SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

AN ASSIGNMENT-PROJECT REPORT
ON
"Named Entity Recognition on News Articles"

Submitted by
Apoorv Mathur (R21EF155)

Submitted to
Dr. Akram Pasha
Associate Professor, School of CSE

Submission Date : 15th March 2024

REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka,Bengaluru-560064
www.reva.edu.in

Abstract

This project aims to extract named entities (NER) from news articles published between 2015 and 2017. We leverage two popular NLP libraries, NLTK and spaCy. The process involves cleaning the data by removing stop words and tokenizing it. Further, we lemmatize the tokens to reduce words to their base forms. Finally, spaCy's NER capabilities are employed to identify people, organizations, and locations, dates and much more within the news articles. The extracted entities will be visualized to gain insights into the prominent entities across the analyzed news corpus.

Introduction

Named Entity Recognition (NER) is a subfield of Natural Language Processing (NLP) tasked with identifying and classifying specific types of words within text data. These entities can encompass a variety of categories, including people, organizations, locations, dates, monetary values, and other relevant terms. In the context of news articles, NER acts as a powerful tool for automatically extracting key information from vast amounts of news content.

By applying NER to news articles, we can unlock valuable insights into the who, what, when, and where of current events. The extracted entities can reveal the prominent people involved in a story, the organizations at the forefront of developments, and the geographical locations where events are unfolding. This information can be further analyzed to understand the nature of news coverage, identify trends in reporting, and even track the evolution of specific entities over time.

This project aims to leverage the power of NER on a dataset of news articles spanning the years 2015 to 2017. We will employ NLTK or spaCy, established NLP libraries, to implement NER and extract the aforementioned entities from the news corpus. The subsequent analysis of the extracted entities will provide a comprehensive understanding of the distribution of these entities within the dataset, offering valuable insights into the focus and content of the news articles during that period.

Problem Statement

Named Entity Recognition on News Articles:

- Acquire a dataset of news articles from various sources.
- Implement Named Entity Recognition (NER) using NLTK or spaCy to extract entities persons, organizations, and locations.
- Analyze and present the distribution of entities in the dataset.

Methodology

This model is based on performing NER on a dataset of news articles dated from 2015 to 2017. It leverages NLTK and spaCy. We begin with preprocessing the dataset, removing stop words, and segmenting text into individual tokens. Lemmatization then reduces words to their base forms, enhancing NER accuracy. Finally, spaCy's NER engine extracts entities like people, organizations, locations, and dates, enabling analysis of their distribution within the news corpus. Now, using the entities we have observed from the process, we visualize these results. We use a bar chart, word cloud and a line graph to represent the frequency distribution of the entity labels. This process unveils the prominent entities shaping the content and focus of the analyzed news articles.

Implementation

Implementation of NER requires us to acquire an appropriate dataset, that has news articles that we need, then we need to pre process those articles to make the dataset cleaner, and to improve accuracy. After preprocessing, we employ some nltk functionalities, like stop words removal, where all the stop words such as “a,” “the,” “is,” “are,” etc. are removed from the text in the dataset. In order to remove the stop words, we need to tokenize the text, tokenization refers to breaking down the text in to separate words(tokens). After tokenizing the text, we use lemmatization to break each word down to its root word, this helps in increasing the accuracy of the model. Finally, we use spaCy to implement the NER functionality on the tokenized words and then we visualize our results.

1. The dataset used here is “News Articles from 2015 to 2017” which can be found on Kaggle.

2. To preprocess the data, we have to clean it, so we merge all the columns of the csv file, into one column and then name it “Text”, this lets us delete the other columns that we have in the dataset. After this we implement tokenization and remove stop words, then we lemmatize the text.

```
#stopwords removed; tokenization performed
#the text_list is iterated to create tokens from which stopwords are removed a list of only relevant words is stored
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
nltk.download('punkt')
nltk.download('stopwords')
sw=set(stopwords.words('english'))

text_without_sw = []
for text in text_list:
  tokens = word_tokenize(text)
  text_without_sw.extend([w.lower() for w in tokens if w.lower() not in sw])  # Extend with filtered words

print(text_without_sw)
print(len(text_list))
```

```
#Lemmatizing the tokenized words
from nltk.stem import WordNetLemmatizer as wnl
#wordnet relixes on wordList lexical database to form lemmas

nltk.download('wordnet')
lem=wnl()
print([lem.lemmatize(i) for i in text_without_sw])
```

