AgBlox Technical Interview

Entity Recognition with Spacy

The provided file `aapl-10k.txt` is an excerpt from Apple's most recent 10-K filing. It is a tab delimited file with two columns: the first column is a unique label of the paragraph, and the second column is the text in that paragraph section.

Utilizing the named entity recognition suite provided by the python package `spacy`, extract named entities from each paragraph and dump the results to a JSON file. The JSON file should contain a mapping from paragraph label to a paragraph object. The paragraph object should contain the full text of the paragraph, as well as a list of entity objects. The entity object should contain the full text of the entity, the start and end positions of the entity, as well as the spacy annotation / label for that particular entity. See the following output file example for the paragraph labeled "Properties":


```
{
  "Properties": {
    "text": "The Company\u2019s headquarters are located in Cupertino, California.",
    "entities": [
    {
      "text": "Cupertino",
      "start": 42,
      "end": 51,
      "label": "GPE"
    },
    {
      "text": "California",
      "start": 53,
      "end": 63,
      "label": "GPE"
    }
    ]
}
```


Note: the language model chosen will highly effect the validity of the named entities extracted. Entity labels may be invalid, you can ignore this; you may use the `en_core_web_sm` model for this exercise.