Issue Report

Issue Link: https://github.com/scrapinghub/extruct/issues/87

Description:

When extruct library parses the HTML page content to JSON-LD format there may be a chance of generating of invalid JSON-JD because of invalid HTML. So, contributor of this issue wants extruct library to have capability of handling bad JSON-LD and give a valid json-ld after parsing.

He gave an example as

"Some web pages contain badly formatted JSON-LD data, e.g., an example . The JSON-LD in this page is:

```
{
  "@context": "http://schema.org",
  "@type": "Product",
  "name": "Black 'Clint' FT0511 cat eye sunglasses",
  "image": "https://debenhams.scene7.com/is/image/Debenhams/60742_1515029001",
   "brand":
{ "@type": "Thing",
            "name": "Tom Ford"            },
"offers":{          "@type": "Offer",          "priceCurrency": "GBP",          "price": "285.00",
          "itemCondition": "http://schema.org/NewCondition",
"availability": "http://schema.org/InStock"
}
}
}
```

In the above JSON-LD there are the last "}" is extra. And extruct or json.loads won't handle it.

Error:

The json.loads in Python after 3.5 will give detailed error information as JSONDecodeError: Extra data: line 19 column 1 (char 624).

Solution For the given issue:

We observed this issue and make some changes in the library as given below:

We locate the place where extruct library is scrapping the HTML page content and changing it to a JSON-LD format. So extruct is doing in by its python module in jsonld.py. There is a function named **JsonLdExtractor(),** It is taking argument as scrapped HTML string and change it to a json-ld format. Here it uses to validate whether the HTML string is able to parse to JSON-LD or not in try- except statement. If HTML string is not parse able then it jumps from try block to except block here, we embedded our code as **validate_json()** to auto correct the parsed json-ld string.

validate_json(): It takes argument as json-ld string and with the help of this method, We validate the json and if there are any problem with more '{'/'}' then our embedded code try to remove the appropriate extra brackets and take some appropriate decisions to make this json-ld string correct.

After correcting this JSON-LD string we can parse it to JSON-LD format and return a valid JSON-LD.

Limitations of this solution:

Because of the time constraint we made a small part of validation that is This solution is the specific with the more bracket issue error. If there are errors regarding to other content it will not so perfect to use.