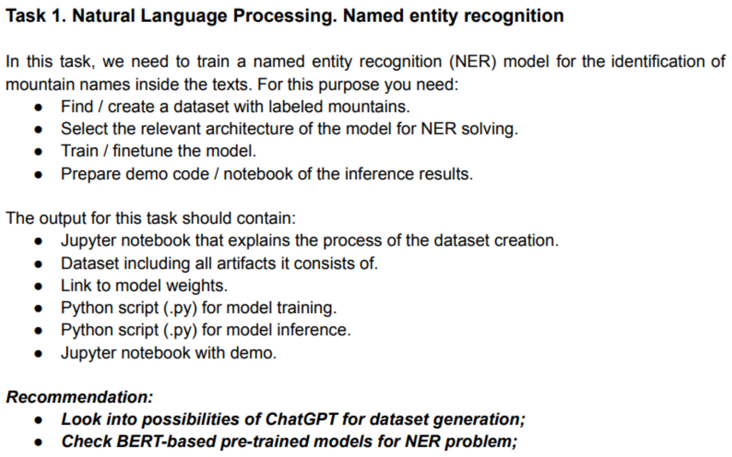


Task 1. Natural Language Processing. Named entity recognition

In this task, we need to train a named entity recognition (NER) model for the identification of mountain names inside the texts. For this purpose you need:

- Find / create a dataset with labeled mountains.
- Select the relevant architecture of the model for NER solving.
- Train / finetune the model.
- Prepare demo code / notebook of the inference results.

The output for this task should contain:

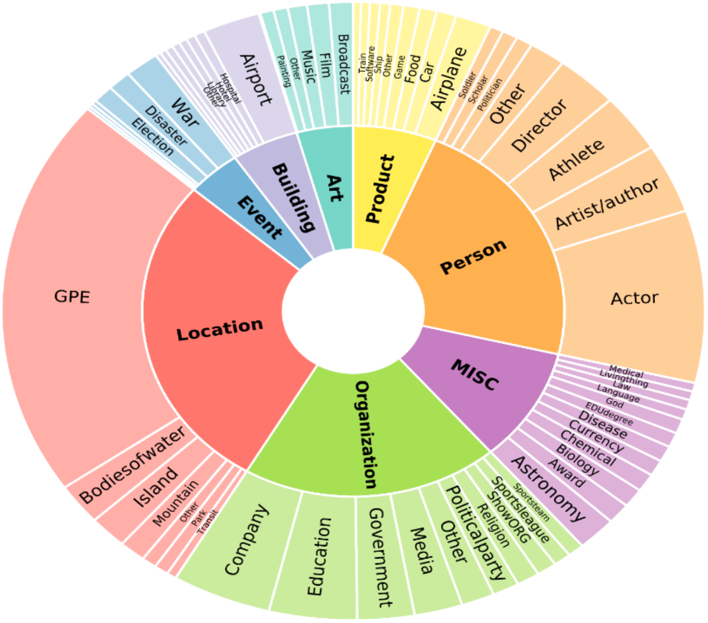
- Jupyter notebook that explains the process of the dataset creation.
- Dataset including all artifacts it consists of.
- Link to model weights.
- Python script (.py) for model training.
- Python script (.py) for model inference.
- Jupyter notebook with demo.

Recommendation:

