Convert Datataurks NER JSON ouptut to Spacy training data.

A simple script to create dataset in Spacy NER format.

Dataturks NER output is very close to the format used by Spacy, just that Spacy used Python tuples which are not supported by JSON standard, hence just use the below function to convert Dataturks JSON to Spacy training data.

Dataturks to Spacy

```
############################################  NOTE  ###########################################
#
#           Creates NER training data in Spacy format from JSON downloaded from Dataturks.
#
#           Outputs the Spacy training data which can be used for Spacy training.
#
###############################################################################################
def convert_dataturks_to_spacy(dataturks_JSON_FilePath):
    try:
        training_data = []
        lines=[]
        with open(dataturks_JSON_FilePath, 'r') as f:
            lines = f.readlines()

        for line in lines:
            data = json.loads(line)
            text = data['content']
            entities = []
            for annotation in data['annotation']:
                #only a single point in text annotation.
                point = annotation['points'][0]
                labels = annotation['label']
                # handle both list of labels or a single label.
                if not isinstance(labels, list):
                    labels = [labels]

                for label in labels:
                    #dataturks indices are both inclusive [start, end] but spacy is not [start,
                    entities.append((point['start'], point['end'] + 1 ,label))


            training_data.append((text, {"entities" : entities}))

        return training_data
    except Exception as e:
        logging.exception("Unable to process " + dataturks_JSON_FilePath + "\n" + "error = " +
        return None
```

s/71e66f7ce5ce7c101f0900da7be915da/raw/8d92541511856c925d2af0b1c524c8f9b14ac758/convert_dataturks_to_spacy.py)
convert_dataturks_to_spacy.py (https://gist.github.com/DataTurks/71e66f7ce5ce7c101f0900da7be915da#file-convert_dataturks_to_spacy-py) hosted with ♡ by **GitHub (https://github.com)**

Here is a sample code on to train a Spacy model from the above data:

Train Spacy

```python
import spacy
################### Train Spacy NER.###########
def train_spacy():
    TRAIN_DATA = convert_dataturks_to_spacy("dataturks_downloaded.json");
    nlp = spacy.blank('en')  # create blank Language class
    # create the built-in pipeline components and add them to the pipeline
    # nlp.create_pipe works for built-ins that are registered with spaCy
    if 'ner' not in nlp.pipe_names:
        ner = nlp.create_pipe('ner')
        nlp.add_pipe(ner, last=True)

    # add labels
    for _, annotations in TRAIN_DATA:
        for ent in annotations.get('entities'):
            ner.add_label(ent[2])

    # get names of other pipes to disable them during training
    other_pipes = [pipe for pipe in nlp.pipe_names if pipe != 'ner']
    with nlp.disable_pipes(*other_pipes):  # only train NER
        optimizer = nlp.begin_training()
        for itn in range(1):
            print("Statring iteration " + str(itn))
            random.shuffle(TRAIN_DATA)
            losses = {}
            for text, annotations in TRAIN_DATA:
                nlp.update(
                    [text],  # batch of texts
                    [annotations],  # batch of annotations
                    drop=0.2,  # dropout - make it harder to memorise data
                    sgd=optimizer,  # callable to update weights
                    losses=losses)
            print(losses)

    #do prediction
    doc = nlp("Samsing mobiles below $100")
    print ("Entities= " + str(["" + str(ent.text) + "_" + str(ent.label_) for ent in doc.ents])
```

DataTurks/f6035b1e58497d52bf88517ff7bf64cf/raw/8c583dbb8e426bebf3db8db6096bc19b0a9e4b68/train_spacy_NER.py)
train_spacy_NER.py (https://gist.github.com/DataTurks/f6035b1e58497d52bf88517ff7bf64cf#file-train_spacy_ner-py)
hosted with ♡ by GitHub (https://github.com)

You can also download a full python script to generate the Spacy training data and store it as a Pickle file. Download from: GitHub (https://gist.github.com/DataTurks/97ff613967e8139e57091f9299c3a104)

About

› Privacy Policy (/privacy.php)
› Open Data Policy (/open-data-policy.php)
› Help (/help/help.php)
› FAQs (/faq.php)

Blog

› All posts (/blog/blog.php)
› OCR APIs comparison (/blog/compare-image-text-recognition-apis.php)
› Face reco APIs comparison. (/blog/face-verification-api-comparison.php)
› Best image moderation APIs. (/blog/image-moderation-api-comparison.php)

How to

› Image Bounding Box. (/help/image-rectangle-bounding-box-annotation.php)
› Document Annotation. (/help/document-annotation-POS-NER.php)
› Polygon Bounding Box. (/help/image-polygon-bounding-box-annotation.php)
› POS Tagging. (/help/pos-text-annotations.php)

Features

› Image Annotations (/features/image-bounding-box.php)
› Text Annotations (/features/document-ner-annotation.php)
› API Docs (https://docs.dataturks.com/)

Documentation

› Export to Pascal VOC (/help/ibbx_dataturks_to_pascal_voc_format.php)

› Export in TensorFlow Format (https://medium.com/dataturks/converting-dataturks-image-classifier-tools-output-to-tensorflow-format-6f569e085bf3)
› NER in Spacy Format (/help/dataturks-ner-json-to-spacy-train.php)
› Docs (/help/help.php)

ML Tutorial

› Introduction to ML (/blog/intro-to-machine-learning-NER-deep-dive.php)
› ML based troll filter (https://medium.com/@dataturks/using-machine-learning-to-fight-cyber-trolls-9bf0fa1c5df9)
› ML and GDPR (https://medium.com/@dataturks/how-does-gdpr-impact-machine-learning-keystrokes-pascal-voc-and-much-more-1625b8a1147b)
› TensorFlow vs Keras? (https://hackernoon.com/tensorflow-vs-keras-comparison-by-building-a-model-for-image-classification-f007f336c519)

(+91) 080-331-72755, +91-99010-49915, +91-88614-08222

contact@dataturks.com

Say Hi: