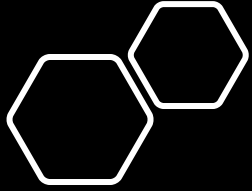


ADDRESS CLASSIFICATION

Trần Tuấn Anh



1. PROBLEM DEFINITION: Address representation

✗ Bad

ADDRESS LINE 1

4321 Sesame Street

ADDRESS LINE 2

Suite 900

2 Text Fields

Input in separate fields
Presses 'tab' for new line
Follows system format
Doesn't support all formats

✓ Good

ADDRESS

4321 Sesame Street
Suite 900

1 Text Area

All input in one field
Presses 'enter' for new line
Follows real world format
Supports international formats

Dr., Mr., Mrs., Ms.	<input type="text"/>
First Name, MI	<input type="text" value="John"/>
Last Name	<input type="text" value="Doe"/>
Street Address	<input type="text" value="My Street"/>
Apt No., Suite No., Box No.	<input type="text"/>
City	<input type="text" value="My City"/>
State or Province	<input type="text" value="MYSTATE"/>
Zip or Postal Code	<input type="text" value="55555"/>
Country	<input type="text" value="USA"/>
Telephone No.	<input type="text" value="800-555-1212"/>
Cellular Phone No.	<input type="text" value="888-555-1212"/>
Fax No.	<input type="text" value="877-555-1212"/>
Email	<input type="text" value="mywork@soandso.com"/>
Email (alternate)	<input type="text" value="myhome@soandso.net"/>

CONTACT

Dropdown

Add new site

Site name

Site name

Address

Type in an address

find

Street number

Number

Street

Street

Town

town

City

City

Region

Region

Postal Code

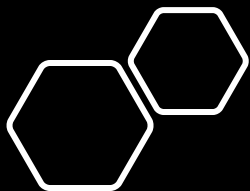
Postal Code

Country

Country

Close

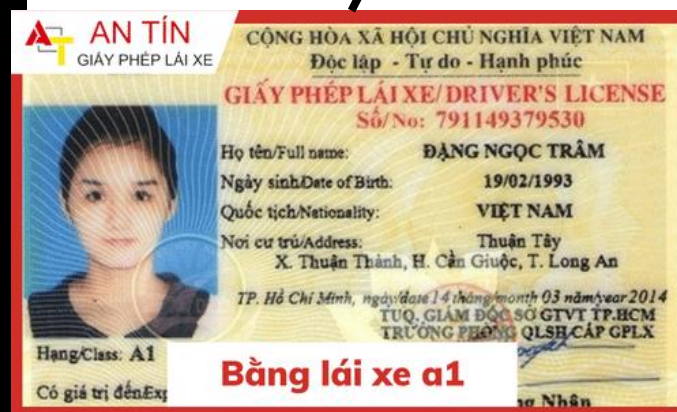
Clarity of
information

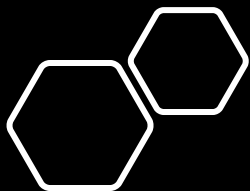


1. PROBLEM DEFINITION: Human Reading and representation



HUMAN READING &
UNDERSTANDING: confusing
rules





1. PROBLEM DEFINITION: Human & Machine Interaction

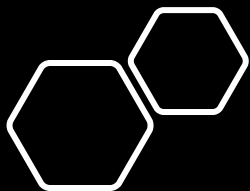
All processes
are correct?