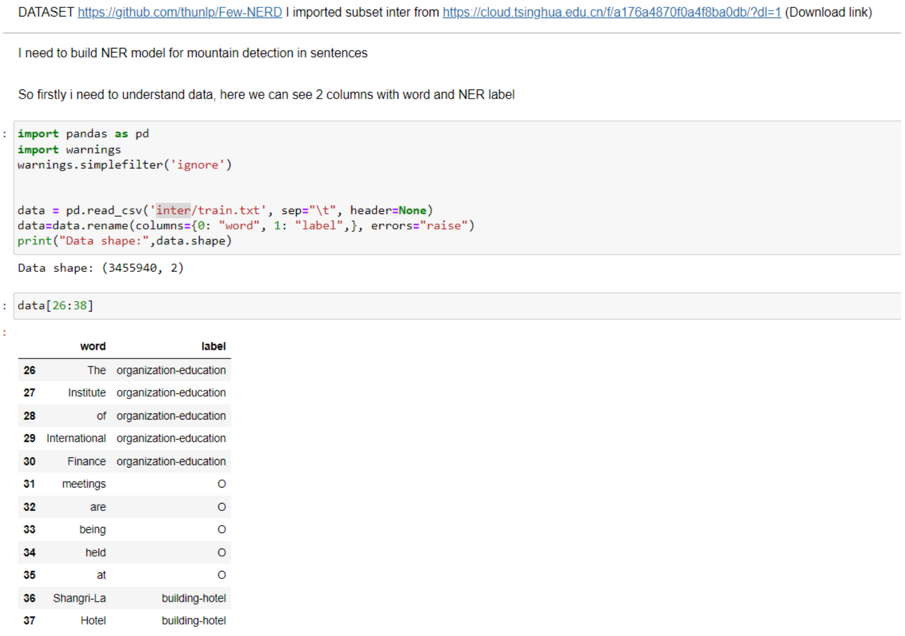
- ***Look into possibilities of ChatGPT for dataset generation;***
- ***Check BERT-based pre-trained models for NER problem;***

In my test task, I decided to use Few-NERD is a large-scale, fine-grained manually annotated named entity recognition dataset, which contains 8 coarse-grained types, 66 fine-grained types, 188,200 sentences, 491,711 entities and 4,601,223 tokens. I'm interested only in Location mountain to build NER for mountain identification.

The schema of Few-NERD is:



After downloading and opening, I need to look and understand what data do I have. Here we can see we have 2 columns word (with words in order) and labels for them.



DATASET https://github.com/thunlp/Few-NERD I imported subset inter from https://cloud.tsinghua.edu.cn/f/a176a4870f0a4f8ba0db/?dl=1 (Download link)

I need to build NER model for mountain detection in sentences

So firstly i need to understand data, here we can see 2 columns with word and NER label

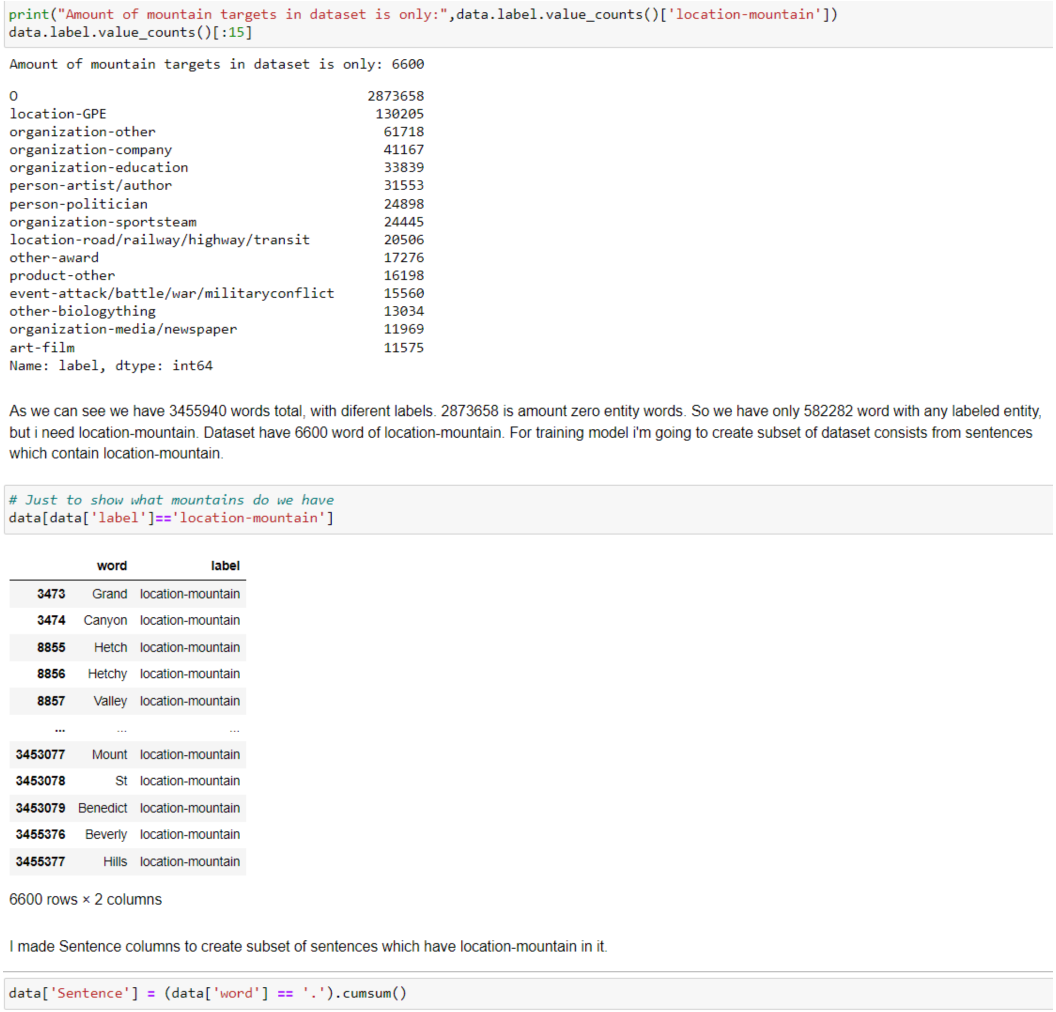
```
import pandas as pd
import warnings
warnings.simplefilter('ignore')


data = pd.read_csv('inter/train.txt', sep="\t", header=None)
data=data.rename(columns={0: "word", 1: "label",}, errors="raise")
print("Data shape:",data.shape)
```

```
Data shape: (3455940, 2)
```

```
data[26:38]
```

