Project Name:

Named Entity Recognition for Vietnamese and applying for real application

Project Requirement

- Python 3.7x or higher
- Pandas
- Matplotlib
- Scikit-learn
- Pytorch/Tensflows

Dataset:

- VLSP 2018 NER dataset (https://vlsp.org.vn/vlsp2018/eval/ner)
- Real Problem and data training for domain that you choose. Please define Name entity and Name entity types. Then you could manually label Name entity and their types for training data.

Project Overview

In any text document, there are particular terms that represent specific entities that are more informative and have a unique context. These entities are known as Named Entities, which more specifically refer to terms that represent real-world objects like people, places, organizations, and so on, which are often denoted by proper names. A naive approach could be to find these by looking at the noun phrases in text documents.

Named entity recognition (NER), also known as entity chunking/extraction, is a popular technique used in information extraction to identify and segment the named entities and classify or categorize them under various predefined classes. With named entity recognition, you can extract key information to understand what a text is about, or merely use it to collect important information to store in a database.

This project focuses Vietnamese language. Data using VLSP 2018 NER dataset and yourself dataset for evaluation (https://vlsp.org.vn/vlsp2018/eval/ner).

Problem statement

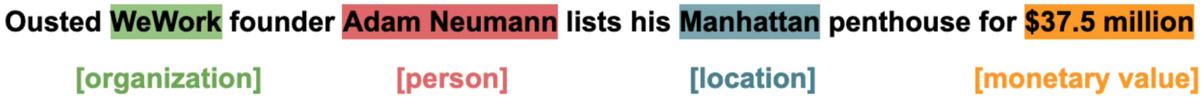
Define the problem which needs to be solved clearly. Describe the input, output. Is it a classification or clustering problem?

Named entity recognition (NER) – also called entity identification or entity extraction – is a natural language processing (NLP) technique that automatically identifies named entities in a text and classifies them into predefined categories.

Input:

Ousted WeWork founder Adam Neumann lists his Manhattan penthouse for $37.5 million.

Output:



Ousted WeWork founder Adam Neumann lists his Manhattan penthouse for $37.5 million

MODUL 1 (Requirement Model):

Input:

- Project's requirement
- Standard structure for a machine learning project

Output and assessment:

#	Deliverable	Assessment
1	Requirement and solution report	• Providing a high-level overview of the project. Background information such as the problem domain, the project origin, and related data sets or input data is provided. • Describe the input, output. Is it a classification or clustering problem? • Evaluation metrics
2	Project plan	• Detailed plan with WBS, working model

3	Working environment for development	• All the required libraries and development tool are installed • The project is structured appropriately

Task list and skills to gain after completion:

#	Tasks	Skills to gain
1	Problem Overview: Research state-of-the-art methods, related works for solving the problems presented: academic papers, products, etc.	• Reading scientific papers • Explore and navigate useful resources in the Internet • Understand industry best-practices
2	Design a specific plan for approaching the problem	• Specify the scope of the problem • Time management • Planning • Teamwork and collaboration
3	Set up the developing environment	• Python environment setup and management • Install libraries • Structure machine learning projects

MODUL 2 (Data Understanding):

Describe the data which has been acquired, including: the format of the data, the quantity of data, for example number of records and fields in each table, the identities of the fields and any other surface features of the data which have been discovered. Does the data acquired satisfy the relevant requirements?

Input:

- Project' requirement
- Public datasets
- Template for data understanding report
- Template for data labeling guideline

Data Preprocessing

- Clean data set, handle missing values if need
- Data augmentation and data labeling if need

- Produce derived attributes (features), entire new records or transformed values for existing attributes
- Splitting data
- Handle imbalanced data sets

Output and assessment:

#	Deliverable	Assessment
2	Data pipeline design	• The pipeline is scalable (when a new source is added)
3	Data Storing and clean data, Data preprocessing	• Data management and versioning scheme is suitable and easy to follow • The normalized structure encompasses all the features and easy to use • Virtualize data • Handle imbalanced data sets
4	Data understanding report	• Can follow the template and fill out the required information
5	Feature extraction report	• Feature design understanding
6	Training data, dev data, test data set	• Methodology of dividing this dataset
7	Newly labeled dataset	• Quantity and quality pass by test domain expert

Task list and skills to gain after completion:

#	Tasks	Skills to gain
1	Dive in Data set. Describe the data which has been acquired, including: the format of the data, the quantity of data, for example number of records and fields in each table, the identities of the fields, etc.	• Understanding of content, businness of data. • Data acquisition process • Document the collected data
2	Build a data pipeline that could gather data from multiple sources and preprocessing them into the designed structure	• Data preprocessing • Data pipeline implementation • Data management • Data versioning
3	Describe the general data statistics and visualize and data analysis	• Statistics • Data visualization • Data technique analysis

4	Feature set extraction	• Feature Engineering • NLP background
5	Labeling and describe the new dataset in a detailed report	• Define of name entity, label and guides for us • Labeling process • Evaluation metric of quality data • NLP background • Write new dataset report

MODUL 3 (Modelling):

- *Select the actual modeling technique that is to be used. If multiple techniques are applied, perform this task for each technique separately. Document the actual modeling technique that is to be used.*
- *Build model: implement your model that you selected.*
- *Create an evaluation measure for test dataset: need to generate a procedure or mechanism to test the model's quality and validity.*
- *With any modeling technique, there are often a number of parameters that can be adjusted. List the parameters and turning this parameters.*

Input:

- Train, validation, and test set
- Feature set
- Metric evaluation

Output and assessment:

#	Deliverable	Assessment
1	Model selection report	• Can follow the template and fill out the required information
2	Project source code	• Follow coding convention and best practices • The evaluation metrics are implemented • All functions are tested • Have APIs for training, evaluating, and running the models given the input. • The source code are documented in detailed so that it can be reused

		• All configurations are maintained in a config file so that they can be adjusted without modifying the source code
3	Based Trained model and improving models	• Basic features set, Basic parameters • Hyper-parameters are optimized
4	Model training report	• Can follow the template and fill out the required information

Task list and skills to gain after completion:

#	Tasks	Skills to gain
1	Select the actual modeling technique that is to be used. If multiple techniques are applied, perform this task for each technique separately.	• Model selection
2	Document the actual modeling technique that is used.	• Write model selection document
3	Create an evaluation measure for test dataset: need to generate a procedure or mechanism to test the model's quality and validity	• Coding
4	Build model: implement your model that you selected.	• Coding • Debugging/Testing
5	Model's parameters fine-tuning. With any modeling technique, there are often a number of parameters that can be adjusted. List the parameters and turning this parameters	• Optimized Technique of Parameters

MODUL 4 (Model Evaluation and Validation):

Input:

- Baseline models
- Feature set
- Train, validation (dev), and test set
- Evaluation metrics

Output and assessment:

#	Deliverable	Assessment
1	Model evaluation report	• Can follow the template and fill out the required information • The errors/weak points of model are analyzed thoroughly by examples, logic or visualization
2	Optimized training model	• This model should better than the baseline model • Model Improving and Model's parameters fine-tuning solution
3	Improving model report	• Errors analysis and suggest solution of improving model.

Task list and skills to gain after completion:

#	Tasks	Skills to gain
1	Evaluate models and compare with the baseline	• Model evaluation metrics
2	Errors analysis and improving model	• Error analysis by examples, logic, visualization • Matplotlib, TensorBoard
3	Improving model solutions	• NLP background • Solutions for fix errors

MODUL 5 (Deploy Model Name Entity Recognition):

Input:

- Training model
- Predicting model

Output and assessment:

#	Deliverable	Assessment
1		• Correct Front-end app

	Deploy model report and management	• Correct Back-end API • Version management
2	Software demo	• Mobie or Web End-to-End application with correct output and function

Task list and skills to gain after completion:

#	Tasks	Skills to gain
1	Introduction to Deployment	• Gain familiarity with cloud and deployment environment. • Understand the machine learning workflow in production • React framework •
2	Deploy a Model	• Deploy a model within Web, Mobie app • Learn to provide access to an endpoint from a website • API in Back-End for AI model in Serving • Use API to integrate ML models into a web app
3	Updating a Model	• Update your model to account for changes in the data

Modul 6:

Review and Conclusion

Input:

- All moduls of the project outputs and deliverables

Output and assessment:

#	Deliverable	Assessment
1	Project summary report (presentation)	• The project is summarized with all the main points • Strength and weakness of the models are explained • The lessons learned bring new insights
2	Technical report (paper???)	• Written in the format of a scientific paper with detailed supplementary
3	Software demo	• Mobie or Web End-to-End application with correct output and function

Task list and skills to gain after completion:

#	Tasks	Skills to gain
1	Write summary report and lesson learned. The final results are discussed in detail	• Result analysis
2	Exploration as to why some techniques worked better than others, or how improvements	• Project summary • Future work proposal
3	Software demo	• System Testing and function testing