TesseractOCRStats

Here are some stats contributed by Mark Kerzner and Amanda Towler from Hyperion Gray.

```
Total number of images to process: about 300,000
Average time per image: about 1 sec
Total run time required: about 10 days
Our run times on various bathes: about 1 day total
OCR quality: decent
```

Future Work

- Use Tika, rather than do Tesseract directly
- Scale it up with Spark or Hadoop
- A few polishes, with the view on other teams/projects using it later