CISC3025 Project 3

NER with Maximum Entropy Model

WONG KAI YUAN (DC026157) | CHAN KA WAI (DC226165)

Table of contents

Table of contents

01

Introduction

Quickly introduce the project and its significance in the field of NLP.

Purpose

Build a **maximum entropy model** (MEM) for identifying person names in newswire texts (Label=PERSON or Label=O), which is also called named entity recognition (NER) with high accuracy using feature sets.

Why NER ?

Data Analysis

Information Retrieval

Sentiment Analysis

NER is good at extracting vital information from unstructured text. Many application on Natural Language Processing.

02

Methodology

Explanation of the Maximum Entropy Model, the advanced features added to improve the model, and the technical implementation details using Python and NLTK.

Maximum Entropy

Utilizes a **probabilistic framework** that predicts class labels (PERSON or not) based on the statistical properties of features extracted from text.

Feature Engineering

Feature	What it does ?	Why ?
1) ALLCAP	Checks if the entire word is in uppercase.	Words fully in uppercase are often acronyms or headings, rarely person names, helping reduce false positives.
2) Lowercase	Saturn is a gas giant composed mostly of hydrogen and helium	Person names usually start with a capital letter; full lowercase often indicates common nouns or other parts of speech.
3) After Symbol	Neptune is the farthest planet from the Sun	Names often follow certain punctuations in written text, indicating a new sentence or clause where names are more likely to appear.

Feature Engineering

Feature	What it does ?	Why ?
4) Number	Identifies presence of numerical characters in a word.	Numerical characters in a word typically signify that it is not a person name, useful for filtering out numerical data or mixed content.
5) Pretitle	Checks if the previous word is a common title or honorific.	Titles precede names, providing a strong contextual hint that the following word is likely a person name, thus enhancing recognition accuracy.

03

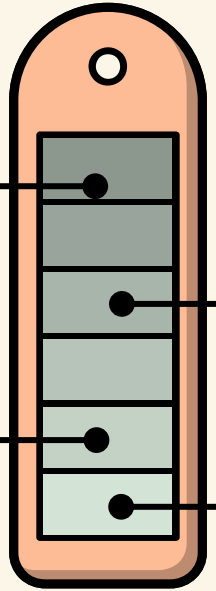
Training & Testing

Description of the training process, data handling, and performance metrics used to evaluate the classifier's effectiveness in identifying person names.

How do we do?

1) Data Preparation

2) Model Training



3) Model Testing

4) Model Evaluation

Using F1 Score, Accuracy, Recall, Precision.

Result (Training)

Iteration	Log Likelihood	Final Accuracy
5	-0.04436	0.981
20	-0.01954	0.998
25	-0.01670	0.999
30	-0.01464	0.999

Result (Testing)

Iteration	F1 Score	Accuracy	Recall	Precision
5	0.9219	0.9739	0.7924	0.9780
20	0.9657	0.9871	0.9002	0.9862
25	0.9657	0.9871	0.9008	0.9858
30	0.9660	0.9872	0.9019	0.9859

Fairly good performance to identify person’s name.

Words	P(PERSON)	P(O)
EU	0.0119	*0.9881
rejects	0.0001	*0.9999
German	0.0453	*0.9547
call	0.0001	*0.9999
to	0.0001	*0.9999
boycott	0.0001	*0.9999
British	0.0464	*0.9536
lamb	0.0001	*0.9999
.	0.0000	*1.0000
Peter	*0.8437	0.1563
Blackburn	*0.5750	0.4250
BRUSSELS	0.2250	*0.7750
1996-08-22	0.0000	*1.0000
The	0.0542	*0.9458
European	0.0446	*0.9554
Commission	0.0450	*0.9550
said	0.0001	*0.9999
on	0.0001	*0.9999
Thursday	0.0437	*0.9563
it	0.0001	*0.9999

During show sample stage, model got every single one correct !

04

Web Demo

A quick demonstration of the web application built to showcase the model in action, illustrating how users can input text and view the NER results.

Overview of Web Application

Demonstrates the **practical application** of the enhanced NER model, allowing users to input text and see the named entity recognition results in **real time.**

—Purpose

Built using **Flask**, a lightweight **Python** web framework, which facilitates the creation of web applications quickly and with minimal code.

