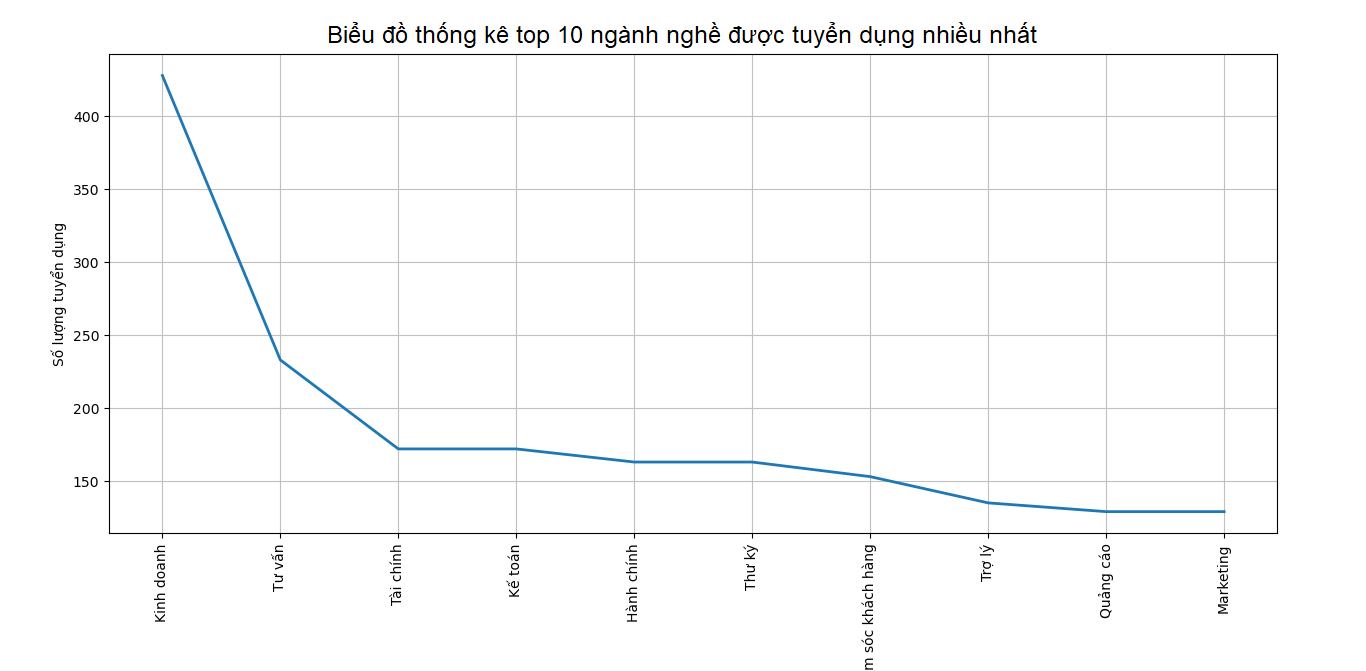
* Sau khi có dữ liệu, tiến hành phân tích và vẽ biểu đồ

Xử lí, thống kê các ngành nghề và vẽ biểu đồ



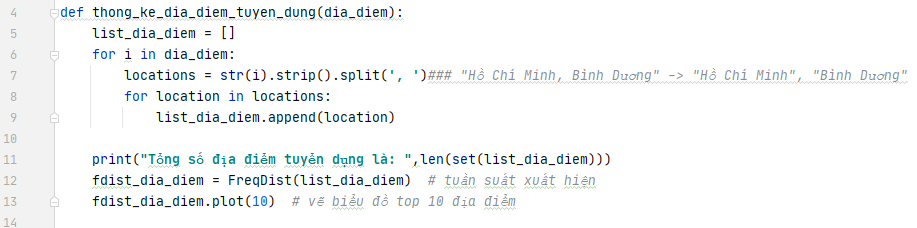
Biểu đồ thống kê top 50 ngành nghề được tuyển dụng nhiều nhất