	word	label
26	The	organization-education
27	Institute	organization-education
28	of	organization-education
29	International	organization-education
30	Finance	organization-education
31	meetings	O
32	are	O
33	being	O
34	held	O
35	at	O
36	Shangri-La	building-hotel
37	Hotel	building-hotel

Then I grouped word by sentences by using “.” Symbol and pandas .cumsum()



```
print("Amount of mountain targets in dataset is only:",data.label.value_counts()['location-mountain'])
data.label.value_counts()[:15]
```

```
Amount of mountain targets in dataset is only: 6600

O                                             2873658
location-GPE                                   130205
organization-other                              61718
organization-company                            41167
organization-education                          33839
person-artist/author                            31553
person-politician                               24898
organization-sportsteam                         24445
location-road/railway/highway/transit           20506
other-award                                     17276
product-other                                   16198
event-attack/battle/war/militaryconflict        15560
other-biologything                              13034
organization-media/newspaper                    11969
art-film                                        11575
Name: label, dtype: int64
```

As we can see we have 3455940 words total, with diferent labels. 2873658 is amount zero entity words. So we have only 582282 word with any labeled entity, but i need location-mountain. Dataset have 6600 word of location-mountain. For training model i'm going to create subset of dataset consists from sentences which contain location-mountain.

```
# Just to show what mountains do we have
data[data['label']=='location-mountain']
```

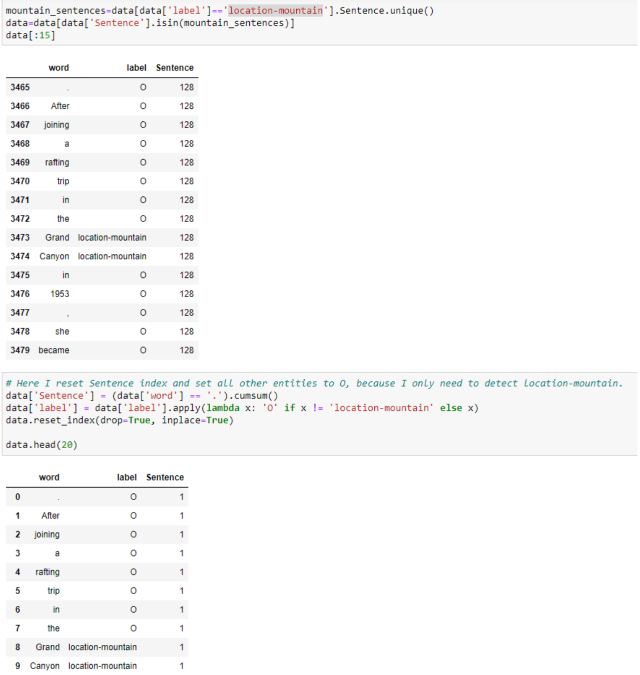
	word	label
3473	Grand	location-mountain
3474	Canyon	location-mountain
8855	Hetch	location-mountain
8856	Hetchy	location-mountain
8857	Valley	location-mountain
...	...	...
3453077	Mount	location-mountain
3453078	St	location-mountain
3453079	Benedict	location-mountain
3455376	Beverly	location-mountain
3455377	Hills	location-mountain

