Competition #2: Data Science

Address Elements Extraction

13 March 2021

Competition Details

Start time: 13 March 2021, 1pm (GMT+7) / 2pm (GMT+8)

End time: 20 March 2021, 10.59pm (GMT+7) / 11.59pm (GMT+8)

Duration: 1 Week

Table of Contents

Background

At Shopee, we strive to ensure our customers' highest satisfaction for their shopping and delivery experience - fast and accurate delivery of goods. This can be better achieved if we have key address elements for each user address which allows us to accurately geocode it to obtain geographic coordinates to ship the parcel to our customers. These key address elements include Point of Interest (POI) Names and Street Names. However, most addresses that Shopee receives are unstructured and in free text format, not following a certain pattern. Thus it is important for us to develop a model to precisely extract the key address elements from it.

Task

In this competition, you'll work on addresses collected by us to build a model to correctly extract Point of Interest (POI) Names and Street Names from unformatted Indonesia addresses.

Participants are expected to build their own model for this competition, submissions by teams which directly call any third party APIs on the test set will not be taken into consideration.

Scoring Metric

Categorization Accuracy

Evaluation Description

We will be using accuracy as our metric for this competition. It is defined as follows:

$$accuracy((p_i, s_i), (\hat{p}_i, \hat{s}_i)) = \begin{cases} 1 \ \ if \ p_i == \hat{p}_i \ and \ s_i == \hat{s}_i \\ 0 \ \ otherwise \end{cases}$$

Where:

p_i = the actual POI name for ith address
$\hat{p}_i$ = the predicted POI name for ith address
s_i = the actual street name for ith address
$\hat{s}_i$ = the predicted street name for ith address

There are addresses with missing POI or street elements. For such cases, you should leave it empty for that specific element of that address. The formula for the overall metric is defined as the average of accuracy score for each address as following:

$$score = \frac{1}{n}\sum_{i=1}^{n}(accuracy((p_i,\ \hat{p}_i),(s_i,\hat{s}_i)))$$

Where:

$n =$ the total number of addresses

$accuracy =$ the function provided above

Sample Dataset

id	raw_address	POI/street
0	urip sumoharjo 59 mangkujayan cv. tri saka buana ponorogo	cv. tri saka buana/urip sumoharjo
1	karang mulia bengkel mandiri motor raya bosnik 21 blak kota	bengkel mandiri motor/raya bosnik
2	primkob pabri adiwerna	primkob pabri/
3	jalan mh thamrin, sei rengas i kel. medan kota	/jalan mh thamrin
4	smk karya pemban, pon	smk karya pembangunan/pon

Submission Format

Extract the Point of Interest (POI) Names and Street Names, submission file format should be 'csv' file only.

For each 'id' in the test dataset, you need to provide two prediction results, one for POI and one for street. POI and street should be separated with a "/" character without any spaces in between. There are cases where POI/street elements in the raw_address are not complete, for this case, you also need to predict the complete subwords before returning the result.

Examples - assume that:

1) The POI is "bengkel mandiri motor" and street name is "raya bosnik" the returned POI/street should be:
    - bengkel mandiri motor/raya bosnik
2) The POI is "primkob pabri" and no street name is found the returned POI/street should be:
    - primkob pabri/
3) No POI is found and the street name is "jalan mh thamrin" the returned POI/street should be:
    - /jalan mh thamrin
4) The word "pembangunan" in raw_address "smk karya pemban, pon" is not complete. The correct POI will be "smk karya pembangunan" and the returned result should be:
    - smk karya pembangunan/pon

If there is no result for a certain address element, please leave it as empty. The returned result should also be lowercase. The ‘csv’ file should contain a header and have the following format:

id	POI/street
0	/
1	pt tunas artha gardatama/lenteng agung barat
2	/lenteng agung barat
3	muh sigit wahyu p dr sp kj/jalan adi sucipto
4	senam melinda/