Kết quả thống kê ngành nghề tuyển dụng

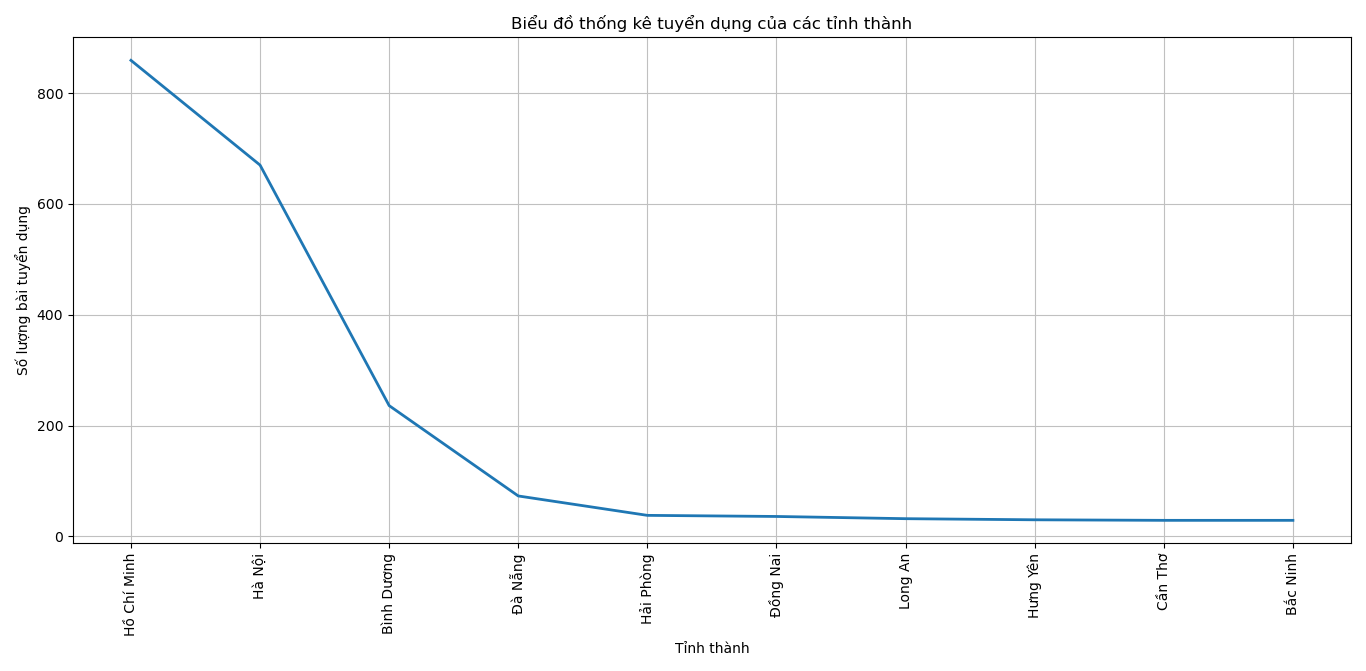
*⏵Kinh doanh* là ngành nghề có nhu cầu tuyển dụng nhiều nhất