6600 rows × 2 columns

I made Sentence columns to create subset of sentences which have location-mountain in it.

```
data['Sentence'] = (data['word'] == '.').cumsum()
```

This solution is not working perfect, because sentences starts from “.”, but its any solve my problem. Also I deleted other entity classes, because I only need to detects mountains.



```
mountain_sentences=data[data['label']=='location-mountain'].Sentence.unique()
data=data[data['Sentence'].isin(mountain_sentences)]
data[:15]
```

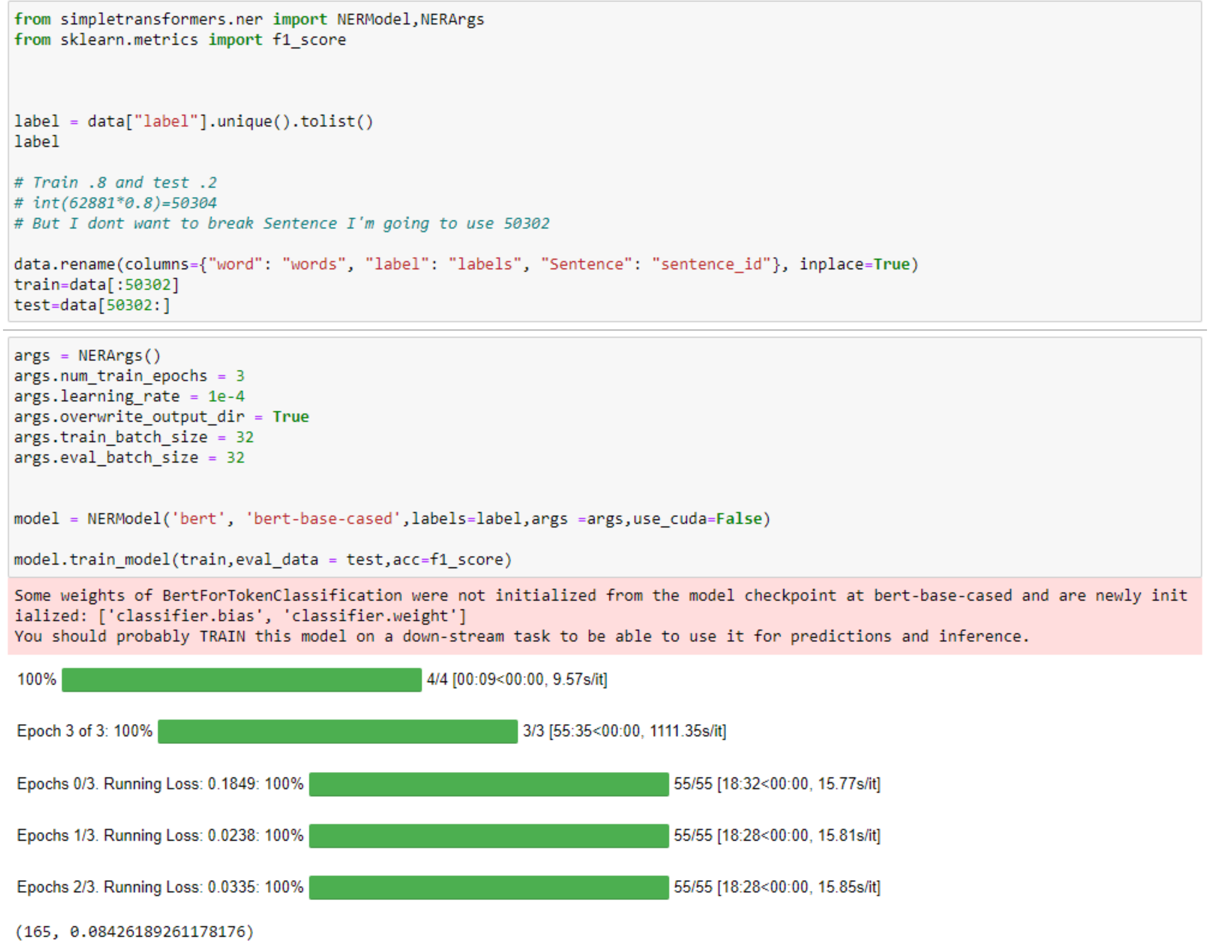
	word	label	Sentence
3465	.	O	128
3466	After	O	128
3467	joining	O	128
3468	a	O	128
3469	rafting	O	128
3470	trip	O	128
3471	in	O	128
3472	the	O	128
3473	Grand	location-mountain	128
3474	Canyon	location-mountain	128
3475	in	O	128
3476	1953	O	128
3477	,	O	128
3478	she	O	128
3479	became	O	128

```
# Here I reset Sentence index and set all other entities to O, because I only need to detect location-mountain.
data['Sentence'] = (data['word'] == '.').cumsum()
data['label'] = data['label'].apply(lambda x: 'O' if x != 'location-mountain' else x)
data.reset_index(drop=True, inplace=True)

data.head(20)
```

