**Feature Name**: Address Extraction from Text Blobs

**Feature Statement**

Extract a subset of text representing a property address (street address, city, state, zip code) from an HOCR file, using text clusters which describe what the address will “look like” in the HOCR file.

An HOCR file is an open standard of data representation for formatted text. It is created when optical character recognition is applied to a document image. OCR encodes the image in terms of text, style, layout information, and other information using XML.

**Acceptance Criteria**

* Final code should be shared with heavywaterdevops Github username
* The Property Address Extraction algorithm should use any open source machine learning algorithm
* The Property Address Extraction algorithm should be trained on the subset of 20 sample HOCR files + text files containing test clusters provided in the October Codeathon I github repo:
  + https://github.com/codegarage/october-codeathon/tree/master/SampleData
* Once trained, the Property Address Extraction Algorithm should identify the property addresses in the HOCR files, based on the related text clusters
* As output, you should produce:
* A text file containing the property addresses found in the HOCR file
* A copy of the HOCR file, with tags on the nodes that were used to extract the property addresses
* The infrastructure should be on Amazon Web Services
* The algorithm needs to conform to machine learning standards
* The code for the Property Address Extraction Algorithm should be deployed using Continuous Integration and Infrastructure as Code tools available on AWS
* Functionality and infrastructure should be deployed on Amazon Web Services [without any manual steps](https://www.youtube.com/watch?v=WL2xSMVXy5w#t=39m30)
* The Property Address Extraction Algorithm should be deployed and demoed on AWS EC2 server
* A video file or video link should be committed to your Github repo, which includes the following components:
  + Instructions of how to run your code
  + Design and development steps that were taken
  + Demo of the functionality
  + Mention of some challenges that were faced
  + Proof on what makes your solution effective beyond the sample commands provided