Top 10 ngành nghề được tuyển dụng nhiều nhất

* Thống kê nhu cầu tuyển dụng của các tỉnh thành trên cả nước

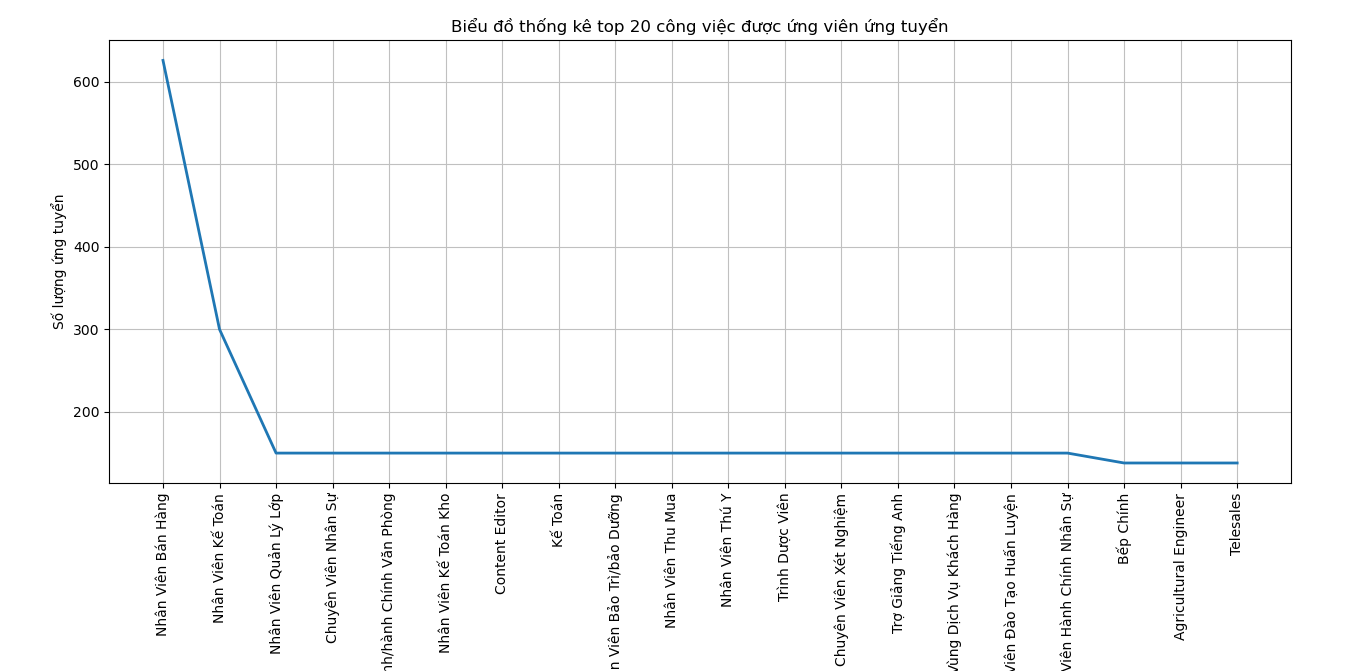


Thống kê các địa điểm tuyển dụng

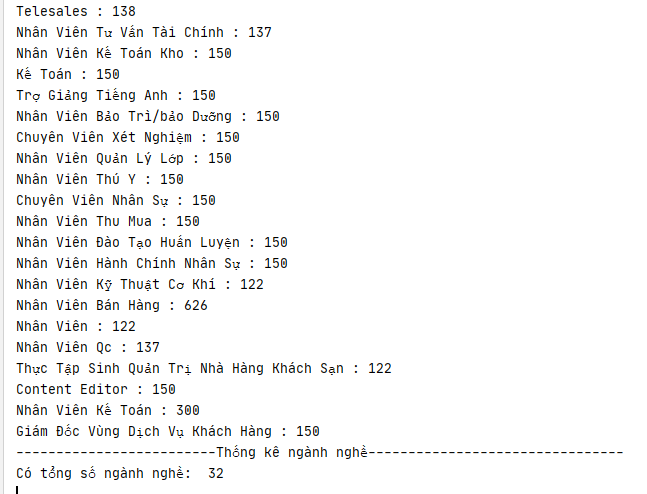


Biểu đồ thống kê nhu cầu việc làm của các tỉnh thành

⏵ Tỉnh thành có nhu cầu tuyển dụng cao nhất là *Thành Phố Hồ Chí Minh*, sau đó là *Hà Nội, Bình Dương, Đà Nẵng,* . . .

