ADVANCED ALGORITHMS

# Address Classification


**Instructor:**
*Thầy Trần Tuấn Anh*

**Group:**

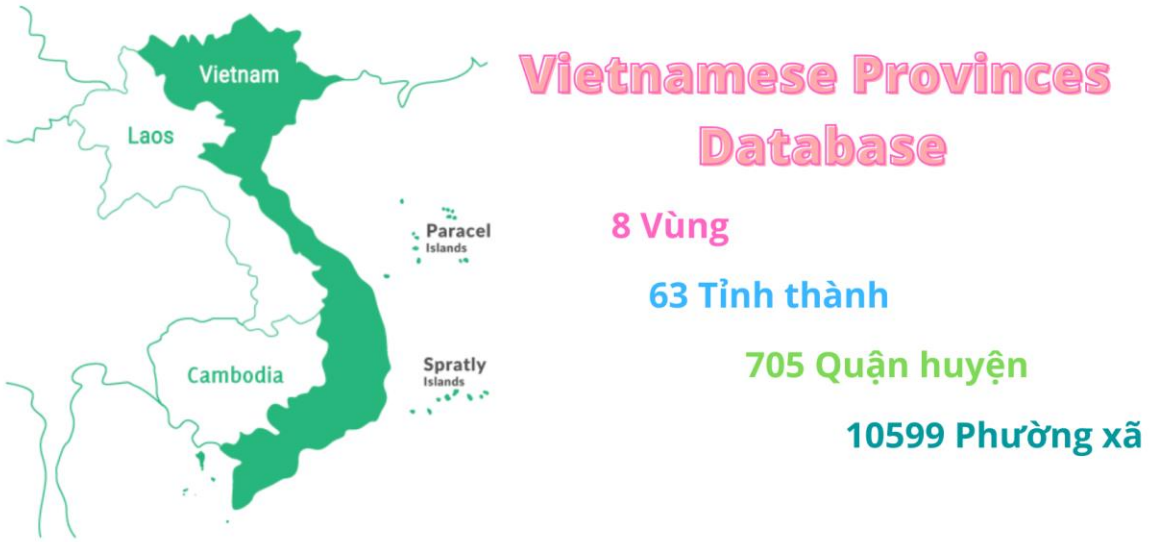
| | | |
|---|---|---|
| Mai Chí Bảo | - | 2370691 |
| Huỳnh Nhật Anh | - | 2370689 |
| Lý Ngọc Trân Châu | - | 2370496 |
| Huỳnh Minh Khôi | - | 2370501 |
| Đặng Lâm Tùng | - | 2370506 |
| Trà Trung Tín | - | 2010702 |

# Agenda

1. Data preparation
2. BM25 introduction
3. Address retrieve algorithm
4. Pre-process
5. Choosing the best match
6. Post-process
7. Result.

# Data preparation

We use the Vietnam province database as the basic of the solution




https://github.com/ThangLeQuoc/vietnamese-provinces-database/blob/master/README_vi.md

# Data preparation

We read all the provinces-database and divide the data into 4 subsection:

Ward: Smallest division name

District: Second division name

Province: Largest division name

Full address: Combination of all 3 levels of divisions by the order of : Ward, District and Province

# Information Retrieval Algorithms

- Retrieval algorithms excel in address classification by weighing the importance of terms like administrative divisions, crucial for accurate categorization.
- BM25 stands out as the best choice within this group due to its adaptive nature, accommodating varying address lengths and structures effectively.
- BM25's customizable parameters enable fine-tuning, leading to enhanced classification precision.
- Offer scalability, efficiently processing large address datasets

--> **With proven effectiveness in text classification tasks, BM25 is a reliable choice for address classification**.

