Assignment 2.1: NLP, NER and PoS Tagging

Mostafa Zamaniturk

Instructions In this assignment, you will apply Natural Language Processing (NLP), Named Entity Recognition (NER), and Part-of-Speech (PoS) techniques in NLP to analyze the Climate Fever dataset. The dataset contains climate change-related articles, and your task is to extract named entities and assign PoS tags to different parts of speech in the text.

Instructions

In this assignment, you will apply Natural Language Processing (NLP), Named Entity Recognition (NER), and Part-of-Speech (PoS) techniques in NLP to analyze the Climate Fever dataset. The dataset contains climate change-related articles, and your task is to extract named entities and assign PoS tags to different parts of speech in the text.

```
# Import needed libraries
import pandas as pd
import matplotlib.pyplot as plt
import os
```

Required Details

1- Apply NER techniques to identify named entities (such as persons, organizations, locations, etc.) within the text.

```
import kagglehub

# Download dataset
path = kagglehub.dataset_download("bouweceunen/climate-fever-dataset")

print("Path to dataset files:", path)
```

```
Using Colab cache for faster access to the 'climate-fever-dataset'
dataset.
Path to dataset files: /kaggle/input/climate-fever-dataset
```

```
# List files in the downloaded dataset folder
os.listdir(path)

# Suppose the CSV file is named 'climate-fever.csv'
```

```
df = pd.read_csv(os.path.join(path, 'climate-fever.csv'))
df.head()
```

```
{"type":"dataframe","variable_name":"df"}
```

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

# Download resources (run once)
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('punkt_tab') # Download punkt_tab resource


# define a function to do all tokenization, Lowercasing and removing
stopwords:
def preprocess_text(text):

    # Handle missing or non-string values
    if not isinstance(text, str):
        return []

    # Tokenization
    tokens = word_tokenize(text)

    # Lowercasing
    tokens = [word.lower() for word in tokens]

    # Remove stopwords and keep only alphabetic words
    stop_words = set(stopwords.words('english'))
    filtered_tokens = [word for word in tokens if word.isalpha() and
word not in stop_words]

    return filtered_tokens

# Apply only to text (string/object) columns, keep originals intact
for col in df.select_dtypes(include=["object"]).columns:
    df[f"{col}_processed"] = df[col].apply(preprocess_text)

# Show a preview
print(df.head())
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
```

```
   claim_id                                              claim
claim_label  \
0         0  Global warming is driving polar bears toward e...
SUPPORTS
1         5  The sun has gone into ‘lockdown’ which could c...
SUPPORTS
2         6        The polar bear population has been growing.
REFUTES
3         9  Ironic' study finds more CO2 has slightly cool...
REFUTES
4        10  Human additions of CO2 are in the margin of er...
REFUTES

                    evidences/0/evidence_id evidences/0/evidence_label
\
0   Extinction risk from global warming:170            NOT_ENOUGH_INFO

1                                Famine:386                    SUPPORTS

2                           Polar bear:1332            NOT_ENOUGH_INFO

3                    Atmosphere of Mars:131            NOT_ENOUGH_INFO

4  Carbon dioxide in Earth's atmosphere:140            NOT_ENOUGH_INFO


                    evidences/0/article  \
0   Extinction risk from global warming
1                                Famine
2                           Polar bear
3                    Atmosphere of Mars
4  Carbon dioxide in Earth's atmosphere

                                  evidences/0/evidence
evidences/0/entropy  \
0  "Recent Research Shows Human Activity Driving ...
0.693147
1  The current consensus of the scientific commun...
0.000000
2  "Ask the experts: Are polar bear populations i...
0.693147
3  CO2 in the mesosphere acts as a cooling agent ...
0.693147
4  While CO 2 absorption and release is always ha...
0.693147

  evidences/0/votes/0 evidences/0/votes/1  ...
evidences/3/votes/4_processed  \
0            SUPPORTS     NOT_ENOUGH_INFO  ...
[]
1            SUPPORTS            SUPPORTS  ...
```