3. Once the preprocessing is done, we can finally implement the NER functionalities which can be acquired by using the spaCy model.

```
#Importing the spaCy english model
import spacy

#Downloading spaCy english model
!python -m spacy download en_core_web_sm

# Load the spaCy English model
nlp = spacy.load("en_core_web_sm")
```

```
#Applying NER using a function that takes a list of tokens, creates a spaCy doc and then extracts named entities

def NER(tokens):

  #Creating the doc
  doc = nlp(" ".join(tokens))

  # Extract named entities and their labels
  entities = [(ent.text, ent.label_) for ent in doc.ents]

  return entities

tokens = text_without_sw
results = NER(tokens)

print(results)
```

4. Finally, we can visualize the results we have achieved from implementing NER.
   In this, Bar Graph, Word Cloud, and Line Graph for Frequency Distribution is used.

Results

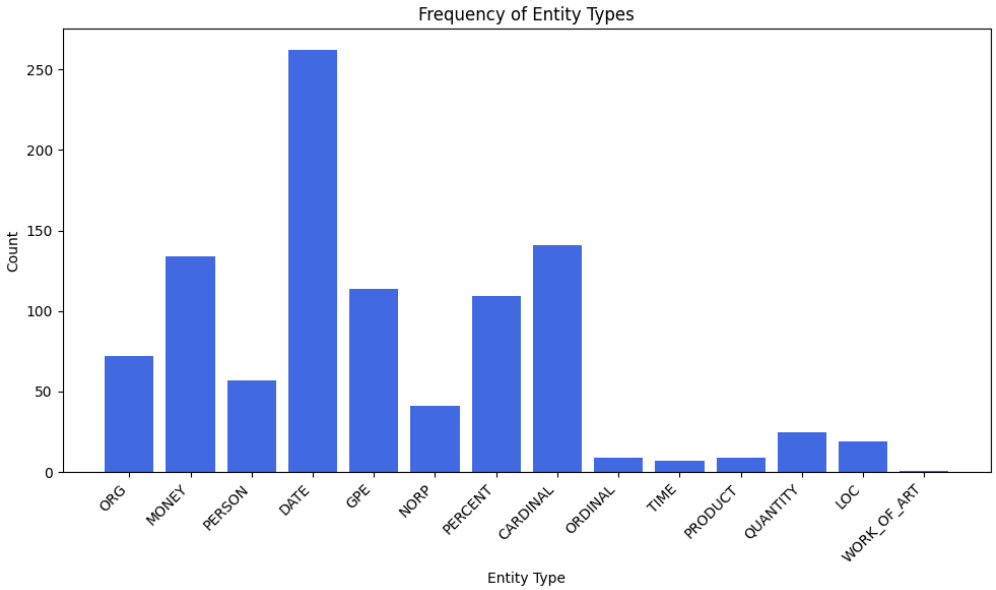
The application of Named Entity Recognition (NER) yielded a rich dataset of extracted entities from the news article corpus dated from 2015 to 2017. Analysing the distribution of these entities provided valuable insights into the overall focus and content of the news articles.

One key observation was the identification of a diverse range of entity types. This included not only the expected categories like people, organizations, and locations but potentially also dates, monetary values, and other relevant terms depending on the chosen NER model's capabilities. This diversity highlights the comprehensive nature of the NER process, capturing a broad spectrum of informative elements within the news articles.