	word	label	Sentence
0	.	O	1
1	After	O	1
2	joining	O	1
3	a	O	1
4	rafting	O	1
5	trip	O	1
6	in	O	1
7	the	O	1
8	Grand	location-mountain	1
9	Canyon	location-mountain	1

Bert based NER was chosen for this task, I used simpletransformers library to fine-tune model, main metric is f1-score, because target is only 11% of total amount of data. Before it, I made train/test split and renamed columns.

(unfortunately I couldnt export weights from model)



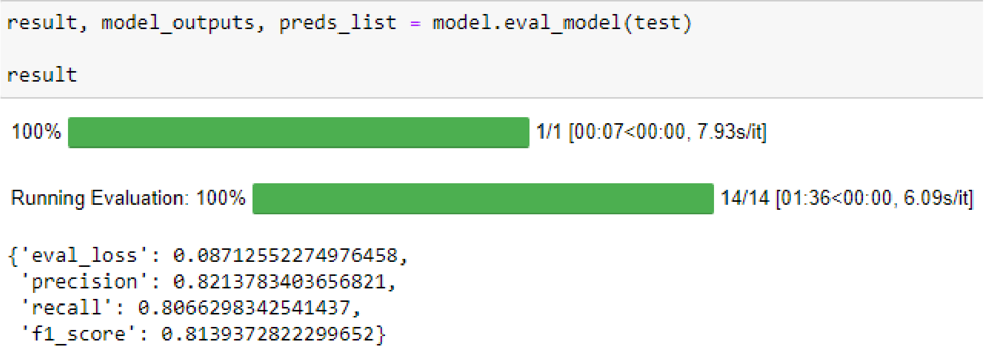
```
from simpletransformers.ner import NERModel,NERArgs
from sklearn.metrics import f1_score


label = data["label"].unique().tolist()
label

# Train .8 and test .2
# int(62881*0.8)=50304
# But I dont want to break Sentence I'm going to use 50302

data.rename(columns={"word": "words", "label": "labels", "Sentence": "sentence_id"}, inplace=True)
train=data[:50302]
test=data[50302:]
```

```
args = NERArgs()
args.num_train_epochs = 3
args.learning_rate = 1e-4
args.overwrite_output_dir = True
args.train_batch_size = 32
args.eval_batch_size = 32


model = NERModel('bert', 'bert-base-cased',labels=label,args =args,use_cuda=False)

model.train_model(train,eval_data = test,acc=f1_score)
```

```
Some weights of BertForTokenClassification were not initialized from the model checkpoint at bert-base-cased and are newly init
ialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

```
100% 4/4 [00:09<00:00, 9.57s/it]

Epoch 3 of 3: 100% 3/3 [55:35<00:00, 1111.35s/it]

Epochs 0/3. Running Loss: 0.1849: 100% 55/55 [18:32<00:00, 15.77s/it]

Epochs 1/3. Running Loss: 0.0238: 100% 55/55 [18:28<00:00, 15.81s/it]

Epochs 2/3. Running Loss: 0.0335: 100% 55/55 [18:28<00:00, 15.85s/it]

(165, 0.08426189261178176)
```

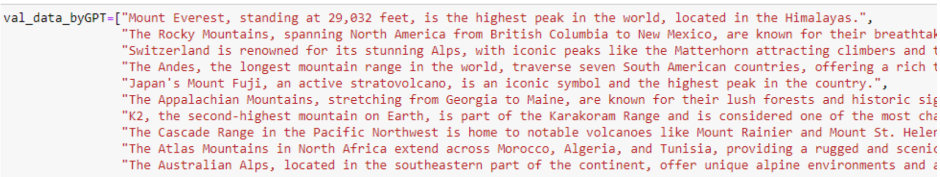
Model performance



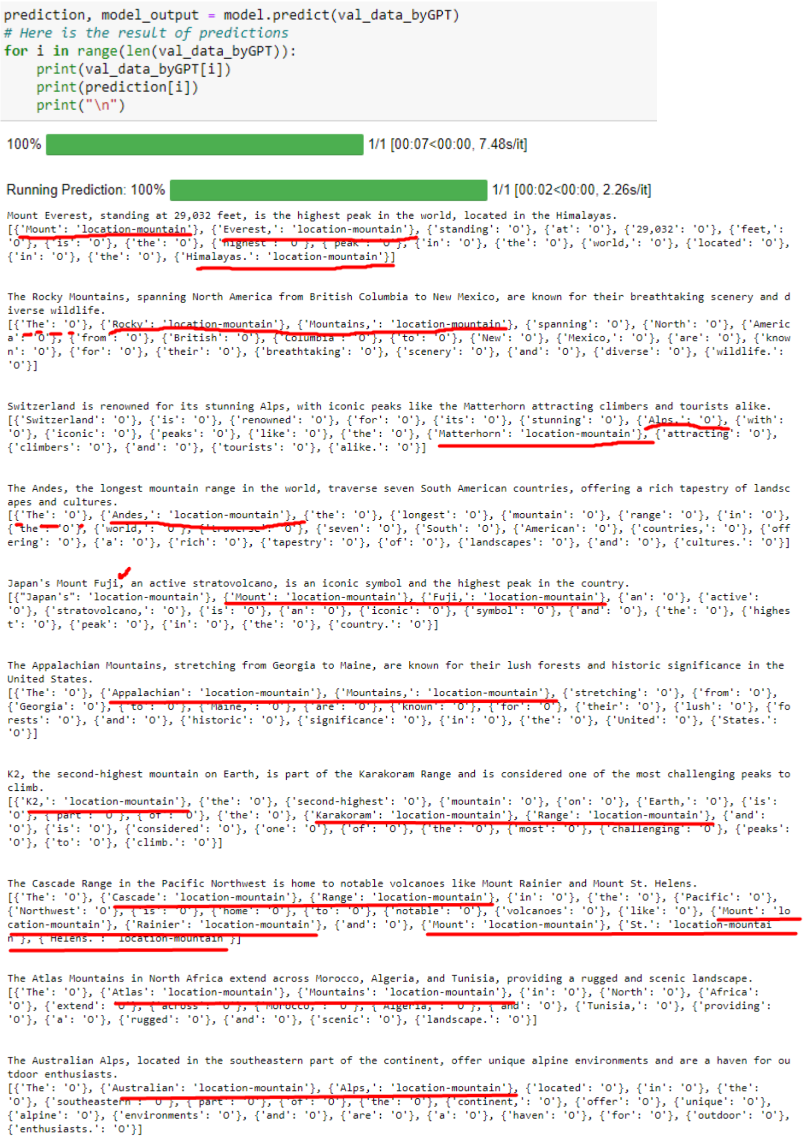
```
result, model_outputs, preds_list = model.eval_model(test)

result
```

```
100% 1/1 [00:07<00:00, 7.93s/it]

Running Evaluation: 100% 14/14 [01:36<00:00, 6.09s/it]

{'eval_loss': 0.08712552274976458,
 'precision': 0.8213783403656821,
 'recall': 0.8066298342541437,
 'f1_score': 0.8139372822299652}
```

To evaluate model I also ask to ChatGPT to create sentences with mountains



