12/11/2022

Continuous education social mining

Author: Abhikumar Patel

Georgian college
Student ID: 200486594

Abstract

In Natural Language Processing, the phrase "Named Entity," first coined by Grishman and Sundheim, is frequently used (NLP). The researchers were concentrating on the data extraction task, which involves taking unstructured text, such as newspaper articles, and extracting structured information on business operations and defense-related activities. Recognizing information aspects like place, person, organization, time, date, money, percent expression, etc. is crucial to understanding "Named Entity." Some academics refer to the process of recognizing these things in the unstructured text as "Named Entity Recognition" (NER). The Stanford Named Entity Recognizer (SNER), Illinois Named Entity Tagger (INET), the Alias-i LingPipe (LIPI), and OpenCalais are excellent tools to carry out this work now that NER technology is developed (OCWS). Each of these has various benefits and is made for particular types of data. Our ultimate objective for this term project is to create a NER module for the IDEAL project based on a specific NER tool, like SNER, in order to apply NER to the Twitter and web page data sets. This project report details our efforts toward achieving this objective, including a review of the relevant literature, the requirements, the algorithm, the development strategy, the system architecture, the implementation, the user manual, and the development manual. Results for several collections are also provided, along with discussion and future plans.

Overview

What is Named Entity Recognition? The technique of detecting and classifying words into various categories has grown in importance in the field of natural language processing (NLP), and it is known as named entity recognition. Recognizing and extracting such data is a fundamental task and a core process of NLP, mainly for two reasons First, NER is directly used in many applied research domains [10]. For example, proteins and genes can be considered as named entities, and many works in medicine focus on the analysis of scientific articles to find out hidden relationships between them, and drive experimental research. Second, NER is used to do preprocessing for more advanced NLP problems, such as relationship or information extraction.

Problem and challenges:

Despite the high F1 measure value found in the MUC-7 dataset, named entity recognition remains a challenging issue. The focus of current studies is on lowering the labor-intensive nature of annotation by utilizing semi-supervised learning [13], resilient performance across domains, and scaling up to fine-grained entity types. The crowdsourcing field has received a lot of attention over the last several years, and it offers a potential way to get high-quality aggregate human assessments for supervised and semi-supervised machine learning approaches to NER. A different intriguing challenge is to find "significant terms" in text and cross-link them to Wikipedia, which can be considered as an example of highly fine-grained named entity recognition, with the types being the actual Wikipedia articles explaining potential ambiguous concepts. The major limitation comes in when training for large datasets and still not much is available to scale the NER to large datasets.

Our Goal:

General Objective:

There have been numerous studies on NER, as seen here. However, these works each have significant drawbacks as well, such as a strong performance on some datasets but failing miserably on others. Therefore, we question and enhance past efforts on various data sets, specifically Twitter data sets where the writing is informal and doesn't pay attention to grammar and syntax.

Specific Objective:

Here the specific objective is to understand popular courses among social media followers of Georgian college.

In this term the project, our work contains three parts: 1. We extract the data from these social media platforms 2. We clean the dataset to remove noise 3. Annotating the custom entities 4. Applying the spacy model on the annotated data 5. Evaluating the predictions of the custom model.

Methodology:

Planning phase:

The planning phase encompasses all aspects of project and product management. This typically includes resource allocation, capacity planning, project scheduling, cost estimation, and provisioning.

In the capacity planning work for 8 hrs. per week and for a total of 12 weeks

In the project scheduling part started 19th September till date.

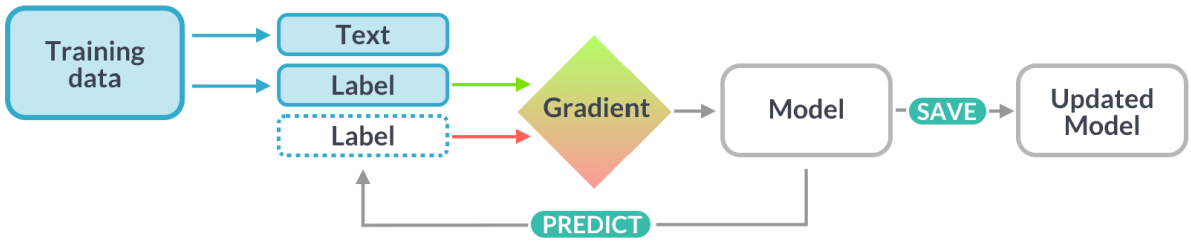
System Analysis

During the planning, I interacted with Ross to further understand the project requirement and the dataset preparation part. Once I had a clear picture of what all social media platforms, I can get the data as follows:

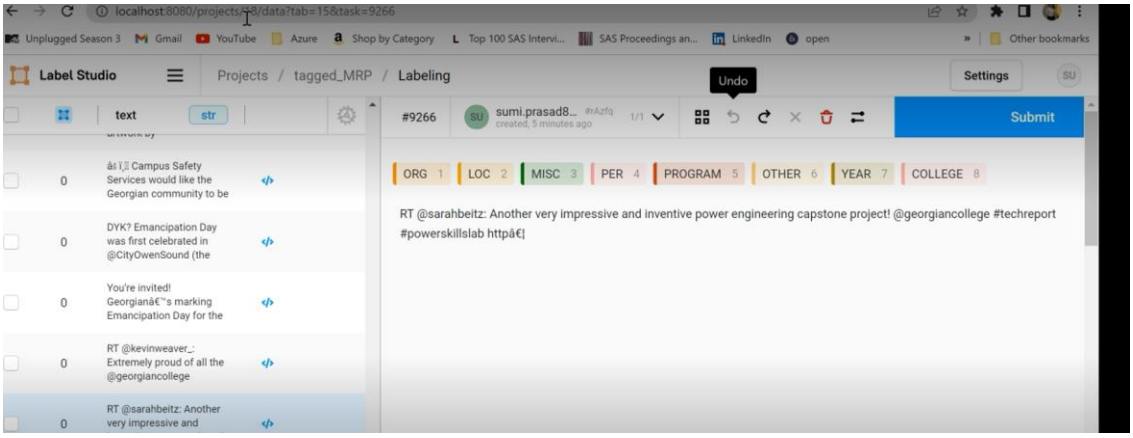
1) Scrap data from Twitter
2) Scrap data from Facebook

Scripting for both of this process was done and the final outcome CSV as attached with this project

System Design



open-source library-Label Studio for annotation:



System implementation:

- Based on the data scraped from Facebook and Twitter. The dataset was cleaned to remove the noise present in the dataset. we used label studio(https://labelstud.io/) to annotate our cleaned data. The outcome Json is attached to this project and can be seen below:

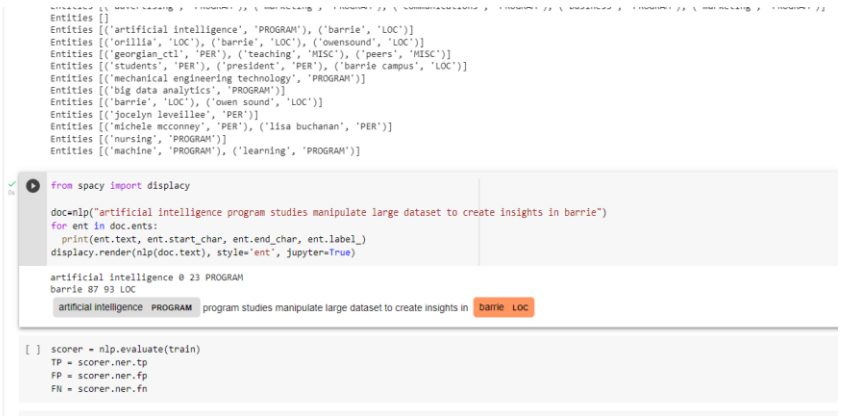
[{"id":607,"annotations":[{"id":114,"completed_by":1,"result":[{"value":{"start":3,"end":16,"text":"southsimcoeps","labels":["PER"]},"id":"7Npvi5na55","from_name":"label",
"to_name":"text","type":"labels","origin":"manual"}],"was_cancelled":false,"ground_truth":false,"created_at":"2022-11-28T23:51:01.042005Z","updated_at":"2022-11-28T23:51:01.042045Z",
"lead_time":6.13,"prediction":{},"result_count":0,"task":607,"parent_prediction":null,"parent_annotation":null}],"file_upload":"3b4003ee-Text-data-for-label-studio.csv","drafts":[],
"predictions":[],"data":{"Unnamed: 0":606,"text":"rt southsimcoeps great see constable rahiman profiled way go wazeer sspsfamily proud httpstcosrsuac"},"meta":{},
"created_at":"2022-11-23T03:43:44.591781Z","updated_at":"2022-11-28T23:51:01.110081Z","inner_id":607,"total_annotations":1,"cancelled_annotations":0,"total_predictions":0,
"comment_count":0,"unresolved_comment_count":0,"last_comment_updated_at":null,"project":3,"updated_by":1,"comment_authors":[]},{"id":590,"annotations":[{"id":113,"completed_by":1,
"result":[{"value":{"start":9,"end":14,"text":"riley","labels":["PER"]},"id":"pHBIy_IErr","from_name":"label","to_name":"text","type":"labels","origin":"manual"}],
"was_cancelled":false,"ground_truth":false,"created_at":"2022-11-28T23:50:13.837478Z","updated_at":"2022-11-28T23:50:13.837524Z","lead_time":2.927,"prediction":{},"result_count":0,
"task":590,"parent_prediction":null,"parent_annotation":null}],"file_upload":"3b4003ee-Text-data-for-label-studio.csv","drafts":[],"predictions":[],"data":{"Unnamed: 0":589,
"text":"congrats riley httpstcotcamuyaej"},"meta":{},"created_at":"2022-11-23T03:43:44.590529Z","updated_at":"2022-11-28T23:50:13.906170Z","inner_id":590,"total_annotations":1,
"cancelled_annotations":0,"total_predictions":0,"comment_count":0,"unresolved_comment_count":0,"last_comment_updated_at":null,"project":3,"updated_by":1,"comment_authors":[]},
{"id":589,"annotations":[{"id":112,"completed_by":1,"result":[{"value":{"start":3,"end":18,"text":"michele_newton_","labels":["PER"]},"id":"zSxE1tjOVU","from_name":"label",
"to_name":"text","type":"labels","origin":"manual"}],"was_cancelled":false,"ground_truth":false,"created_at":"2022-11-28T23:50:08.377282Z","updated_at":"2022-11-28T23:50:08.377319Z",
"lead_time":3.102,"prediction":{},"result_count":0,"task":589,"parent_prediction":null,"parent_annotation":null}],"file_upload":"3b4003ee-Text-data-for-label-studio.csv","drafts":[],
"predictions":[],"data":{"Unnamed: 0":588,"text":"rt michele_newton_ today im presenting inclusive communications workshop communications team georgiancollege love wh"},"meta":{},
"created_at":"2022-11-23T03:43:44.590450Z","updated_at":"2022-11-28T23:50:08.449160Z","inner_id":589,"total_annotations":1,"cancelled_annotations":0,"total_predictions":0,
"comment_count":0,"unresolved_comment_count":0,"last_comment_updated_at":null,"project":3,"updated_by":1,"comment_authors":[]},{"id":588,"annotations":[{"id":111,"completed_by":1,
"result":[{"value":{"start":3,"end":14,"text":"bruce_power","labels":["PER"]},"id":"R79rqwHEbw","from_name":"label","to_name":"text","type":"labels","origin":"manual"}],
"was_cancelled":false,"ground_truth":false,"created_at":"2022-11-28T23:50:04.180206Z","updated_at":"2022-11-28T23:50:04.180238Z","lead_time":2.633,"prediction":{},"result_count":0,
"task":588,"parent_prediction":null,"parent_annotation":null}],"file_upload":"3b4003ee-Text-data-for-label-studio.csv","drafts":[],"predictions":[],"data":{"Unnamed: 0":587,
"text":"rt bruce_power interested career electrician bruce power partnering georgiancollege cusw offer free we"},"meta":{},"created_at":"2022-11-23T03:43:44.590376Z",
"updated_at":"2022-11-28T23:50:04.248621Z","inner_id":588,"total_annotations":1,"cancelled_annotations":0,"total_predictions":0,"comment_count":0,"unresolved_comment_count":0,
"last_comment_updated_at":null,"project":3,"updated_by":1,"comment_authors":[]},{"id":583,"annotations":[{"id":110,"completed_by":1,"result":[{"value":{"start":0,"end":11,
"text":"newell_jenn","labels":["PER"]},"id":"sEY2JMcyWa","from_name":"label","to_name":"text","type":"labels","origin":"manual"}],"was_cancelled":false,"ground_truth":false,
"created_at":"2022-11-28T23:49:48.420163Z","updated_at":"2022-11-28T23:49:48.420194Z","lead_time":3.091,"prediction":{},"result_count":0,"task":583,"parent_prediction":null,
"parent_annotation":null}],"file_upload":"3b4003ee-Text-data-for-label-studio.csv","drafts":[],"predictions":[],"data":{"Unnamed: 0":582,"text":"newell_jenn thanks sharing us "},
"meta":{},"created_at":"2022-11-23T03:43:44.590004Z","updated_at":"2022-11-28T23:49:48.487707Z","inner_id":583,"total_annotations":1,"cancelled_annotations":0,"total_predictions":0

The annotated entities were Program, LOC, PER, Misc

- Based on the annotation a custom Spacy model has trained over 200 iterations and the model was then tested over unseen test data.

