Leeds Status Report Mar 21 to Mar 28

Weekly status report for Mar 21-Mar 28 2021

Scope	Schedule	Cost	Risks	Quality

Key Performance Indicators

- **Schedule:** ON SCHEDULE
  - Schedule Variance: 0 Days
  - Percent complete : **85%**
- **Labor:** ON SCHEDULE
  - Labor hours: 10 hours/person
- **Administration:** COMPLETE
  - Sponsor meeting, group meeting, attendance, hour log

Summary

As this week was the “Spring Pause” we decided to take last week off and start back up this week. For this week we have started to shift focus in how we are identifying company headers. We now have two methods, which are regex preformed on the optimal blob and now are using machine learning. I have spoken a lot about the regex approach so there is not much more to say about that one except that we are tweaking it and seeing the performance changes. For the new machine learning method, we are using a Tensorflow model. This model is being trained by us to identify the company headers on each image and display the bounding boxes. Our plan from there is to take the bounding boxes, and use them to find the words that correlate to their coordinates with Diego’s data.

Work planned for this week

- Make future improvements to extract info, CAP-78
- Collect info for company names found, CAP-80

Work completed this week

Make future improvements to extract info, CAP-78

Tensorflow model trained using this tutorial (https://tensorflow-object-detection-api-tutorial.readthedocs.io/en/latest/index.html). Currently it is being ran locally so the next step is to set it up to run on the research clusters GPU’s and have it process all the images.

Planned future improvements:

- It seems that many companies missed are foreign companies. Need to see if there are any common ending words for foreign companies as there are for English.
- Some of the company names span multiple lines so need to include line breaks

Collect info for company names found, CAP-80

Diego wants us to nail down identifying company names before we start to collect info for the company names we have found.

Plans for next week

- Improve regex methods to find ~5-10% more company names per year - CAP-82
- Run Tensorflow on cluster - CAP-81

Open Issues

Deliverables and Milestones for sprint (3/21 to 4/4)

Deliverable or Milestone	WBS	Planned	Forecasted	Actual	Status
Run Tensorflow on cluster	CAP-81	4/4/21	4/4/21		In progress
Improve regex methods to find ~5-10% more company names per year	CAP-82	4/4/21	4/4/21		In Progress

Open Change Requests

Change Request Name	Change Request Number	Requested Date	Current Status