```
val_data_byGPT=["Mount Everest, standing at 29,032 feet, is the highest peak in the world, located in the Himalayas.",
                "The Rocky Mountains, spanning North America from British Columbia to New Mexico, are known for their breathtak
                "Switzerland is renowned for its stunning Alps, with iconic peaks like the Matterhorn attracting climbers and t
                "The Andes, the longest mountain range in the world, traverse seven South American countries, offering a rich t
                "Japan's Mount Fuji, an active stratovolcano, is an iconic symbol and the highest peak in the country.",
                "The Appalachian Mountains, stretching from Georgia to Maine, are known for their lush forests and historic sig
                "K2, the second-highest mountain on Earth, is part of the Karakoram Range and is considered one of the most cha
                "The Cascade Range in the Pacific Northwest is home to notable volcanoes like Mount Rainier and Mount St. Helen
                "The Atlas Mountains in North Africa extend across Morocco, Algeria, and Tunisia, providing a rugged and scenic
                "The Australian Alps, located in the southeastern part of the continent, offer unique alpine environments and a
```



```python
prediction, model_output = model.predict(val_data_byGPT)
# Here is the result of predictions
for i in range(len(val_data_byGPT)):
    print(val_data_byGPT[i])
    print(prediction[i])
    print("\n")
```

```
100% 1/1 [00:07<00:00, 7.48s/it]

Running Prediction: 100% 1/1 [00:02<00:00, 2.26s/it]
Mount Everest, standing at 29,032 feet, is the highest peak in the world, located in the Himalayas.
[{'Mount': 'location-mountain'}, {'Everest,': 'location-mountain'}, {'standing': 'O'}, {'at': 'O'}, {'29,032': 'O'}, {'feet,':
'O'}, {'is': 'O'}, {'the': 'O'}, {'highest': 'O'}, {'peak': 'O'}, {'in': 'O'}, {'the': 'O'}, {'world,': 'O'}, {'located': 'O'},
{'in': 'O'}, {'the': 'O'}, {'Himalayas.': 'location-mountain'}]


The Rocky Mountains, spanning North America from British Columbia to New Mexico, are known for their breathtaking scenery and d
iverse wildlife.
[{'The': 'O'}, {'Rocky': 'location-mountain'}, {'Mountains,': 'location-mountain'}, {'spanning': 'O'}, {'North': 'O'}, {'Americ
a': 'O'}, {'from': 'O'}, {'British': 'O'}, {'Columbia': 'O'}, {'to': 'O'}, {'New': 'O'}, {'Mexico,': 'O'}, {'are': 'O'}, {'know
n': 'O'}, {'for': 'O'}, {'their': 'O'}, {'breathtaking': 'O'}, {'scenery': 'O'}, {'and': 'O'}, {'diverse': 'O'}, {'wildlife.':
'O'}]


Switzerland is renowned for its stunning Alps, with iconic peaks like the Matterhorn attracting climbers and tourists alike.
[{'Switzerland': 'O'}, {'is': 'O'}, {'renowned': 'O'}, {'for': 'O'}, {'its': 'O'}, {'stunning': 'O'}, {'Alps,': 'O'}, {'with':
'O'}, {'iconic': 'O'}, {'peaks': 'O'}, {'like': 'O'}, {'the': 'O'}, {'Matterhorn': 'location-mountain'}, {'attracting': 'O'},
{'climbers': 'O'}, {'and': 'O'}, {'tourists': 'O'}, {'alike.': 'O'}]


The Andes, the longest mountain range in the world, traverse seven South American countries, offering a rich tapestry of landsc
apes and cultures.
[{'The': 'O'}, {'Andes,': 'location-mountain'}, {'the': 'O'}, {'longest': 'O'}, {'mountain': 'O'}, {'range': 'O'}, {'in': 'O'},
{'the': 'O'}, {'world,': 'O'}, {'traverse': 'O'}, {'seven': 'O'}, {'South': 'O'}, {'American': 'O'}, {'countries,': 'O'}, {'off
ering': 'O'}, {'a': 'O'}, {'rich': 'O'}, {'tapestry': 'O'}, {'of': 'O'}, {'landscapes': 'O'}, {'and': 'O'}, {'cultures.': 'O'}]