```
[]
2     NOT_ENOUGH_INFO              REFUTES  ...
[]
3     NOT_ENOUGH_INFO             SUPPORTS  ...
[]
4     NOT_ENOUGH_INFO              REFUTES  ...
[]

  evidences/4/evidence_id_processed
evidences/4/evidence_label_processed  \
0                            [polar]
[]
1                                 []
[]
2                            [polar]
[refutes]
3                           [carbon]
[]
4                                 []
[refutes]

  evidences/4/article_processed  \
0                 [polar, bear]
1                      [winter]
2                 [polar, bear]
3             [carbon, dioxide]
4                         [sea]

                       evidences/4/evidence_processed  \
0   [bear, hunting, caught, global, warming, debate]
1  [many, regions, winter, associated, snow, free...
2  [recognized, polar, bear, subpopulations, one,...
3  [less, energy, reaches, upper, atmosphere, the...
4  [recently, anthropogenic, activities, steadily...

  evidences/4/votes/0_processed evidences/4/votes/1_processed  \
0                    [supports]                            []
1                     [refutes]                            []
2                     [refutes]                     [refutes]
3                            []                            []
4                     [refutes]                     [refutes]

   evidences/4/votes/2_processed evidences/4/votes/3_processed  \
0                             []                            []
1                             []                            []
2                             []                            []
3                             []                            []
4                             []                            []

  evidences/4/votes/4_processed
```

```
0                                  []
1                                  []
2                                  []
3                                  []
4                                  []

[5 rows x 100 columns]
```

```
# Converts all columns in the pandas DataFrame df to the data type str
(string)
df = df.astype(str)

import spacy
from spacy import displacy

# Load the English language model
nlp = spacy.load('en_core_web_sm')

# Define a function for NER
def extract_entities(text):
    if not isinstance(text, str):
        return []
    doc = nlp(text)
    return [(ent.text, ent.label_) for ent in doc.ents]

# List of columns for NER process
text_columns = ['claim']  # replace with or add desired columns

# Apply NER column by column
for col in text_columns:
    df[col + '_entities'] =
df[col].astype(str).apply(extract_entities)

# Show original columns and their corresponding _entities columns
cols_to_show = text_columns + [col + '_entities' for col in
text_columns]
print(df[cols_to_show].head(10))
```

```
                                               claim  \
0  Global warming is driving polar bears toward e...
1  The sun has gone into ‘lockdown’ which could c...
2        The polar bear population has been growing.
3  Ironic' study finds more CO2 has slightly cool...
4  Human additions of CO2 are in the margin of er...
5  They tell us that we are the primary forces co...
6  The Great Barrier Reef is experiencing the mos...
7  it’s not a pollutant that threatens human civi...
8  If CO2 was so terrible for the planet, then in...
9  Sea level rise has been slow and a constant, p...

                                      claim_entities
```

```
0                                                    []
1                                                    []
2                                                    []
3                        [(Ironic, ORG), (CO2, PRODUCT)]
4  [(CO2, PRODUCT), (CO2, PRODUCT), (the last ice...
5                                                    []
6                                                    []
7                                                    []
8                        [(CO2, PERSON), (CO2, PRODUCT)]
9                                                    []
```

2- Implement PoS tagging to assign appropriate parts of speech to different words in the text.

Analyze the results and provide insights on the named entities and their corresponding parts of speech in the Climate Fever dataset.

```
# Define a function for POS tagging
def extract_pos(text):
    if not isinstance(text, str):
        return []
    doc = nlp(text)
    # Return a list of tuples (word, POS tag)
    return [(token.text, token.pos_) for token in doc]

# List of columns for POS process
text_columns = ['claim']  # add columns for POS purpose

# Apply POS tagging column by column
for col in text_columns:
    df[col + '_pos'] = df[col].astype(str).apply(extract_pos)

# Show original columns and their corresponding _pos columns
cols_to_show = text_columns + [col + '_pos' for col in text_columns]
print(df[cols_to_show].head(10))
```

```
                                               claim  \
0  Global warming is driving polar bears toward e...
1  The sun has gone into ‘lockdown’ which could c...
2        The polar bear population has been growing.
3  Ironic' study finds more CO2 has slightly cool...
4  Human additions of CO2 are in the margin of er...
5  They tell us that we are the primary forces co...
6  The Great Barrier Reef is experiencing the mos...
7  it’s not a pollutant that threatens human civi...
8  If CO2 was so terrible for the planet, then in...
9  Sea level rise has been slow and a constant, p...