System Testing:

Raw unseen data was passed on to model to predict the outcome from the model



Acceptance, Implementation, and deployment:

- The following is the confusion matrix generated for each entity. The results seem to be fine with the given small dataset.

```
scorer = nlp.evaluate(train)
for ent_type, scorer_ent_type in scorer.ner_per_ents.items():
    TP = scorer_ent_type.tp
    FP = scorer_ent_type.fp
    FN = scorer_ent_type.fn
    print('Ent_type:', ent_type, 'TP:', TP, 'FP:', FP, 'FN:', FN
```

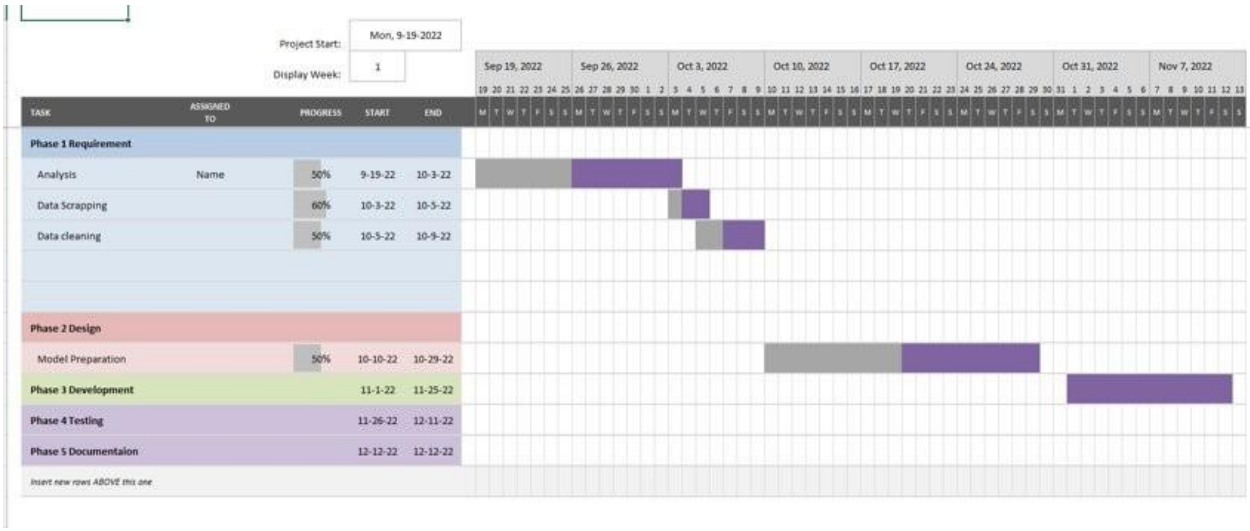
```
Ent_type: PROGRAM TP: 18 FP: 1 FN: 0
Ent_type: LOC TP: 10 FP: 0 FN: 0
Ent_type: PER TP: 6 FP: 0 FN: 0
Ent_type: PROJECT TP: 2 FP: 0 FN: 0
Ent_type: ORG TP: 0 FP: 0 FN: 1
```

- More data will be required to further do rigorous training and train the model on different variants of data.
- More data labels need to be annotated so that a pool of annotated data can be used in training.

Required Tools:

- For the extraction of the Twitter account a developer account was created with elevated access for the extraction
- Python was mainly used for scripting

Gantt chart:



Tools used:

Python: For scrapping script, Data Clensing and final project implementation

Label Studio: For annotation

NLTK/TF-IDF: For tokenization

Spacy/BERT entity model: For Information extraction language models

Excel: For data wranglings

Literature review:

We also came across a couple of excellent publications on NER. Nadeau and Sekine give a review of the NER technologies that are currently available. A chunk tagger for NER based on HMM was proposed by Zhou and Su. Ratinov and Roth looked into the typical difficulties and NER misunderstandings, along with some answers. INET, which they developed, is based on six different supervised learning techniques, including hidden Markov models (HMM), multilayered neural networks, and other statistical techniques. [4] introduces the technology behind the Stanford NER, which is based on linear chain conditional fields. The LIPI [7] system employs an n-gram character language model and is trained using conditional random field and HMM techniques. The most recent tool we are aware of is OCWS [8], a NER Web Service.

References

[1] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30.1 (2007): 3-26.

[2] Zhou, GuoDong, and Jian Su. "Named entity recognition using an HMM-based chunk tagger." Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002