# BM25

- **Term Weighting:** Calculating weights based on term frequency and inverse document frequency
- **Parameters:** BM25's parameters, like k1 and b, allow for adaptability to different address lengths and structures
- **Fit Function to the Dataset:** Enable customization to fit the specific characteristics of address data, enhancing classification accuracy by fine-tuning the algorithm to focus on relevant features of addresses
- **Compute Score**: Computes a relevance score between a query and a document by considering factors such as term frequency, inverse document frequency, document length, and average document length
- **Rank**: The computed score is then used to rank documents in information retrieval systems, ensuring the most relevant documents are presented to the user

# Address retrieve algorithm

# Pre Process

+ Detect shorten name of District or Town.

"TXChu Se" → "Chu Se".

+ Lower case, remove dots, commas:

“Trúc Bạch, Ba Đình, Hà Nội” → “trúc bạch  ba đình  hà nội”

Remove Stopword:

“Quận 10, Thành phố hồ chí minh" → “Quận 10 hồ chí minh"

# Choosing best match from BM25

First we choose k best match of BM25

```
scores = self.bm25.search(query)

scores_index = np.argsort(scores)[::-1]

top_results = np.array([self.addresses[i] for i in scores_index][:k])
```

After that, we use editdistance to find the most similar address from k results

```
final_result = min(top_results, key=lambda result: editdistance.eval(result['full_address'].replace(",",""), text))

return final_result
```

# Post Process

- In the post process stage, we use the edit distance to find the nearest address of each level of the result to the nearest address in the database and replace them, in order to correct any remain grammar error.

# Result

- The result with public test:

| correct | total | score / 10 | max_time_sec | avg_time_sec |
|---|---|---|---|---|
| 1092 | 1350 | 8.09 | 0.1047 | 0.0249 |

- Detail result:

| ID | text | province | province_student | nce_c | district | district_student | district_correct | ward | ward_student | ward_correct | total_correct | time_sec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | TT Tân Bình Huyện Yên So | Tuyên Quang | Tuyên Quang | 1 | Yên Sơn | Yên Sơn | 1 | Tân Bình | | 0 | 2 | 0.0272136 |
| 1 | 357/28,Ng-T- Thuật,P1,Q3 | Hồ Chí Minh | | 0 | | 3 | 0 | | | 1 | 1 | 0.011351 |
| 2 | 284DBis Ng Văn Giáo, P3, | Tiền Giang | | 0 | Mỹ Tho | | 0 | 3 | | 0 | 0 | 0.0136934 |
| 3 | Nà Làng Phú Bình, Chiêm I | Tuyên Quang | Tuyên Quang | 1 | Chiêm Hóa | Chiêm Hóa | 1 | Phú Bình | Phú Bình | 1 | 3 | 0.015911 |
| 4 | 59/12 Ng-B-Khiêm, Đa Ka | Hồ Chí Minh | Hồ Chí Minh | 1 | 1 | 1 | 1 | Đa Kao | | 0 | 2 | 0.0133453 |
| 5 | 46/8F Trung Chánh 2 Trun | Hồ Chí Minh | Hồ Chí Minh | 1 | Hóc Môn | Hóc Môn | 1 | Trung Chánh | Trung Chánh | 1 | 3 | 0.0140814 |
| 6 | T18,Cẩm Bình, Cẩm Phả, C | Quảng Ninh | Quảng Ninh | 1 | Cẩm Phả | | 0 | Cẩm Bình | Cẩm Hải | 0 | 1 | 0.0146864 |
| 7 | Thanh Long, Yên Mỹ Hưng | Hưng Yên | Hưng Yên | 1 | Yên Mỹ | Yên Mỹ | 1 | Thanh Long | Thanh Long | 1 | 3 | 0.0149168 |
| 8 | D2, Thạnh Lợi, Vĩnh Thạnh | Cần Thơ | Cần Thơ | 1 | Vĩnh Thạnh | Vĩnh Thạnh | 1 | Thạnh Lợi | Thạnh Lợi | 1 | 3 | 0.012912 |
| 9 | Cổ Lũy Hải Ba, Hải Lăng, C | Quảng Trị | Quảng Trị | 1 | Hải Lăng | Hải Lăng | 1 | Hải Ba | Hải Ba | 1 | 3 | 0.0145074 |
| 10 | Khu phố 4 Thị trấn, Dương | Tây Ninh | Tây Ninh | 1 | Dương Minh Ch | Dương Minh Châu | 1 | | | 1 | 3 | 0.0149893 |
| 11 | Khu phố 3, Trảng Dài, Thà | Đồng Nai | Đồng Nai | 1 | Biên Hòa | | 0 | Trảng Dài | | 0 | 1 | 0.017213 |
| 12 | Số Nhà 38, Tổ 9 Tô Hiệu, T | Sơn La | Sơn La | 1 | Sơn La | | 0 | Tô Hiệu | | 0 | 1 | 0.0180808 |
| 13 | CH F1614-HH2-Khu ĐTM D | Hà Nội | Hà Nội | 1 | Hà Đông | Hà Đông | 1 | Yên Nghĩa | | 0 | 2 | 0.0180728 |
| 14 | Khu 3 Suối Hoa, Thành phố | Bắc Ninh | Bắc Ninh | 1 | Bắc Ninh | | 0 | Suối Hoa | | 0 | 1 | 0.0130634 |
| 15 | Phú Lộc Phú Thạnh, Phú T | An Giang | An Giang | 1 | Phú Tân | Phú Tân | 1 | Phú Thạnh | Phú Thạnh | 1 | 3 | 0.0173433 |
| 16 | Nguyễn Khuyến Thị trấn V | Hà Nam | Hà Nam | 1 | Lý Nhân | Lý Nhân | 1 | Vĩnh Trụ | Vĩnh Trụ | 1 | 3 | 0.0186025 |
| 17 | Nam chính Tiền hải, Thái b | Thái Bình | Thái Bình | 1 | Tiền Hải | Tiền Hải | 1 | Nam Chính | Nam Chính | 1 | 3 | 0.0132195 |
| 18 | Đá Hàng Hiệp Thạnh, Gò D | Tây Ninh | Tây Ninh | 1 | Gò Dầu | Gò Dầu | 1 | Hiệp Thạnh | Hiệp Thạnh | 1 | 3 | 0.0129673 |
| 19 | 371/11 Thoại Ngọc Hầu H | Hồ Chí Minh | Hồ Chí Minh | 1 | Tân Phú | Tân Phú | 1 | Hiệp Tân | | 0 | 2 | 0.0174779 |
| 20 | Số 93, khu phố 9, thị trấn I | Phú Yên | Phú Yên | 1 | Sông Hinh | Sông Hinh | 1 | Hai Riêng | Hai Riêng | 1 | 3 | 0.0158615 |
| 21 | Tổ Dân Phố 3 Thị trấn Chư | Gia Lai | Gia Lai | 1 | Chư Prông | Chư Prông | 1 | Chư Prông | Chư Prông | 1 | 3 | 0.0143723 |
| 22 | Xã Minh Đạo, Huyện Tiên | Bắc Ninh | Bắc Ninh | 1 | Tiên Du | Tiên Du | 1 | Minh Đạo | Minh Đạo | 1 | 3 | 0.0140465 |
| 23 | Xã Bồng Khê Huyện Con C | Nghệ An | Nghệ An | 1 | Con Cuông | Con Cuông | 1 | Bồng Khê | Bồng Khê | 1 | 3 | 0.0143931 |
| 24 | Thôn 16 Cư Prông, Ea Kar, | Đắk Lắk | Đắk Lắk | 1 | Ea Kar | Ea Kar | 1 | Cư Prông | Cư Prông | 1 | 3 | 0.0133999 |
| 25 | Số 259/54/8, Tổ 28, KP1, L | Đồng Nai | Đồng Nai | 1 | Biên Hòa | | 0 | Long Bình Tân | | 0 | 1 | 0.0206638 |

# THANK YOU