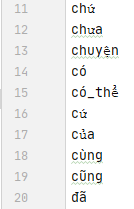
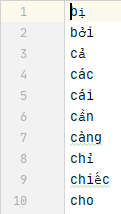
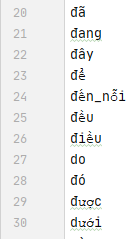
* Thống kê các ngành nghề cần việc làm(ứng viên ứng tuyển)

Biều đồ thống kê ngành nghề cần ứng viên ứng tuyển

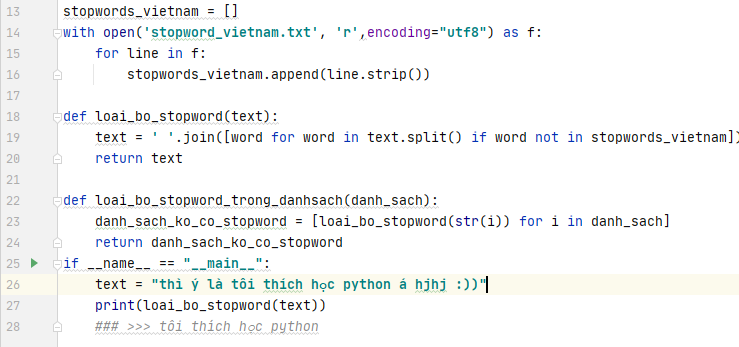


Kết quả thống kê ngành nghề ứng viên ứng tuyển

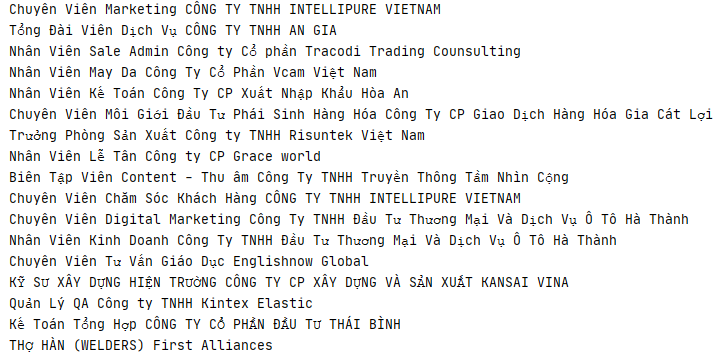
## 1.5. Dự đoán

* Xử lí *StopWord*
* StopWords là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng việt StopWords là những từ như: để, này, kia... Tiếng anh là những từ như: is, that, this...

Danh sách StopWord sưu tầm trên wikipedia

* Sau khi thu thập stopword, tiến hành loại bỏ chúng trong tiêu đề, mô tả của dữ liệu

Xử lí, loại bỏ StopWord

* Kết quả sau khi loại bỏ stopword:

Tiêu đề sau khi loại bỏ stopword

* Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của tư duy và giao tiếp.
* Bag of Words (Bow)
* Bag of Words là một thuật toán hỗ trợ xử lý ngôn ngữ tự nhiên và mục đích của BoW là phân loại text hay văn bản. Ý tưởng của BoW là phân tích và phân nhóm dựa theo “Bag of Words”(corpus). Với test data mới, tiến hành tìm ra số lần từng từ của test data xuất hiện trong "bag".
* TF-IDF
* tf–idf, viết tắt của thuật ngữ tiếng Anh term frequency – inverse document frequency, của một từ là một con số thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.
* TF- term frequency – tần số xuất hiện của 1 từ trong 1 văn bản. Cách tính:

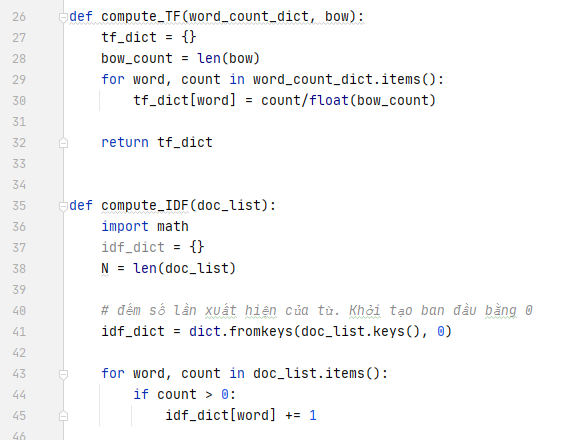


Trong đó:

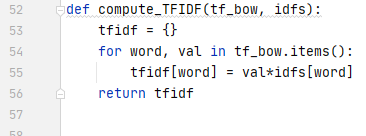
* f(t,d) - số lần xuất hiện từ t trong văn bản d
* max{f(w,d):w∈d} - số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản.
* IDF – inverse document frequency. Tần số nghịch của 1 từ trong tập văn bản (corpus).
* Tính IDF để giảm giá trị của những từ phổ biến. Mỗi từ chỉ có 1 giá trị IDF duy nhất trong tập văn bản. Cách tính:



Trong đó:

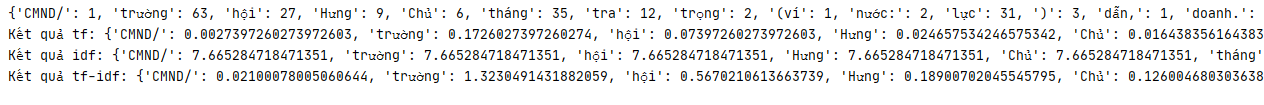
* idf(t, D): giá trị idf của từ t trong tập văn bản
* |D|: Tổng số văn bản trong tập D
* |{d ∈ D : t ∈ d}|: thể hiện số văn bản trong tập D có chứa từ t.
* Cụ thể, chúng ta có công thức tính tf-idf hoàn chỉnh như sau: tfidf(t, d, D) = tf(t, d) x idf(t, D)
* Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).
* Tiến hành cài đặt tf, idf, tf-idf :

