Format our invoice dataset from the web according to chargrid requirements

Chargrid is an encoder-decoder based algorithm that extracts meaningful instances from document images.

Dataset

Our invoice dataset is composed of invoice images collected from the Internet. It contains two folders, `images` and `tagged`. `images` contains 231 invoice images. `tagged` contains the corresponding labels of these document images. The labels of a document image are saved under a `json` file with the same names as the invoice image file.

```
web-data
└──images
|   |   bild1.png
|   |   bild2.jpg
|   |   ...
└──tagged
    |   bild1.json
    |   bild2.json
    |   ...
```

Each labels file contains three components, `words`, `bbox` and `labels`.

```
[{"words":["ROCKETMEN","CED","DER","XXL-ONLINE","KURS","FÜR","SUPER-
SEX!","TIONSSTRASSE"...], "bbox":  [[253,9,446,27],[452,7,549,30],[253,37,282,48],
[285,37,374,48]...], "labels": ["O","O","O",...,"O","B-COMPANY","I-COMPANY","I-
COMPANY","I-COMPANY",...,"B-DATE", ...]
```

`words` is the list of text chunks detected by an `ocr` algorithm from the corresponding invoice image. `bbox` is a list of text chunk positions in the corresponding invoice image, each in the format [x_min, y_min, x_max, y_max]. `labels` is a list of text chunk labels. Labels including but not limited to: "B-COMPANY", "I-COMPANY", "B-TAG-INVOICE-NO", "I-TAG-INVOICE-NO", "B-INVOICE-NO", "I-INVOICE-NO", "B-TAG-CUSTOMER-NO", "I-TAG-CUSTOMER-NO", "B-CUSTOMER-NO", "I-CUSTOMER-NO", "B-TAG-TOTAL", "I-TAG-TOTAL", "B-TOTAL", "I-TOTAL", "B-TAG-IBAN", "I-TAG-IBAN", "B-IBAN", "I-IBAN", "O"...

Chargrid

Chargrid is an encoder-decoder based algorithm that extracts meaningful instances from document images.

Tasks

1. Find and read paragraphs related to data preparation in the chargrid paper.
2. What kind of input is required by the chargrid model? E.g. Chargrid requires image or text chunks or both as input? Does it require any location information? What kind of location information is required? In which case will

characters overlap with each other? How does chargrid deal with overlapping characters? e.t.c.

3. Please write as a single Python script the code that loads the data from disk and transforms the labels to the format required by the chargrid model. The code should be shared via email.