—Technology Used

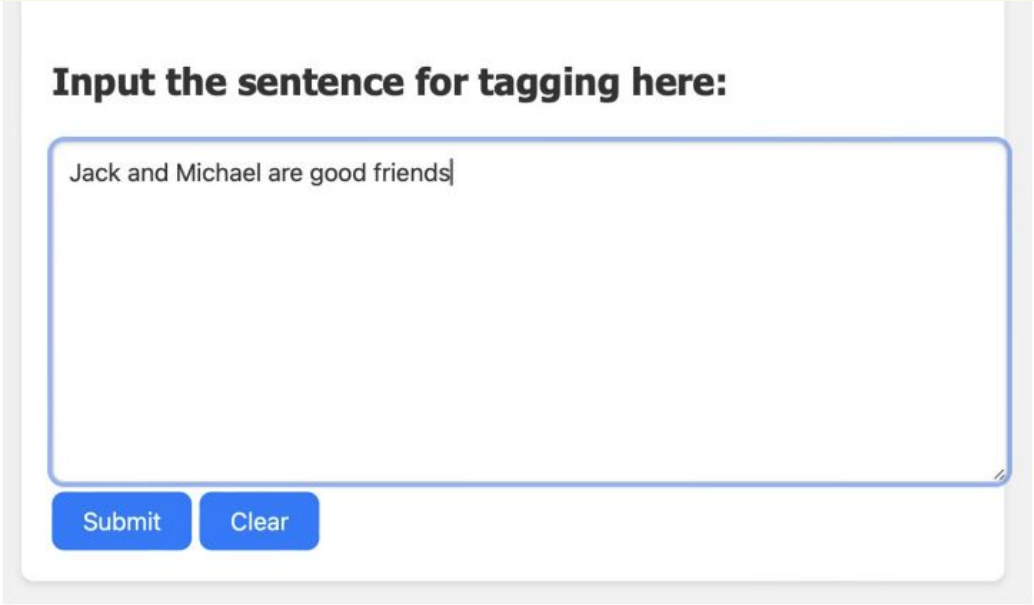
Functionality

1. Users can type or paste text into a text box.

2. Application processes the text using the trained NER model.

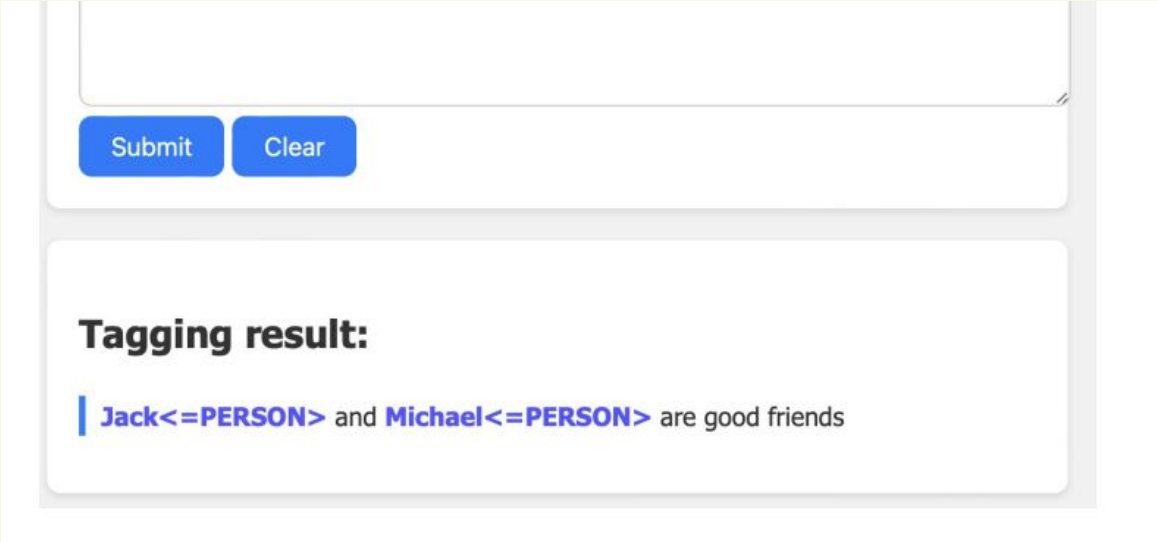
3. The results display each word tagged accordingly as 'PERSON' or 'O' for non-person entities.