Tính tf, idf

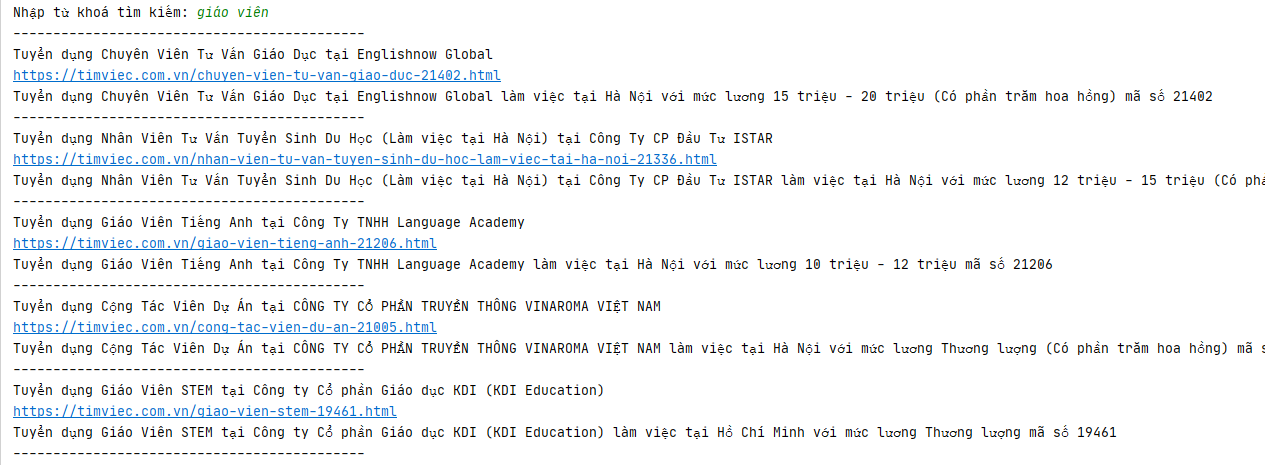
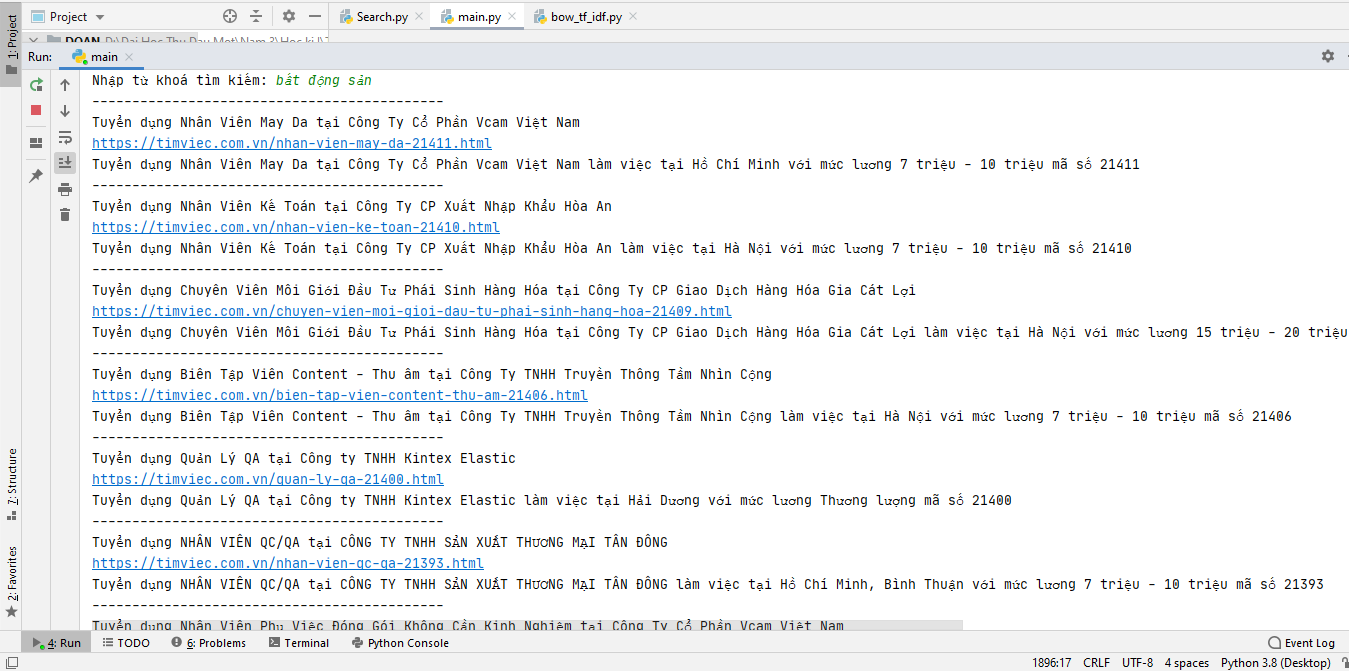
* Tính tf-idf bằng cách nhân tf với idf lại với nhau:

Tính tf-idf

Kết quả tính tf-idf trong một bài tuyển dụng ngẫu nhiên

* Tìm kiếm

Tìm kiếm các thông tin liên quan đến từ khoá tìm kiếm



Kết quả tìm kiếm với từ khoá “bất động sản”

Kết quả tìm kiếm với từ khoá “giáo viên”

# PHẦN 2: TỰ CHẤM

|  |  |  |  |
| --- | --- | --- | --- |
| Nội dung | Yêu cầu | Thang điểm | Điểm Thành viên 1 |
| Phần 1 | Dữ liệu | 3 điểm | 3 |
|  | Phân tích | 3 điểm | 3 |
|  | Dự báo | 4 điểm | 3 |
| **Tổng** | | 10 | 9 |

# DANH MỤC TÀI LIỆU THAM KHẢO

<https://www.nltk.org/book/ch01.html>

<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

<https://newspaper.readthedocs.io/en/latest/>

<https://viblo.asia/p/xu-ly-ngon-ngu-tu-nhien-voi-python-p4-WAyK8RymlxX>

<https://codetudau.com/machine-learning-nlp-scikit-learn/index.html>

<https://codetudau.com/bag-of-words-tf-idf-xu-ly-ngon-ngu-tu-nhien/index.html>

<https://www.sqlitetutorial.net/>

<https://maelfabien.github.io/machinelearning/NLP_2/#2-term-frequency-inverse-document-frequency-tf-idf>