|  |  |
| --- | --- |
| A logo on a black background  Description automatically generated | |
| **Faculty of Science and Technology**  **CISC3025 – Natural Language Processing** | |
| **Project 3: Implementation of a Maximum Entropy Model for Named Entity Recognition** | |
| **Group Members:** | |
| Huang Yanzhen | DC126732 |
| Chen Zirui | DC127901 |

Table of Contents

[1 Introduction 3](#_Toc163999609)

[2 Methods and Implementation 3](#_Toc163999610)

[3 Evaluations and Discoveries 3](#_Toc163999611)

[3.1 Advantages and Drawbacks 3](#_Toc163999612)

[3.2 Interesting Phenomenon 3](#_Toc163999613)

[4 Conclusion 3](#_Toc163999614)

# 1 Introduction

Being a part of the syntax level processing techniques in NLP, Named-Entity Recognition is a mechanism for a subtask of information retrieval that’s been applied in various fields. One popular usage is its application in search engines. In this context, NER assists to recognize user queries and extract accurate information from them.

To be more specific, Named Entity Recognition involves assigning a class label for entities within a piece of plain text. Those classes could include person names, firm or organization names, locations, dates, and more customed tags as you want. This project made use of the Maximum Entropy Classifier from *scikit-learn* to implement a simple NER task, which is to recognize person names.

The initial task for building a MaxEnt model is to select proper features. Once the features are selected, the training set, which are the word tokens pre-assigned a label, are passed along with the features to train the model. The model will assign an importance weight to each feature for each class. Lastly, to test the model, for each entity, the model calculates the vote of each class using the features and the weights, finally deciding the class label.

In this project, the training part is done by the encapsulated model for *scikit-learn.* The main workload of this process is to process training data to select features, and to build a website that demonstrates the performance of the model, which runs at the backend in the local server.

# 2 Methods and Implementation

## 2.1 Features Selection

Feature selection aims to find representative features of a word that could produce the maximum entropy of the classification, that is, to best distinguish the classes. There are three types of features, which are Internal Pattern Features, Library Features, and Contextual Features. For the binary case, there are positive and negative features. Positive features indicate that a word is more likely to be a human name. Negative features, on the contrary, unrecommends to assign the label of human name to the word.

### 2.1.1 Internal Pattern Features

Internal Pattern Features put their scopes in the word itself. It analyses the patterns of a word and assumes that some specific pattern can distinguish a person’s name from others.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table.** Internal Pattern Features | | | | |
|  | **Feature Name** | **Feature Description** | **Reason for selection** | **Match Examples** |
| + | p\_cap\_low | Start with Capital letter, and the rest letters are lowercase. There may be a prime after the first capital letter. There may be a second capital letter in the third letter’s space. | In English, names always start with a capital letter. There are also some special name styles, like McArthr, and O’Brien. | Jenny, McArthur, O’Brien |
| + | p\_cap\_period | A single capital letter followed by a period. | In English, human name initials | D., J., M., … |
| - | p\_noun\_like | Have some noun-like ending patterns, like *-tion, -sion, -ance,* etc. | Human names doesn’t always ends with the same pattern as some special nouns that are derived from verbs or adjectives. | mention, motion, addiction, expansion, allowance, … |

The corresponding regular expressions of these features are as follows.

|  |  |  |
| --- | --- | --- |
| **Table.** Regular expressions for internal patterns | | |
|  | **Feature Name** | **Regular Expression** |
| + | p\_cap\_low | ^[A-Z](\'[A-Z]?|[a-z][A-Z])?[a-z]+ |
| + | p\_cap\_period | ^[A-Z]\.$ |
| - | p\_noun\_like | (([aio]?tion|ment|ness|ship|hood|\w+age|[ae]nce)[sd]?$)/i |

### 2.1.2 Library Features

There are a pack of words that is likely not to be classified into human names. For instance, week names like *Monday* and *Tuesday* are not likely to be human names. Country names, like *Korea, China, Japan* or *Britain* are also not likely to be classified to human names. Therefore, we added some libraries that contains these kinds of words and combined them into one evaluation criteria. That is to say, as long as a word belongs to one of the libraries in this set, it will be assigned a negative feature. The negative library set includes the following.

|  |  |
| --- | --- |
| **Table.** Negative feature libraries | |
| **Libraries** | **Python Packages** |
| Week Names | *self-defined* |
| Month Names | *self-defined* |
| Country Names | *geonamescache* |
| City Names | *geonamescache* |
| Stop Words | *nltk.stopwords* |

On the opposite side, there are pack of words that are likely be classified into names. One of the most important packs are, exactly, names. Therefore, an *nltk* corpora is used to identify whether a token is always presented as names. Likewise, if a token belongs to this library, it will be assigned a positive feature.

### 2.1.3 Contextual Features

Contextual Features looks beyond the word itself, and seeks evidence in the contextual environment where the word is in. For instance, a name tends to be in the first and second place of a sentence. A name also tends to be in the main reference of an attributive clause. Unlike Library Features, Contextual Features are discrete, that is each pattern matches only one feature, unlike in Library Features, one feature is matched to numerous patterns.

# 3 Evaluations and Discoveries

## 3.1 Trial and Errors

Over-fitting problem, too much negative features,

Merge Similar Features, Cut unuseful features, …

## 3.2 Advantages and Drawbacks

Perform bad for non-english names, good if it is after non-english names.

Good for attributive clause.

## 3.3 Interesting Phenomenon

# 4 Conclusion