Live Demo



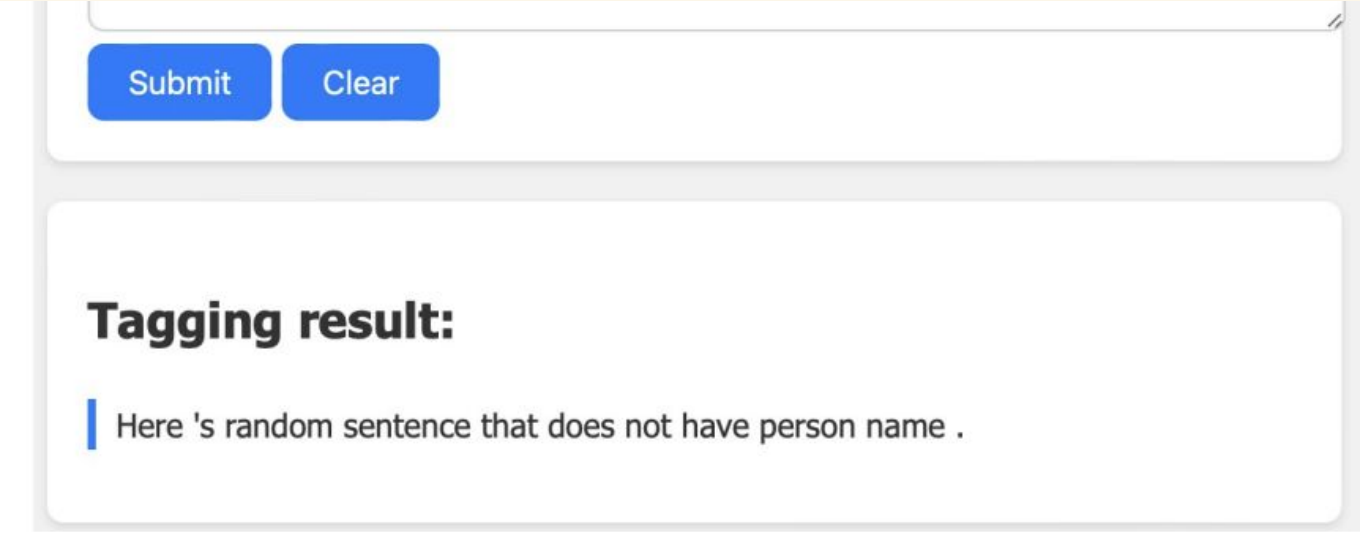
This is the input sentence in web application

Live Demo



This is the output in web application

Live Demo



This is the input sentence without person name in web application

05

Future Work

Brief thoughts on potential future improvements.

Areas of Improvement

Explore the use of **neural networks**, such as LSTM or BERT models, to capture deeper contextual meanings and improve accuracy.

1) Integration of Deep Learning

Enhance the **web application**'s interface to support more interactive features.

2) User Interface Enhancements

Extend the model to recognize more entity types beyond person names, such as **locations, organizations**, and dates.

3) Expand Entity Types

06

Conclusion

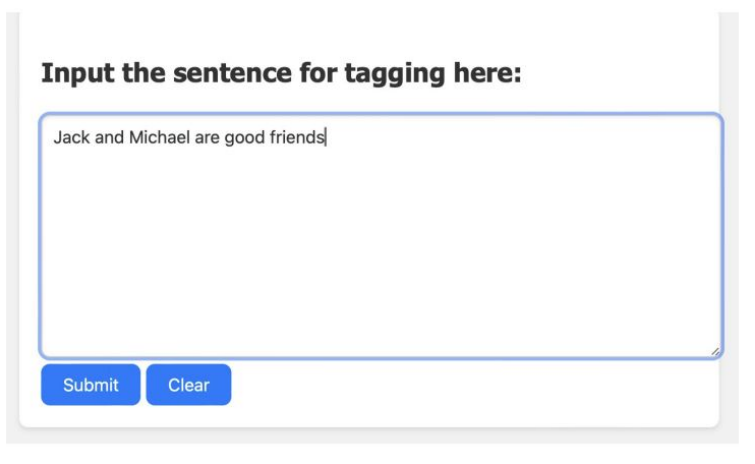
Summary of the project outcomes, highlighting the enhancements made to the NER system and its implications for real-world applications.

Project Recap

- **Objective Achieved:** Successfully enhanced a Named Entity Recognition system using the Maximum Entropy Model to accurately identify person names in newswire texts.

- **Key Innovations:** Implemented advanced features like ALLCAP, lowercase, after symbol, number, and pretitle, significantly improving the model's accuracy and reliability.

Major Accomplishments



- **Improved Performance:** Demonstrated through rigorous training and testing, the enhanced model shows superior performance metrics compared to the baseline, particularly in precision and recall.

- **Practical Application:** Developed a user-friendly web application that showcases the model's real-world utility by allowing users to interactively test the NER system.

Thanks!

Does anyone have any questions?