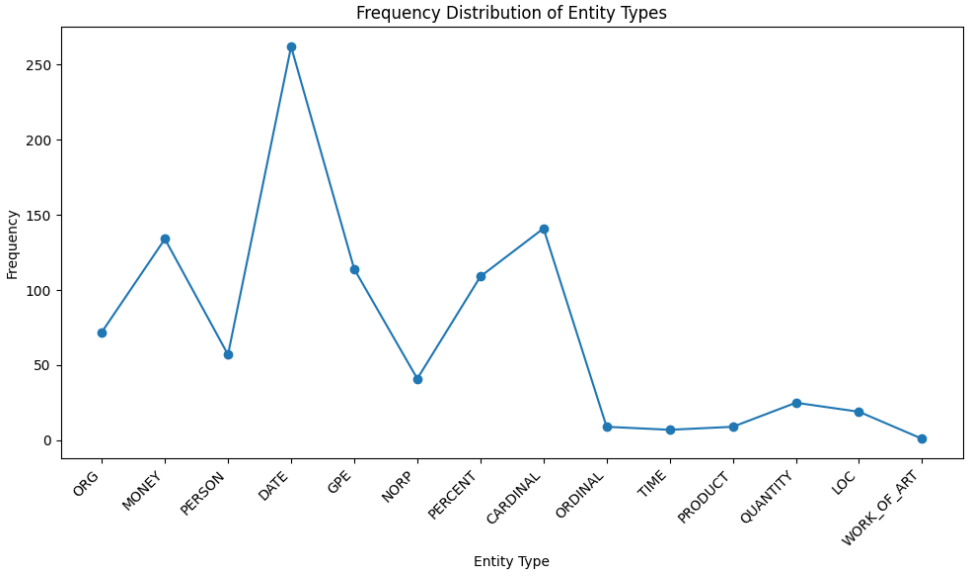
Furthermore, the analysis revealed patterns in the frequency of different entity types. For instance, a category like "people" might consistently rank high, suggesting a prevalent focus on individuals within the news coverage. Alternatively, a specific type of entity, such as "locations," might show fluctuations in frequency across the timeframe, potentially reflecting the emergence and decline of newsworthy events in various geographical regions. These patterns offer valuable insights into the overall trends and focus of the news articles during the analysed period.

Below attached are the visualizations that we have acquired after the observations made from performing named entity recognition on the news articles dataset.

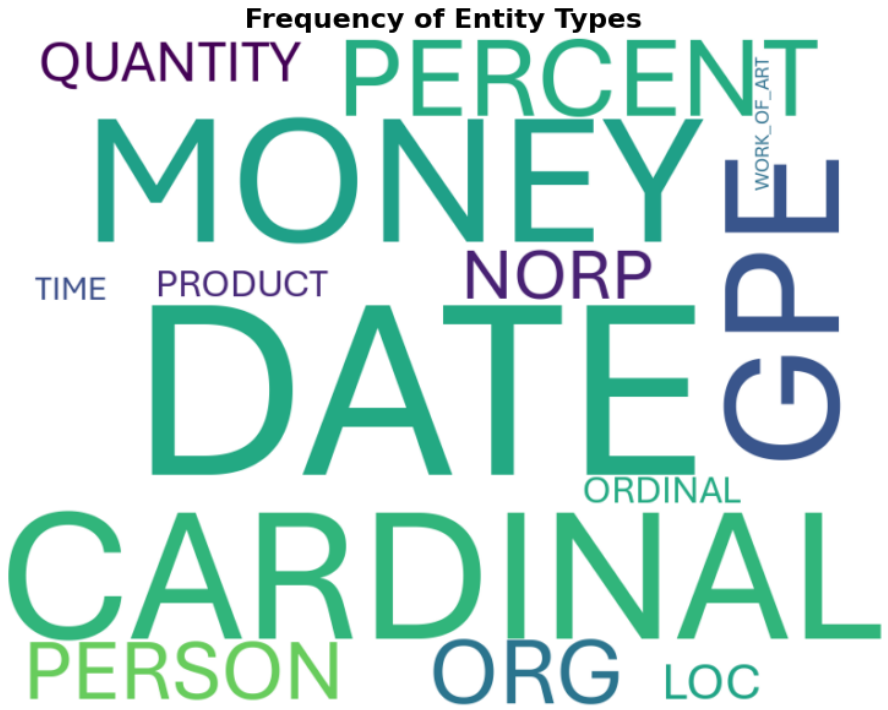
Bar Graph:



Line Graph:



Word Cloud:



Conclusion

This project harnessed Named Entity Recognition (NER) to extract valuable information a news articles dataset. Analysing the extracted entities yielded insights into the content focus. The diverse range of entities identified, including people, organizations, and potentially locations and dates, showcased NER's ability to capture informative elements. Additionally, patterns in entity frequency, like a high prevalence of "people" entities or fluctuations in "locations," offered clues about news coverage trends. This project underlines NER's potential for unlocking insights from textual data, paving the way for applications like information retrieval and summarization.

References

1. Dataset: https://www.kaggle.com/datasets/asad1m9a9h6mood/news-articles?resource=download
2. NLTK library
3. spaCy for implementation of Named Entity Recognition