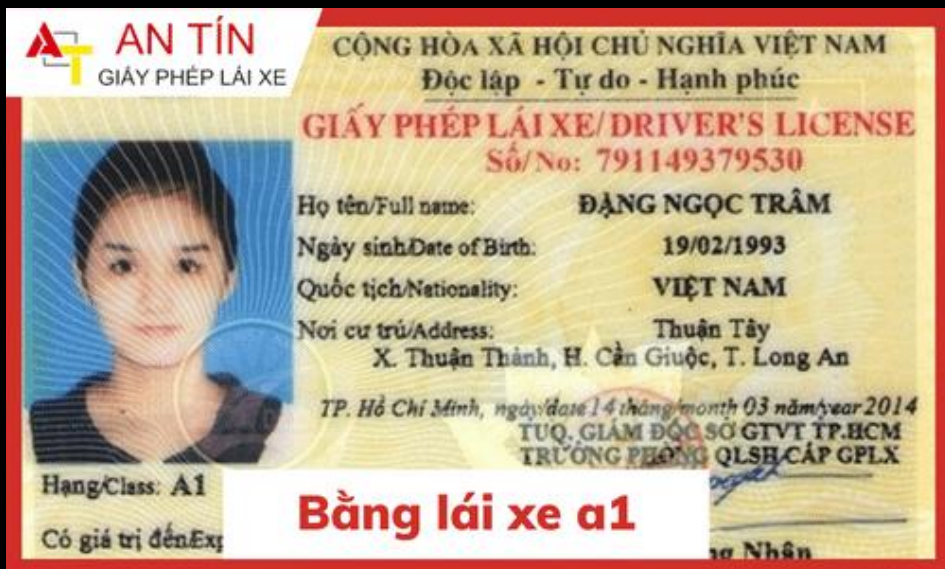
Detect address area

OCR

Classify address

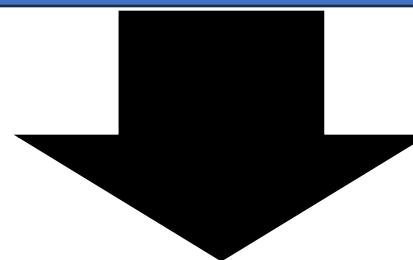


1. PROBLEM DEFINITION: Input vs Output

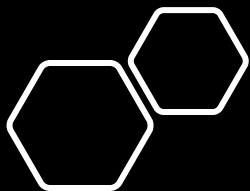


OCR:

X. Thuận Thành, H. Cần Giuộc, T. Long An
Thuận Thanh, HCần Giuộc, Tlong An
Thuận Thành, H Cần Giuộc T. Long An
X Thuận Thành H. Cần Giuộc, Long An
X ThuanThanh H. Can Giuoc, Long An
.....



Address classification:
Xã: Thuận Thành
Huyện: Cần Giuộc
Tỉnh: Long An



Example

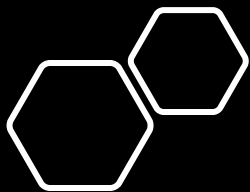
1. PROBLEM DEFINITION: Input vs Output

```
all_raw.json  X
Schema: <No Schema Selected>

1  [
2  {
3    "ground_truth": "Xã Thịnh Sơn H. Đô lương T. Nghệ An",
4    "pred_no_correct": "Xã Thịnh Sơn H., Đô lương T. Nghệ An",
5    "pred_and_correct": "Xã Thịnh Sơn, Đô Lương, T. Nghệ An",
6    "address_info": {
7      "province": "Nghệ An",
8      "district": "Đô Lương",
9      "ward": "Thịnh Sơn"
10   }
11 },
12 {
13   "ground_truth": "X.Hoi Thượng - Đồng Hỷ - Thái 2012",
14   "pred_no_correct": "X.Hoi Thượng - Đồng Hỷ- Thái 20yên",
15   "pred_and_correct": "X.Hóa Thượng",
16   "address_info": {
17     "province": "Thái Nguyên",
18     "district": "Đồng Hỷ",
19     "ward": "Hóa Thượng"
20   }
21 },
22 {
23   "ground_truth": "Tổ 73 - Hoàng Cầu - 1 Chợ Xu - Đống Đa - Hà Ni",
24   "pred_no_correct": "Tổ 73 - Hoàng Cầu -TP.1 C Xm Đống Đa - Hà Ni",
25   "pred_and_correct": "Tổ 73-Hoàng Cầu-TP.1 C Xm, Đống Đa, Hà Nội",
26   "address_info": {
27     "province": "Hà Nội",
28     "district": "Đống Đa",
29     "ward": ""
30   }
31 },
32 }
```

Input

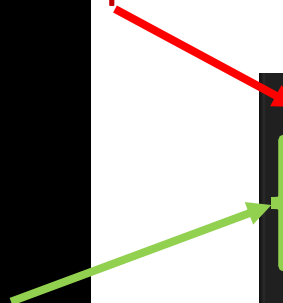
Output








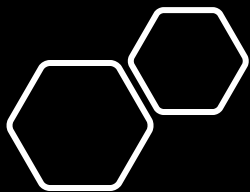
2. DEVELOPMENT & TESTING

Development provided: 1000 samples

Testing provided (private)



	all_raw	10/4/2023 3:39 AM	JSON File	579 KB
	list_district	10/4/2023 3:56 AM	Text Document	8 KB
	list_province	10/4/2023 3:56 AM	Text Document	1 KB
	list_ward	10/4/2023 3:57 AM	Text Document	97 KB
	README	10/4/2023 4:10 AM	MD File	1 KB



3. REQUIREMENTS

General:

- Submit all source code and related data.
- Readme (how to run): The Readme file should be detailed and easily understood.
- The source code will be run on a unique machine without internet.
- **Any copy or cheating will get 0 for the whole course.**

For the exercise:

- Maximum time for 1 request ≤ 0.1 s
- Average time for 1 request: ≤ 0.01 s
- Duration: 3 weeks (a part of private data will be released 3-4 days before the deadline)
- Support 01 free test.
- Do not apply a machine learning approach, just algorithms.
- **Score= 80% private test + 20% public test**



Q&A