# Leeds Status Report Mar 1 to Mar 7

Weekly status report for Mar 1-Mar 7 2021

| Scope | Schedule | Cost | Risks | Quality |
|---|---|---|---|---|

## Key Performance Indicators

- **Schedule:** ON SCHEDULE
  - Schedule Variance: 0 Days
  - Percent complete : **80%**
- **Labor:** ON SCHEDULE
  - Labor hours: 11 hours/person
- **Administration:** COMPLETE
  - Sponsor meeting, TA meeting, attendance, hour log

## Summary

This week we focused on using the code we have written to produce results in the Moody's Manuals. We ran our optimal blob and company name program on many years to see its initial performance. The results will not be perfect right away but this gives us a baseline and allows us to see areas where we could improve our code in order to recognize and pick out company names and sections better. Other than using Diego's data, we are also pursuing using our own models to generate data so that we can see if one OCR model works better than the other.

---

## Work planned for this week

- Extract information from optimal blob, CAP-77

---

## Work completed this week

### Extract information from optimal blob, CAP-77

It turns out many of the all capital strings are not company names. We tried a named entity NLP model to pick out company names but that did not work. We also tried to make a bag of words that are words that were commonly seen all caps strings but weren't company names but there were too many. The thing that ended up working was to keep any string that ended in something such as 'LTD, CO, CORP,' etc. This seems to recognize all of the company names.

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

### Add optimal blob for each page to DB

We have begun to add the optimal blob for each page to the database. This will allow us to not have to constantly be doing the same calculations and begin to focus on extracting the company info we are after.

---

## Plans for next week

## Open Issues

## Deliverables and Milestones for sprint (2/15 to 3/1)

| 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 |
| Make future improvements to extract info | CAP-78 | 3/14/21 | 3/14/21 | | In progress |

## Open Change Requests

| Change Request Name | Change Request Number | Requested Date | Current Status |
|---|---|---|---|
| | | | |