                                           claim_pos
0  [(Global, ADJ), (warming, NOUN), (is, AUX), (d...
1  [(The, DET), (sun, NOUN), (has, AUX), (gone, V...
```

```
2  [(The, DET), (polar, ADJ), (bear, NOUN), (popu...
3  [(Ironic, ADJ), (', PUNCT), (study, NOUN), (fi...
4  [(Human, ADJ), (additions, NOUN), (of, ADP), (...
5  [(They, PRON), (tell, VERB), (us, PRON), (that...
6  [(The, DET), (Great, PROPN), (Barrier, PROPN),...
7  [(it, PRON), ('s, VERB), (not, PART), (a, DET)...
8  [(If, SCONJ), (CO2, NOUN), (was, AUX), (so, AD...
9  [(Sea, NOUN), (level, NOUN), (rise, NOUN), (ha...
```

```
# --- Visualization using displacy for a single example ---
# Choose a row to visualize (e.g., the 45th row)
sample_text = df.loc[45, 'claim']
doc = nlp(sample_text)

# Visualize Named Entities
displacy.render(doc, style='ent', jupyter=True)
```

```
<IPython.core.display.HTML object>
```

3- Visualize the findings using appropriate graphs, charts, or tables to enhance understanding.

```
# --- Visualization using displacy for POS tags ---
# Choose a row to visualize (e.g., the second row)
sample_text = df.loc[1, 'claim']
doc = nlp(sample_text)

# Visualize POS tags and dependencies
displacy.render(doc, style='dep', jupyter=True)
```

```
<IPython.core.display.HTML object>
```

4- Summarize your approach, the findings, and any challenges faced during the process.

Note: You are free to use any additional techniques or libraries to enhance the NER and PoS tagging tasks. Make sure to provide proper documentation and references for any external resources used.

Approach Summary

I started by using the Kaggle library to import the dataset into my workspace. To get familiar with the dataset, I first explored the CSV file in detail, reviewing how the information was arranged in rows and columns, and identifying the key features that would be relevant for Natural Language Processing (NLP) tasks. This initial inspection helped me understand the structure of the data and determine what preprocessing steps were necessary.

Next, I prepared the dataset for NER (Named Entity Recognition) and POS (Part-of-Speech) tagging. The preprocessing involved several steps in just one function:

Tokenization: splitting the text into smaller units (tokens) such as words or punctuation marks.

Lowercasing: converting all text to lowercase to ensure consistency and reduce redundancy caused by case differences.

Stop-word removal: filtering out common words (such as “the,” “is,” or “and”) that do not contribute significant meaning to the NLP models.

Once the dataset was cleaned and structured, I designed a workflow to selectively apply NER and POS tagging only to the columns most relevant to the task. This selective approach helped reduce unnecessary processing and improved efficiency.

Finally, to make the results more interpretable, I used spaCy’s displacy visualization tool. This allowed me to graphically illustrate how the text was parsed, showing entities, dependencies, and part-of-speech tags in a clear and visual format. These visualizations were useful for validating the correctness of preprocessing steps and understanding how the model interprets the data.

Findings and Challenges

- Python version compatibility: Different versions of Python worked inconsistently with certain libraries. To resolve this, I switched from Visual Studio to Google Colab, which provided a more stable and compatible environment.
- Tokenization and reusability: Preparing the dataset required building a reusable pipeline. To achieve this, I designed functions in an object-oriented manner, which made the code more modular and easier to update. Since NLP algorithms often need fine-tuning, this approach helps reduce the amount of rework when adjustments are required.
- Working with raw documents: One open question that came in my mind is how to handle documents from reference books or other sources. Specifically, how can these documents be preprocessed and converted into a CSV format so they can be more easily used in NLP tasks? While the course mainly works with ready-to-use datasets, handling raw text sources seems like a valuable skill.
- Efficiency in NER: Initially, I applied NER to the entire dataset, which was highly computationally time-consuming. To optimize performance, I created a list of target columns and applied NER only to those, which significantly reduced processing time. A key question here is whether this selective approach is considered a best practice for implementing NER.

Required Format

To prepare for assignment submission, convert your Jupyter Notebook to a single, clean PDF or HTML document file. Your deliverable should contain your implementations of the tasks above, as well as any additional comments or observations you may have. Please ensure the PDF or MS Word document displays the code and output appropriately.