Japan's Mount Fuji, an active stratovolcano, is an iconic symbol and the highest peak in the country.
[{"Japan's": 'location-mountain'}, {'Mount': 'location-mountain'}, {'Fuji,': 'location-mountain'}, {'an': 'O'}, {'active':
'O'}, {'stratovolcano,': 'O'}, {'is': 'O'}, {'an': 'O'}, {'iconic': 'O'}, {'symbol': 'O'}, {'and': 'O'}, {'the': 'O'}, {'highes
t': 'O'}, {'peak': 'O'}, {'in': 'O'}, {'the': 'O'}, {'country.': 'O'}]


The Appalachian Mountains, stretching from Georgia to Maine, are known for their lush forests and historic significance in the
United States.
[{'The': 'O'}, {'Appalachian': 'location-mountain'}, {'Mountains,': 'location-mountain'}, {'stretching': 'O'}, {'from': 'O'},
{'Georgia': 'O'}, {'to': 'O'}, {'Maine,': 'O'}, {'are': 'O'}, {'known': 'O'}, {'for': 'O'}, {'their': 'O'}, {'lush': 'O'}, {'fo
rests': 'O'}, {'and': 'O'}, {'historic': 'O'}, {'significance': 'O'}, {'in': 'O'}, {'the': 'O'}, {'United': 'O'}, {'States.':
'O'}]


K2, the second-highest mountain on Earth, is part of the Karakoram Range and is considered one of the most challenging peaks to
climb.
[{'K2,': 'location-mountain'}, {'the': 'O'}, {'second-highest': 'O'}, {'mountain': 'O'}, {'on': 'O'}, {'Earth,': 'O'}, {'is':
'O'}, {'part': 'O'}, {'of': 'O'}, {'the': 'O'}, {'Karakoram': 'location-mountain'}, {'Range': 'location-mountain'}, {'and':
'O'}, {'is': 'O'}, {'considered': 'O'}, {'one': 'O'}, {'of': 'O'}, {'the': 'O'}, {'most': 'O'}, {'challenging': 'O'}, {'peaks':
'O'}, {'to': 'O'}, {'climb.': 'O'}]


The Cascade Range in the Pacific Northwest is home to notable volcanoes like Mount Rainier and Mount St. Helens.
[{'The': 'O'}, {'Cascade': 'location-mountain'}, {'Range': 'location-mountain'}, {'in': 'O'}, {'the': 'O'}, {'Pacific': 'O'},
{'Northwest': 'O'}, {'is': 'O'}, {'home': 'O'}, {'to': 'O'}, {'notable': 'O'}, {'volcanoes': 'O'}, {'like': 'O'}, {'Mount': 'lo
cation-mountain'}, {'Rainier': 'location-mountain'}, {'and': 'O'}, {'Mount': 'location-mountain'}, {'St.': 'location-mountai
n'}, {'Helens.': 'location-mountain'}]


The Atlas Mountains in North Africa extend across Morocco, Algeria, and Tunisia, providing a rugged and scenic landscape.
[{'The': 'O'}, {'Atlas': 'location-mountain'}, {'Mountains': 'location-mountain'}, {'in': 'O'}, {'North': 'O'}, {'Africa':
'O'}, {'extend': 'O'}, {'across': 'O'}, {'Morocco,': 'O'}, {'Algeria,': 'O'}, {'and': 'O'}, {'Tunisia,': 'O'}, {'providing':
'O'}, {'a': 'O'}, {'rugged': 'O'}, {'and': 'O'}, {'scenic': 'O'}, {'landscape.': 'O'}]


The Australian Alps, located in the southeastern part of the continent, offer unique alpine environments and are a haven for ou
tdoor enthusiasts.
[{'The': 'O'}, {'Australian': 'location-mountain'}, {'Alps,': 'location-mountain'}, {'located': 'O'}, {'in': 'O'}, {'the':
'O'}, {'southeastern': 'O'}, {'part': 'O'}, {'of': 'O'}, {'the': 'O'}, {'continent,': 'O'}, {'offer': 'O'}, {'unique': 'O'},
{'alpine': 'O'}, {'environments': 'O'}, {'and': 'O'}, {'are': 'O'}, {'a': 'O'}, {'haven': 'O'}, {'for': 'O'}, {'outdoor': 'O'},
{'enthusiasts.': 'O'}]
```