Leeds Status Report Mar 8 to Mar 14

Weekly status report for Mar 8-Mar 14 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

This week we focused a lot on using our initial methods to process the data from the cluster. We processed all Industrial books from 1920 to 1949. Each page we stored the optimal brightness, the correlating blob, and the company names found on that page. We sent them to Diego so that he could compare our results to his results. This comparison will give us a good direction to go in for what improvements to make to our methods. We were also given a gold standard of the company names listed in each book from 1930 to 1939 so we can know what companies we have missed.

Work planned for this week

- Make future improvements to extract info, CAP-78

Work completed this week

Get results from all books with initial methods, CAP-79

Industrial books from 1920 to 1949 were processed. We are waiting on feedback from Diego to see how our results lined up with his. We were able to run each year on the cluster and parallelize all of them. Each year takes somewhere from 3-6 hours so we were able to run each one as an individual job and finished all of them in 6 hours. The database was able to process all of the write requests which was good to see.

Link to data: https://drive.google.com/file/d/1Pv7KMFtbxZYJREs46wMinxDJDdAUNfHo/view?usp=sharing

Make future improvements to extract info, CAP-78

Planned future improvements:

- Choose strings that are not all caps but strings that are 90% all caps
  - This will account for when the model accidently misreads a character
- String pages together so that company info that spans two pages can be captured
- Identify history and executives by words that have an edit distance very close to the original word

Upload golden standard, CAP-78

Diego gave us a golden standard for all the companies from 1930 to 1939 which will help us know how efficient our algorithms are.

Plans for next week

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

Open Issues

Deliverables and Milestones for sprint (3/1 to 3/14)

Deliverable or Milestone	WBS	Planned	Forecasted	Actual	Status
Extract info from optimal blob	CAP-77	3/1/21	3/8/21	3/7/21	Complete
Get results from all books with initial methods	CAP-79	3/12/21		3/12/21	Complete
Make future improvements to extract info	CAP-78	3/14/21	4/1/21		In progress

Open Change Requests

Change Request Name	Change Request Number	Requested Date	Current Status